https://doi.org/10.1186/s13662-023-03797-x

Advances in Continuous and Discrete Models

RESEARCH

Open Access

Convergence of gradient descent for learning linear neural networks

Gabin Maxime Nguegnang[1*], Holger Rauhut[1,3] and Ulrich Terstiege[2]

*Correspondence:
nguegnang@math.lmu.de
[1] Department of Mathematics, Ludwig-Maximilians-Universität München, Theresienstr. 39, 80333, München, Germany
Full list of author information is available at the end of the article

Abstract

We study the convergence properties of gradient descent for training deep linear neural networks, i.e., deep matrix factorizations, by extending a previous analysis for the related gradient flow. We show that under suitable conditions on the stepsizes gradient descent converges to a critical point of the loss function, i.e., the square loss in this article. Furthermore, we demonstrate that for almost all initializations gradient descent converges to a global minimum in the case of two layers. In the case of three or more layers, we show that gradient descent converges to a global minimum on the manifold matrices of some fixed rank, where the rank cannot be determined a priori.

Keywords: Deep Learning; Gradient descent; Boundedness; Balancedness and Convergence

1 Introduction

Deep learning is arguably the most widely used and successful machine learning method, which has led to spectacular breakthroughs in various domains such as image recognition, autonomous driving, machine translation, medical imaging and many more. Despite its widespread use; the understanding of the mathematical principles of deep learning is still in its early stage, and has not yet been fully developed. Particular widely open questions concern the convergence properties of commonly used (stochastic) gradient descent (S)GD algorithms for learning a deep neural network from training data: Does (S)GD always converge to a critical point of the loss function? Does it converge to a global minimum? Does the network learned via (S)GD generalize well to unseen data? We contribute to the first two questions in the case of GD for linear neural networks.

To approach these questions, we study gradient descent for learning a deep *linear* network, i.e., a network with activation function being the identity, or in other words, learning a deep matrix factorization. While linear neural networks are not expressive enough for most practical applications, the theoretical study of gradient descent for linear neural networks is highly nontrivial and, therefore, expected to be very valuable. The difficulty in deriving mathematical convergence guarantees results from the minimizing functional being non-convex in terms of the individual matrices in the factorization. We are convinced that the case of linear networks should be well-understood before passing to the more difficult (but more practically relevant) case of nonlinear networks. We expect that some principles (though not all) will carry over to the nonlinear case, and the mathematical analysis of the linear case will provide valuable insights.

This article is a continuation of the work started in [5], where a theoretical analysis of the gradient flow related to learning a deep linear network via minimization of the square loss has been studied. Extending earlier contributions [2, 3, 7], it was shown in [5] that gradient flow always converges to a critical point of the square loss. Moreover, for almost all initializations, it converges to a global minimizer in the case of two layers. It is conjectured that this result also holds for more than two layers, but currently, it is only shown in [5] that for more layers, gradient flow converges to the global minimum of the loss function restricted to the manifold of matrices of some fixed rank k for almost all initializations, where unfortunately the result does not allow to determine k a priori.

We note here that the square loss in connection with linear networks has the nice property that all local minimizers are global, see [18], so that our analysis boils down to proving that (strict) saddle points are avoided almost surely. This remarkable property of the square loss is very specific and connected to the notion of Euclidean distance degree and properties of the manifold of fixed rank matrices, see [26, Appendix A.2] for more details.

As another interesting discovery, [5] considers the flow of the product matrix resulting from the gradient flow for the individual matrices in the factorization and identifies this flow of the product matrix as a Riemannian gradient flow. More precisely, the flow of the product matrix takes place on the manifold of matrices of a fixed rank k with respect to a nontrivial and explicitly given Riemannian metric on that manifold. This result requires that at initialization, the tuple of individual matrices is *balanced*, a term that the authors of [2] introduced. It is important to note that balancedness is preserved by the gradient flow, i.e., this property is related to the natural invariant set of the flow.

In this article, we extend the convergence analysis in [5] from gradient flow to gradient descent. Under certain conditions on the stepsizes, we show that the gradient descent iterations converge to a critical point of the square loss function. Moreover, for almost all initializations, our convergence is towards a global minimum in the case of two layers, while for more than two layers, we obtain the analog of the main result in [5] that for almost all initializations, the product matrix converges to a global minimum of the square loss restricted to the manifold of rank k matrices for some k.

We believe that the extension of the analysis from the gradient flow case to gradient descent is an important step, which turned out to be much more involved than one might initially expect. In fact, there are many works related to the convergence analysis of (stochastic) gradient descent methods in both convex and non-convex situations see, for instance, [17, 21, 22] and references therein. However, we are not aware of any results that are directly applicable to our setting of deep linear networks (and also not to most nontrivial setups for nonlinear networks). In fact, it is common to assume a loss function with Lipschitz gradient. However, due to factorization of the layers and unbounded domain, such Lipschitz assumptions will not be satisfied. Note that our analysis shows the boundedness of all the iterates so that we could, in principle, restrict to a bounded domain, but this needs to be shown first, which is a major part of this work. Hence, our analysis required work without such Lipschitz gradient assumptions and, therefore, may be of independent interest. Moreover, the existing gradient flow analysis in [3, 5] does not provide any hint on conditions on the stepsizes that ensure convergence of its discrete version gradient descent, which is another reason why we believe that our work can be of value.

The difficulties in establishing the extension of the gradient flow analysis to the gradient descent are due to the fact that the gradient descent iterations no longer satisfy exactly the invariance property related to the balancedness. This property of the gradient flow, however, was heavily used in the convergence proof in [5]. In order to circumvent this problem, we develop an induction argument inspired by the article [11], which covers the significantly simpler special case of two layers. The induction proof tracks, in particular, how much the balancedness condition is perturbed during the iterations. In fact, such perturbations stay bounded under suitable assumptions on the stepsizes. In particular, this allows for the bounding of all the individual factors in the linear network.

Learning linear networks are currently also studied in the context of the so-called implicit bias of gradient descent and gradient flows [2, 8, 14, 15, 19, 24, 28, 30]. We expect that the convergence analysis of gradient descent performed in our paper will also be a useful tool for the detailed analysis of the implicit bias of (stochastic) gradient descent in learning deep overparameterized neural networks.

1.1 Relation to previous work

For the scenario of learning deep linear networks, works done in [2, 6, 13, 27–29, 31] study the convergence of gradient descent. The authors of [13] provided a guarantee of convergence to global minimizers for gradient descent with random balanced near-zero initialization. Their proof proceeds by transferring the convergence properties of gradient flow to gradient descent. In contrast, based on the Lojasiewicz theorem, we directly prove that gradient descent converges to a critical point of the square loss of deep linear networks. Then we extend the result in [5] that for almost all initializations gradient descent converges to the global minimum for networks of depth 2. For three or more layers, we prove that gradient descent converges to a global minimum on a manifold of a fixed rank. The convergence result in [13] is restricted to a simple scalar regression problem with near-zero initialization and constant stepsize, whereas our result works for the general multivariate case, almost all initializations and not necessarily constant stepsize. Under certain conditions, convergence of the stochastic (sub)gradient method to a critical point has been established in [9]. This result requires the subgradient sequence to be bounded and the cost function to be strictly decreasing along any trajectory of the differential inclusion proceeding from a noncritical point. In addition, the authors of [9] comment that the boundedness of the iterates may be enforced by assuming that the constraint set on which the set valued map is defined is bounded or by a proper choice of a regularizer. In contrast, we do not require these conditions. We rather prove the boundedness of the gradient descent sequence and demonstrate the strong descent condition of this sequence. The authors of [10, 16] address a multivariate regression problem and prove that gradient descent with Gaussian resp. orthogonal random initialization and constant stepsize converges to a global minimum. The result in [16] requires that the hidden layer dimension should be greater than the dimension of the input data with orthogonal initialization, and the one in [10] assumes that the hidden layer dimension is greater than the dimension of the output data. Compared to these results, our result is more general in the sense that it does not require these conditions which exclude some important models such as auto-encoders where the dimensions of the intermediate layers are commonly less than the input and output dimensions. Moreover, our result does not require the initialization to be close enough to a global minimum (as in [2]), and the maximum allowed stepsize in

Theorem 2.4 does not decay exponentially with depth (Remark 2.5(b)). In this sense, our theorem is less restrictive.

Our article is structured as follows. Section 2 introduces deep linear networks and gradient descent, recalls the recent results from [5] on gradient flows, and presents our two main results on convergence to a critical point and convergence to a global minimizer for almost all initializations. Section 3 provides the proof of convergence to critical points (in the sense described above), while Sect. 4 is dedicated to the proof of convergence to global minimizers. Finally, Sect. 5 presents numerical experiments illustrating our results.

1.2 Notation

The standard ℓ_p-norm on $\mathbb{R}^d$ will be denoted by $\|x\|_p = (\sum_{j=1}^{d} |x_j|^p)^{1/p}$ for $1 \leq p < \infty$. We write the spectral norm on $\mathbb{R}^{d\times m}$ as $\|A\| = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}(A)$ is the largest singular value of A. Moreover, we let $\sigma_{\min}(A) = \min_{\|x\|_2=1} \|Ax\|_2$ be the smallest singular value of A. The trace of a matrix A is denoted as $\operatorname{tr}(A)$, and its Frobenius norm is defined as $\|A\|_F = \sqrt{\operatorname{tr}(A^TA)} = \sqrt{\sum_{j,k} |A_{j,k}|^2}$. We will often combine matrices $W_1, \ldots, W_N$ into a tuple $\overrightarrow{W} = (W_1, \ldots, W_N)$. We define the Frobenius inner product of two such tuples $\overrightarrow{W}$ and $\overrightarrow{V}$ as $\langle \overrightarrow{W}, \overrightarrow{V} \rangle_F = \sum_{j=1}^{N} \operatorname{tr}(W_j^T V_j)$ and the corresponding Frobenius norm as $\|\overrightarrow{W}\|_F = \sqrt{\langle \overrightarrow{W}, \overrightarrow{W} \rangle_F} = (\sum_{j=1}^{N} \|W_j\|_F^2)^{1/2}$. The operator norm of a mapping $\mathcal{A}$ acting between tuples of matrices will be denoted as $\|\mathcal{A}\|_{F\to F} = \max_{\|\overrightarrow{W}\|_F=1} \|\mathcal{A}(\overrightarrow{W})\|_F$. We introduce $[d_j] = \{1, 2, \ldots, d_j\}$ with $d_j \in \mathbb{N}$.

2 Linear neural networks and gradient descent analysis

A neural network is a function $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ of the form

$$f(x) = f_{W_1,\ldots,W_N,b_1,\ldots,b_N}(x) = g_N \circ g_{N-1} \cdots \circ g_1(x),$$

where the so-called layers $g_j : \mathbb{R}^{d_{j-1}} \to \mathbb{R}^{d_j}$ are the composition of an affine function with a componentwise activation function, i.e.,

$$g_j(z) = \sigma(W_j z + b_j), \quad \text{for } W_j \in \mathbb{R}^{d_j \times d_{j-1}}, b_j \in \mathbb{R}^{d_j},$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ applied to a vector $w \in \mathbb{R}^{d_j}$ acts as $(\sigma(w))_k = \sigma(w_k)$, $k \in [d_j]$. Here, $d_0 = d_x$ and $d_N = d_y$, while $d_1, \ldots, d_{N-1} \in \mathbb{N}$ are some numbers. Prominent examples for activation functions used in deep learning include $\sigma(t) = \operatorname{ReLU}(t) = \max\{0, t\}$ and $\sigma(t) = \tanh(t)$, but we will simply choose the identity $\sigma(t) = t$ in this article.

Learning a neural network $f = f_{W_1,\ldots,W_N,b_1,\ldots,b_N}$ consists in adapting the parameters W_j, b_j based on labeled training data, i.e., pairs (x_i, y_i) of input data $x_1, \ldots, x_m \in \mathbb{R}^{d_x}$ and output data $y_1, \ldots, y_m \in \mathbb{R}^{d_y}$ in a way that $f_{W_1,\ldots,W_N,b_1,\ldots,b_N}(x_i) \approx y_i$ for $i \in [m]$. Ideally, the learned neural network f should generalize well to unseen data, i.e., it should predict well the label y corresponding to new input data x. However, we will not discuss this point further in this article.

The learning process is usually performed via optimization. Given a loss function $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}_+$ (usually satisfying $\ell(y, y) = 0$), one aims at minimizing the empirical risk function

$$\mathcal{L}(W_1,\dots,W_N,b_1,\dots,b_N) = \sum_{i=1}^{m} \ell\big(f_{W_1,\dots,W_N,b_1,\dots,b_N}(x_i), y_i\big)$$

with respect to the parameters $W_1,\dots,W_N,b_1,\dots,b_N$. Gradient descent and stochastic gradient descent algorithms are most commonly used for this task. Convergence analysis of these algorithms is challenging in general since, due to the compositional nature of neural networks, the function $\mathcal{L}$ is not convex in general.

Due to this difficulty, we reduce to the special case of *linear* neural networks in this article, i.e., we assume that $\sigma(t) = t$ is the identity and that $b_j = 0$ for all j. Consequently, a linear neural network takes the form

$$f(x) = f_{W_1,\dots,W_N}(x) = W_N \cdots W_1 x = Wx, \quad \text{where } W = W_N \cdot W_{N-1} \cdots W_1.$$

While linear networks may not be expressive enough for many applications, convergence properties of gradient descent applied to learning linear neural networks are still nontrivial to understand. We will concentrate on the square-loss $\ell(z,w) = \frac{1}{2}\|z-w\|_2^2$ here, so that our learning problem consists in minimizing

$$L^N(W_1,\dots,W_N) = \frac{1}{2}\sum_{i=1}^{m} \|y_i - W_N \cdots W_1 x_i\|_2^2 = \frac{1}{2}\|Y - W_N \cdots W_1 X\|_F^2,$$

where the data matrix $X \in \mathbb{R}^{d_x \times m}$ contains the data points $x_i \in \mathbb{R}^{d_x}$, $i = 1,\dots,m$ as columns and likewise the matrix $Y \in \mathbb{R}^{d_y \times m}$ contains the label points $y_i \in \mathbb{R}^{d_y}$, $i = 1,\dots,m$. The function L^1 is given by

$$L^1(W) = \frac{1}{2}\|Y - WX\|_F^2.$$

Note that the rank of the matrix $W = W_N \cdots W_1$ is at most $r := \min_{i=0,\dots,N} d_i$, which is strictly smaller than $\min\{d_x, d_y\}$ if one of the "hidden" dimensions d_i is smaller than this number. Hence, we can also view the learning problem as one of minimizing $L^1(W)$ under the constraint $\operatorname{rank}(W) \le r$. Instead of directly minimizing over W, we choose an *overparameterized* representation as $W = W_N \cdots W_1$ and consider gradient descent with respect to each factor W_i. While overparameterization seems to be a waste of resources at first sight, it also has certain advantages as it can even accelerate convergence [4] (at least for ℓ_p-losses with $p > 2$) or lead to solutions with better generalization properties [30]. Moreover, we expect that understanding theory for overparameterization in linear neural network will also give insights for overparameterization in nonlinear networks, which is widely used in practice. While the speed of convergence or implicit bias are certainly of interest on their own, we will not delve into this but rather concentrate on mere convergence here.

We consider gradient descent for the loss function L^N with stepsizes η_k, i.e.,

$$W_j(k+1) = W_j(k) - \eta_k \nabla_{W_j} L^N\big(W_1(k),\dots,W_N(k)\big). \tag{2.1}$$

We further define the matrix W at each iteration k by

$$W(k) = W_N(k) \cdots W_1(k).$$

Before discussing gradient descent itself, let us recall previous results for the related gradient flow, which will guide the intuition for the analysis in this paper.

2.1 Gradient flow analysis

The gradient flow $\overrightarrow{W}(t) = (W_1(t), \dots, W_N(t))$, $t \in \mathbb{R}_+$ for the function L^N is defined via the differential equation

$$\frac{d}{dt} W_j(t) = -\nabla_{W_j} L^N\big(W_1(t), \dots, W_N(t)\big), \qquad W_j(0) = W_{j,0}, \quad j = 1, \dots, N, \tag{2.2}$$

for some initial matrices $W_{j,0} \in \mathbb{R}^{d_j \times d_{j-1}}$. This flow represents the continuous analog of the gradient descent algorithm and has been analyzed in [2, 3, 5, 7].

An important invariance property of the gradient flow (2.2) consists in the fact that the differences

$$W_{j+1}^T(t) W_{j+1}(t) - W_j(t) W_j^T(t), \quad j = 1, \dots, N \tag{2.3}$$

are constant in time, see [2, 3, 5, 7]. This motivates to call a tuple $\overrightarrow{W} = (W_1, \dots, W_N)$ *balanced* if

$$W_{j+1}^T W_{j+1} = W_j W_j^T \quad \text{for all } j = 1, \dots, N. \tag{2.4}$$

If $\overrightarrow{W}(0) = (W_{1,0}, \dots, W_{N,0})$ is balanced, then $\overrightarrow{W}(t)$ is balanced for all $t \in \mathbb{R}_+$ as a consequence of the invariance property. Note that by taking the trace on both sides of (2.4), we see that balancedness implies $\|W_j\|_F = \|W_1\|_F$ for all $j = 1, \dots, N$.

It is useful to introduce the "end-to-end" matrix $W(t) = W_N(t) \cdots W_1(t)$, which describes the action of the resulting network and is the object of main interest. It was shown in [3] that if the initial tuple $\overrightarrow{W}(0)$ (and hence $\overrightarrow{W}(t)$ for any $t \geq 0$) is balanced then the dynamics of $W(t)$ can be described without making use of the individual matrices $W_j(t)$. More precisely, it satisfies the differential equation

$$\frac{d}{dt} W(t) = -\mathcal{A}_{W(t)}(\nabla L^1\big(W(t)\big), \tag{2.5}$$

where $\mathcal{A}_W : \mathbb{R}^{d_x \times d_y} \to \mathbb{R}^{d_x \times d_y}$ is the linear map

$$\mathcal{A}_W(Z) = \sum_{j=1}^{N} \big(WW^T\big)^{\frac{N-j}{N}} \cdot Z \cdot \big(W^T W\big)^{\frac{j-1}{N}}.$$

One feature of the flow in (2.5), see [5, Theorem 4.5], is that the rank of $W(t)$ is constant in t, i.e., if $W(0) = W_N(0) \cdots W_1(0)$ has rank r then the $W(t)$ stays in the manifold of rank r matrices for all $t \geq 0$ (but note that the rank may drop in the limit). This property may fail for non-balanced initializations [5, Remark 4.2]. Another interesting observation (which, however, will not be important in our article) is that (2.5) can be interpreted as Riemannian gradient flow with respect to an appropriately defined Riemannian metric on the manifold of rank r matrices, see [5] for all the details.

The convergence properties of the gradient flow (2.2) (in both the unbalanced and balanced case) can be summarized in the following theorems. The first one from [5, Theorem 3.2] significantly generalizes the main result of [7].

Theorem 2.1 [5, *Theorem* 3.2] *Assume that XX^T has full rank. Then, the flow $\overrightarrow{W}(t)$ defined by* (2.2) *is defined and bounded for all $t \geq 0$ and converges to a critical point of L^N as $t \to \infty$.*

This result is shown via the Lojasiewicz theorem [1], which requires, in turn, to show boundedness of all components $W_i(t)$ of $\overrightarrow{W}(t)$. While the boundedness is straightforward to show for $W(t)$, it is a nontrivial property of the $W_i(t)$. In fact, the proof exploits the invariance of the differences in (2.3).

While convergence to a critical point is nice to have, we would like to obtain more information about the type of critical point, whether it is a global or local minimum or merely a saddle point. Note that the function L^N built from the square loss has the nice (but rare) property that a local minimum is automatically a global minimum [18, 26]. This means that we only need to single out saddle points. Also, observe that we cannot expect to have convergence to a global minimizer for any initialization because the flow will not move when initializing in any critical point, so we cannot expect convergence to a global minimizer if that critical point is not already a global minimizer. The following result, valid for almost all initializations, was derived in [5, Theorem 6.12]. In order to state it, we need to introduce the matrix

$$Q = YX^T \left(XX^T\right)^{-1/2}, \tag{2.6}$$

assuming that XX^T has full rank.

Theorem 2.2 [5, *Theorem* 6.12] *Assume that XX^T has full rank, let $q = \operatorname{rank}(Q)$, $r = \min_{j=0,\dots,N} d_j$ and $\bar{r} = \min\{q, r\}$ where Q is the matrix defined in* (2.6).

(a) *For almost all initializations $\overrightarrow{W}(0)$, the flow* (2.2) *converges to a critical point $\overrightarrow{W}^* = (W_1^*, \dots, W_N^*)$ of L^N such that $W^* := W_N^* \cdots W_1^*$ is a global minimizer of L^1 on the manifold of matrices of fixed rank k for some $0 \leq k \leq \bar{r}$.*

(b) *If $N = 2$, then for almost all initial values $W_1(0), \dots, W_N(0)$, the flow converges to a global minimizer of L^N on $\mathbb{R}^{d_0 \times d_1} \times \cdots \times \mathbb{R}^{d_{N-1} \times d_N}$.*

We conjecture that the statement in part (b) also holds for $N \geq 3$, or in other words, that we can always choose the maximal possible rank $k = \bar{r}$ in (a), but unfortunately, the proof method employed in [5] is not able to deliver this extension without making significant adaptations. In fact, the proof relies on an abstract result, see [20] and [5, Theorem 6.3], which states that for almost all initializations, so-called strict saddle points are avoided as limits. Unfortunately, if $N \geq 3$, then minimizers of L^1 restricted to the manifold of matrices of rank $k < \bar{r}$ may correspond to non-strict saddle points of L^N, see [18] and [5, Proposition 6.10], so that the abstract result does not apply to these points.

2.2 Gradient descent analysis

Our main goal is to extend Theorems 2.1 and 2.2 from gradient flow (2.2) to gradient descent (2.1). The balancedness, or more generally, the invariance property, see (2.3), does not appear explicitly in the statements of these theorems for gradient flow, although the invariance property is key in showing boundedness of the flow in the proof of Theorem 2.1. It turns out that balancedness does play an explicit role in the conditions for the stepsizes ensuring convergence. Unfortunately, the invariance of the differences in (2.3) does not carry over to the iterations of gradient descent, which prevents directly following the proof strategy of [5] for showing the boundedness of the iterates. Nevertheless, we will prove that under suitable conditions on the stepsizes, the differences in (2.3) will stay bounded in norm, which then allows us to show the boundedness of the components $W_j(k)$ of $\overrightarrow{W}(k)$ and to apply Lojasiewicz' theorem to show convergence to a critical point.

In order to state our main results, we introduce the following definition.

Definition 2.3 We say that a tuple $\overrightarrow{W} = (W_1, \dots, W_N)$ has *balancedness constant* $\delta \geq 0$ if

$$\left\| W_{j+1}^T W_{j+1} - W_j W_j^T \right\| \leq \delta \quad \text{for all } j = 1, \dots, N-1. \tag{2.7}$$

Obviously, (2.7) quantifies how much the tuple $\overrightarrow{W}$ deviates from being balanced, measured in the spectral norm. Note that the authors of [2] introduced a very similar notion and said $\overrightarrow{W} = (W_1, \dots, W_N)$ to be δ-balanced if (2.7) holds with the spectral norm replaced by the Frobenius norm.

The following Theorem 2.4 indicates that GD with approximately balanced initialization converges to a critical point of L^N. This theorem provides suitable conditions on the stepsizes that guarantee convergence.

Theorem 2.4 *Let $X \in \mathbb{R}^{d_x \times m}$, $Y \in \mathbb{R}^{d_y \times m}$ be data matrices such that XX^T is of full rank. Suppose that the initialization $\overrightarrow{W}(0)$ of the gradient descent iterations (2.1) has balancedness constant $\alpha\delta$ for some $\delta > 0$ and $\alpha \in [0,1)$. Assume that the stepsizes $\eta_k > 0$ satisfy $\sum_{k=0}^{\infty} \eta_k = \infty$ and*

$$\eta_k \leq \frac{2(1-\alpha)\delta}{4L^N(\overrightarrow{W}(0)) + (1-\alpha)\delta B_\delta} \quad \textit{for all } k \in \mathbb{N}_0, \tag{2.8}$$

where

$$B_\delta := 2eNK_\delta^{N-1} \|X\|^2 + \sqrt{e}NK_\delta^{\frac{N}{2}-1} \left\| XY^T \right\|, \tag{2.9}$$

$$K_\delta := M^{\frac{2}{N}} + (N+1)^2 \delta, \tag{2.10}$$

$$M := \frac{\sqrt{2L^N(\overrightarrow{W}(0))} + \|Y\|}{\sigma_{\min}(X)} = \frac{\sqrt{2}\|Y - W_N(0) \cdots W_1(0)X\|_F + \|Y\|}{\sigma_{\min}(X)}. \tag{2.11}$$

Then, the sequence $\overrightarrow{W}(k)$ converges to a critical point of L^N.

The theorem regarding convergence to a critical point of L^N stated above will be proven in the upcoming Sect. 2.4.

Remark 2.5

(a) If $\overrightarrow{W}(0)$ is balanced, i.e., has balancedness constant 0, we can choose $\alpha = 0$ above. Then, for any $\delta > 0$, choosing the stepsizes η_k such that (3.14) below is satisfied ensures convergence to a critical point and that all the iterates $\overrightarrow{W}(k)$, $k \in \mathbb{N}$, have δ, see Proposition 3.4. This latter property will be a crucial ingredient for the proof of the theorem.

(b) Intuitively, the stepsizes η_k should be chosen as large as possible in order to have fast convergence in practice, while it does not seem to be crucial to have the balancedness constant δ as small as possible during the iterations. This suggests maximizing the right-hand side of (2.8) with respect to δ in order to make the condition on the stepsizes as weak as possible. While analytical maximization seems difficult, this may be done numerically in practice. A reasonably good choice for δ seems to be

$$\delta = \frac{1}{N(N+1)^2} M^{\frac{2}{N}}.$$

Then, $K_\delta = (1 + \frac{1}{N})M^{\frac{2}{N}}$ so that $K_\delta^N \leq eM^2$ where e is Euler's constant. Since $2L^N(\overrightarrow{W}(0)) \leq \sigma_{\min}^2(X)M^2$, Condition (2.8) is then satisfied if

$$\eta_k \leq \frac{2}{2(1-\alpha)^{-1}N(N+1)^2M^{2-\frac{2}{N}}\sigma_{\min}^2(X) + 2e^{2-\frac{1}{N}}NM^{2-\frac{2}{N}}\|X\|^2 + e^{1-\frac{1}{N}}NM^{1-\frac{2}{N}}\|XY^T\|}.$$

For a network of depth, this means that δ is of the order $\delta = \mathcal{O}(N^{-2})$, and the stepsizes are required to be of order $\eta = \mathcal{O}(N^{-3})$.

(c) The stepsizes η_k in the theorem can be chosen a priori, for instance, $\eta_k = \eta$ (constant stepsize), or $\eta_k = ck^{-\alpha}$ for some $\alpha \in [0,1)$, or adaptively, i.e., depending on the current iterate $\overrightarrow{W}(k)$, as long as the stepsize condition (2.8) is satisfied. In practice, it seems that a large constant stepsize leads to the best performance in terms of convergence speed.

Of course, more information on the type of critical point to which $\overrightarrow{W}(k)$ converges is desirable. Our next theorem states the analog of Theorem 2.2 that essentially convergence is towards global minimizers for almost all initializations. Since Condition (3.14) on the stepsizes η_k ensuring mere convergence to a critical point depends on the initialization $\overrightarrow{W}(0)$, we can only expect to state a result for almost all initializations for sets of tuples $\overrightarrow{W}$ of matrices for which the balancedness constant δ and M in (3.14) have a uniform upper bound. Consequently, we choose $\mathcal{B} \subset \mathbb{R}^{d_0 \times d_1} \times \cdots \times \mathbb{R}^{d_{N-1} \times d_N}$ to be bounded and let

$$\delta_{\mathcal{B}} = \sup_{\overrightarrow{W} \in \mathcal{B}} \max_{j=1,\dots,N-1} \left\| W_{j+1}^T W_{j+1} - W_j W_j^T \right\|, \tag{2.12}$$

$$L_{\mathcal{B}} = \sup_{\overrightarrow{W} \in \mathcal{B}} L^N(\overrightarrow{W}), \qquad M_{\mathcal{B}} = \left(\sqrt{2L_{\mathcal{B}}} + \|Y\|\right)\sigma_{\min}^{-1}(X). \tag{2.13}$$

Note that $\delta_{\mathcal{B}}$ and $M_{\mathcal{B}}$ are finite (assuming that XX^T has full rank) since L^N is continuous. Let us also recall the definition of the matrix $Q = YX^T(XX^T)^{-1/2}$ in (2.6).

Theorem 2.6 *Let $\mathcal{B} \subset \mathbb{R}^{d_0 \times d_1} \times \cdots \times \mathbb{R}^{d_{N-1} \times d_N}$ be a bounded set with constants $\delta_{\mathcal{B}} \leq \alpha\delta$ as in (2.12) for some $\delta > 0$ and $\alpha \in [0,1)$ and $L_{\mathcal{B}}$, $M_{\mathcal{B}}$ defined by (2.13). Let $q = \operatorname{rank}(Q)$, $r = \min\{d_0, \ldots, d_N\}$ and $\bar{r} = \min\{q, r\}$, and let $(\eta_k)_{k \in \mathbb{N}_0}$ be a sequence of positive stepsizes such that*

$$\eta_k \leq \frac{2(1-\alpha)\delta}{4L_{\mathcal{B}} + (1-\alpha)\delta B_\delta} \quad \text{for all } k \in \mathbb{N}_0, \tag{2.14}$$

where

$$K_\delta := M_{\mathcal{B}}^{\frac{2}{N}} + (N+1)^2\delta, \qquad B_\delta := 2eNK_\delta^{N-1}\|X\|^2 + \sqrt{e}NK_\delta^{\frac{N}{2}-1}\left\|XY^T\right\|.$$

Assume that additionally one of the following conditions is satisfied.

(1) *The sequence (η_k) is constant, i.e., $\eta_k = \eta$ for some $\eta > 0$ for all $k \in \mathbb{N}$.*

(2) *It holds*

$$\eta_k \geq C\frac{1}{k} \quad \text{for some } C > 0 \quad \text{and} \quad \lim_{k\to\infty} \eta_k = 0.$$

Then, the following statements hold.

(a) *For almost all initializations $\overrightarrow{W}(0) = (W_1(0), \ldots, W_N(0)) \in \mathcal{B}$, gradient descent (2.1) with stepsizes η_k converges to a critical point $\overrightarrow{W}$ of L^N such that $W = W_N \cdots W_1$ is a global minimum of L^1 on the manifold $\mathcal{M}_k$ of matrices of rank $k = \operatorname{rank}(W) \in \{0, 1, \ldots, \bar{r}\}$ on $\mathbb{R}^{d_N \times d_0}$.*

(b) *For $N = 2$, gradient descent (2.1) converges to a global minimum of L^N on $\mathbb{R}^{d_0 \times d_1} \times \mathbb{R}^{d_1 \times d_2}$ for almost all $\overrightarrow{W}(0) = (W_1(0), W_2(0)) \in \mathcal{B}$.*

The proof of the global convergence theorem stated above can be found in Sect. 4.

Similar to Theorem 2.2, we conjecture that part (b) extends to $N \geq 3$ or equivalently that part (a) holds with $k = \bar{r}$. As for Theorem 2.2, the current proof method based on a strict saddle point analysis cannot be extended to show this conjecture.

It is currently not clear whether the theorem holds under more general assumptions on the stepsizes η_k, i.e., whether it is necessary that one of the two additional conditions on η_k holds. The current proof can only handle those two cases, for corresponding abstract results are available, see [20, 23]. It seems crucial for these general results that the stepsizes are chosen a priori and independently of the choice of $\overrightarrow{W}(0)$ (or the further iterates). In particular, adaptive stepsize choices are not covered by our theorem. We note that the bounds on the stepsizes are reasonable for practical purposes. In particular, the stepsize choices in our numerical experiments meet these bounds.

3 Convergence to critical points

We will prove Theorem 2.4 in this section. For $\overrightarrow{W} = (W_1, \ldots, W_N)$ will always denote the corresponding product matrix by

$$W = W_N \cdots W_1,$$

and similarly, we denote by $W(k) = W_N(k)\cdots W_1(k)$ the sequence of product matrices associated to a sequence $\overrightarrow{W}(k) = (W_1(k),\dots,W_N(k))$, $k \in \mathbb{N}_0$. We recall from [2, 3, 5, 7] that

$$\nabla L^1(W) = WXX^T - YX^T, \tag{3.1}$$

$$\nabla_{W_j} L^N(W_1,\dots,W_N) = W_{j+1}^T \cdots W_N^T \nabla L^1(W) W_1^T \cdots W_{j-1}^T. \tag{3.2}$$

3.1 Auxiliary bounds

We start with a useful bound for $\|W\|$ in terms of $L^1(W)$.

Lemma 3.1 *Assume that XX^T has full rank. Then, $W \in \mathbb{R}^{d_x \times d_y}$ satisfies*

$$\|W\| \le \big(\|Y - WX\| + \|Y\|\big)\sigma_{\min}^{-1}(X) \le \big(\sqrt{2L^1(W)} + \|Y\|\big)\sigma_{\min}^{-1}(X). \tag{3.3}$$

Consequently, if $L^N(\overrightarrow{W}(k)) \le L^N(\overrightarrow{W}(0))$, then

$$\big\|W(k)\big\| = \big\|W_N(k)\cdots W_1(k)\big\| \le \Big(\sqrt{2L^N\big(\overrightarrow{W}(0)\big)} + \|Y\|\Big)\sigma_{\min}^{-1}(X).$$

Furthermore,

$$\big\|\nabla L^1(W)\big\| \le \|WX - Y\|\|X\| \le \sqrt{2L^1(W)}\|X\|. \tag{3.4}$$

Proof Arguing similarly to the proof of [5, Theorem 3.2] gives

$$\begin{aligned}\|W\| &= \big\|WXX^T\big(XX^T\big)^{-1}\big\| \le \|WX\|\big\|X^T\big(XX^T\big)^{-1}\big\| \le \big(\|Y - WX\| + \|Y\|\big)\sigma_{\min}^{-1}(X)\\ &\le \big(\|Y - WX\|_F + \|Y\|\big)\sigma_{\min}^{-1}(X) = \big(\sqrt{2L^1(W)} + \|Y\|\big)\sigma_{\min}^{-1}(X).\end{aligned}$$

The second claim follows then as an easy consequence recalling that $L^1(W(k)) = L^N(\overrightarrow{W}(k))$.

For the third claim, we use the explicit formula (3.1) for the gradient of L^1 to conclude that

$$\big\|\nabla L^1(W)\big\| = \big\|WXX^T - YX^T\big\| \le \|WX - Y\|\big\|X^T\big\| \le \|WX - Y\|_F\|X\| = \sqrt{2L^1(W)}\|X\|.$$

This completes the proof. □

A crucial ingredient in our proof is to show the boundedness of all matrices $W_j(k)$, $k \in \mathbf{N}_0$. While boundedness for the product $W(k) = W_N(k)\cdots W_1(k)$ follows easily from the previous lemma, it does not immediately imply boundedness of all the factors $W_j(k)$. For instance, multiplying one factor $W_j(k)$ by a constant $\alpha > 0$ and another factor $W_\ell(k)$ by α^{-1} leaves the product $W(k)$ invariant but changes the norm of $W_j(k)$ and $W_\ell(k)$. In particular, letting $\alpha \to \infty$ shows that a bound for $W(k)$ alone does not imply boundedness for $W_j(k)$, $k \in \mathbb{N}_0$. This is where the balancedness comes in. In particular, if a tuple $\overrightarrow{W} = (W_1,\dots,W_N)$ has balancedness constant $\delta \ge 0$, then we can bound $\|W_j\|$, $j = 1,\dots,N$, by an expression (continuously) depending on $\|W\|$. This is the essence of the next statement.

Proposition 3.2 *Let $\overrightarrow{W} = (W_1, \dots, W_N) \in \mathbb{R}^{d_0 \times d_1} \times \cdots \times \mathbb{R}^{d_{N-1} \times d_N}$ with balancedness constant $\delta \geq 0$, and let $W = W_N \cdots W_1$. Then,*

$$\|W_j\|^2 \leq \|W\|^{\frac{2}{N}} + (N+1)^2 \delta \quad \textit{for all } j = 1, \dots, N.$$

Remark 3.3 With a significantly longer proof, one can improve this result to

$$\|W_j\|^2 \leq \|W\|^{\frac{2}{N}} + N^2 \delta \quad \text{for all } j = 1, \dots, N.$$

However, since this does not significantly improve our results, we decided to present the slightly weaker bound in order to keep the proof short.

Proof We will first prove that

$$\|W_1\|^{2N} \leq \|W\|^2 + Q_{N,\delta}\big(\|W_1\|^2 + \delta\big), \tag{3.5}$$

where $Q_{N,\delta}$ is the polynomial of degree $N-1$ defined as

$$Q_{N,\delta}(x) = x(x+\delta)(x+2\delta) \cdots \big(x + (N-1)\delta\big) - x^N.$$

In order to prove this claim, we let $D_j := W_{j-1} W_{j-1}^T - W_j^T W_j$ for $j = 2, \dots, N$ and note that $\|D_j\| \leq \delta$ by assumption. Moreover,

$$\|W_j\|^2 = \big\|W_j^T W_j\big\| = \big\|W_{j-1} W_{j-1}^T - D_j\big\| \leq \|W_{j-1}\|^2 + \delta, \quad \text{for all } j = 2, \dots, N, \tag{3.6}$$

and consequently

$$\|W_j\|^2 \leq \|W_1\|^2 + (j-1)\delta \quad \text{for } j = 1, \dots, N. \tag{3.7}$$

We observe that by basic properties of the spectral norm

$$\begin{aligned}
\|W_1\|^{2N} &= \big\|\big(W_1^T W_1\big)^N\big\| = \big\|W_1^T \big(W_1 W_1^T\big)^{N-1} W_1\big\| \\
&= \big\|W_1^T \big(W_2^T W_2 + D_2\big)^{N-1} W_1\big\| \\
&\leq \big\|W_1^T \big(W_2^T W_2\big)^{N-1} W_1\big\| \\
&\quad + \sum_{k=0}^{N-2} \binom{N-1}{k} \|W_1\| \big\|W_2^T W_2\big\|^k \|D_2\|^{N-k-1} \|W_1\| \\
&\leq \big\|W_1^T \big(W_2^T W_2\big)^{N-1} W_1\big\| \\
&\quad + \|W_1\|^2 \left(\sum_{k=0}^{N-1} \binom{N-1}{k} \|W_2\|^{2k} \delta^{N-k-1} - \|W_2\|^{2(N-1)} \right) \\
&= \big\|W_1^T W_2^T \big(W_2 W_2^T\big)^{N-2} W_2 W_1\big\| \\
&\quad + \|W_1\|^2 \big(\big(\|W_2\|^2 + \delta\big)^{N-1} - \|W_2\|^{2(N-1)}\big).
\end{aligned} \tag{3.8}$$

In the first inequality, we expanded $(W_2^T W_2 + D_2)^{N-1}$ as a (matrix) polynomial in $W_2^T W_2$ and D_2, observing that the highest degree term is $(W_2^T W_2)^{N-1}$. Applying the triangle inequality separates this term from the rest of the polynomial. Applying the submultiplicativity of the spectral norm to all the summands and collecting terms (which now consist of commuting scalars, i.e., the spectral norms $\|W_1\|$, $\|W_2^T W_2\|$ and $\|D_2\|$) gives the sum in (3.8), where the index $k = N-1$ is left out as it was already taken care of in the first term in (3.8).

We continue in this way, replacing $(W_2 W_2^T)^{N-2}$ by $(W_3^T W_3 + D_3)^{N-2}$, and so on. Using also (3.7), we observe that similarly as above, for $j = 2, \dots, N-1$,

$$\begin{aligned}
&\left\| W_1^T \cdots W_j^T \left(W_j W_j^T\right)^{N-j} W_j \cdots W_1 \right\| \\
&\quad \leq \left\| W_1^T \cdots W_{j+1}^T \left(W_{j+1} W_{j+1}^T\right)^{N-j-1} W_{j+1} \cdots W_1 \right\| \\
&\qquad + \|W_j\|^2 \cdots \|W_1\|^2 \left(\left(\|W_{j+1}\|^2 + \delta\right)^{N-j} - \|W_{j+1}\|^{2(N-j)}\right) \\
&\quad \leq \left\| W_1^T \cdots W_{j+1}^T \left(W_{j+1} W_{j+1}^T\right)^{N-j-1} W_{j+1} \cdots W_1 \right\| \\
&\qquad + \|W_1\|^2 \left(\|W_1\|^2 + \delta\right) \cdots \left(\|W_1\|^2 + (j-1)\delta\right) \\
&\qquad \times \left(\left(\|W_1\|^2 + (j+1)\delta\right)^{N-j} - \left(\|W_1\|^2 + j\delta\right)^{N-j}\right) \\
&\quad \leq \left\| W_1^T \cdots W_{j+1}^T \left(W_{j+1} W_{j+1}^T\right)^{N-j-1} W_{j+1} \cdots W_1 \right\| \\
&\qquad + \left(\|W_1\|^2 + \delta\right)\left(\|W_1\|^2 + 2\delta\right) \cdots \left(\|W_1\|^2 + j\delta\right) \\
&\qquad \times \left(\left(\|W_1\|^2 + (j+1)\delta\right)^{N-j} - \left(\|W_1\|^2 + j\delta\right)^{N-j}\right).
\end{aligned}$$

Hereby, we have also used that the function $x \mapsto (x+\delta)^{N-j} - x^{N-j}$ is monotonically increasing in $x \geq 0$. With this estimate, we obtain, noting below that the sum in the second line is telescoping, that

$$\begin{aligned}
\|W_1\|^{2N} &\leq \left\| W_1^T \cdots W_N^T W_N \cdots W_1 \right\| \\
&\quad + \sum_{j=1}^{N-1} \left(\|W_1\|^2 + \delta\right)\left(\|W_1\|^2 + 2\delta\right) \cdots \left(\|W_1\|^2 + j\delta\right) \\
&\quad \times \left(\left(\|W_1\|^2 + (j+1)\delta\right)^{N-j} - \left(\|W_1\|^2 + j\delta\right)^{N-j}\right) \\
&= \|W_N \cdots W_1\|^2 + \left(\|W_1\|^2 + \delta\right)\left(\|W_1\|^2 + 2\delta\right) \cdots \left(\|W_1\|^2 + N\delta\right) \\
&\quad - \left(\|W_1\|^2 + \delta\right)^N \\
&= \|W\|^2 + Q_{N,\delta}\left(\|W_1\|^2 + \delta\right).
\end{aligned}$$

This proves claimed inequality (3.5).

The fact that for all $z, \alpha \in \mathbb{R}$ it holds $z(z+\alpha) \leq (z + \frac{\alpha}{2})^2$ implies that

$$\begin{aligned}
(x+\delta)(x+2\delta)\cdots(x+N\delta) &= \big((x+\delta)(x+N\delta)\big) \cdot (x+2\delta)\big(x+(N-1)\delta\big) \cdots \\
&\leq \left(x + \frac{N+1}{2}\delta\right)^N .
\end{aligned} \tag{3.9}$$

Setting $x = \|W_1\|^2$, $a = \|W\|^2$ and $b = \frac{N+1}{2}\delta$ and combining inequality (3.5) and the definition of $Q_{N,\delta}$ with (3.9), leads to $x^N \leq a + (x+b)^N - (x+\delta)^N$ and hence

$$x^N \leq a + (x+b)^N - x^N. \tag{3.10}$$

The mean-value theorem applied to the map $x \mapsto x^N$ gives

$$(x+b)^N = x^N + N\xi^{N-1}b \quad \text{for some } \xi \in [x, x+b].$$

Hence,

$$x^N \leq a + N\xi^{N-1}b \leq a + N(x+b)^{N-1}b.$$

We assume now that $a > 0$ and will comment on the case $a = 0$ below. Then, the previous inequality implies

$$\frac{x^N}{a} \leq 1 + N\frac{(x+b)^{N-1}b}{a},$$

which is equivalent to

$$\left(\frac{x}{a^{\frac{1}{N}}}\right)^N \leq 1 + N\left(\frac{x}{a^{\frac{1}{N}}} + \frac{b}{a^{\frac{1}{N}}}\right)^{N-1}\frac{b}{a^{\frac{1}{N}}}. \tag{3.11}$$

Setting $z = a^{-\frac{1}{N}}x$ and $c = a^{-\frac{1}{N}}b$, we obtain

$$z^N \leq 1 + Nc(z+c)^{N-1}.$$

We claim that $z \leq 1 + 2Nc$. Assume on the contrary that $z > 1 + 2Nc$. Then, (3.11) gives

$$\begin{aligned} z &\leq \frac{1}{z^{N-1}} + Nc\left(1 + \frac{c}{z}\right)^{N-1} < 1 + Nc\left(1 + \frac{c}{1+2Nc}\right)^{N-1} \leq 1 + Nc\left(1 + \frac{1/2}{N}\right)^N \\ &\leq 1 + Nce^{\frac{1}{2}}. \end{aligned}$$

The last inequality implies $z \leq 1 + 2Nc$, which is a contradiction. Thus, we showed the claim that $z \leq 1 + 2Nc$, that is, $xa^{-\frac{1}{N}} \leq 1 + 2Na^{-\frac{1}{N}}b$, which is equivalent to

$$x \leq a^{\frac{1}{N}} + 2Nb. \tag{3.12}$$

The last inequality also holds in the case $a = 0$, since for $a = 0$ inequality (3.10) remains true if we replace a by any positive number ε and then by our reasoning above $x \leq \varepsilon^{\frac{1}{N}} + 2Nb$. Since this is true for any $\varepsilon > 0$, it follows that for $a = 0$ we have $x \leq 2Nb = a^{\frac{1}{N}} + 2Nb$, thus (3.12) also holds for $a = 0$.

Using the definitions of a, b and x, we obtain from (3.12) that

$$\|W_1\|^2 \leq \|W\|^{\frac{2}{N}} + N(N+1)\delta.$$

For any $j = 1,\dots,N$, (3.7) implies then that

$$\|W_j\|^2 \le \|W_1\|^2 + (j-1)\delta \le \|W\|^{\frac{2}{N}} + N(N+1)\delta + (j-1)\delta \le \|W\|^{\frac{2}{N}} + (N+1)^2\delta.$$

This completes the proof. □

3.2 Preservation of approximate balancedness

The key ingredient to the proof of Theorem 2.4 is the following proposition. It is a highly nontrivial extension of [11, Lemma 3.1] from $N = 2$ layers to an arbitrary number of layers.

Proposition 3.4 *Assume that XX^T has full rank and $\overrightarrow{W}(0) = (W_1(0),\dots,W_N(0))$ has balancedness constant $\alpha\delta$ for some $\delta > 0$ and $\alpha \in [0,1)$. Assume that the positive stepsizes η_k satisfy* (3.14)*. Then, the gradient descent iterates $\overrightarrow{W}(k) = (W_1(k),\dots,W_N(k))$ defined by* (2.1) *satisfy, for all $k \in \mathbb{N}_0$:*

(1) *$\overrightarrow{W}(k)$ has balancedness constant δ, i.e.,*

$$\left\| W_{j+1}^T(k)W_{j+1}(k) - W_j(k)W_j^T(k)\right\| \le \delta \quad \textit{for all } j = 1,\dots,N-1; \tag{3.13}$$

(2) $L^N(\overrightarrow{W}(k)) \le L^N(\overrightarrow{W}(0))$;

(3) $\|W_j(k)\|^2 \le K_\delta = M^{\frac{2}{N}} + (N+1)^2\delta$ *for* $j = 1,\dots,N$;

(4) $L^N(\overrightarrow{W}(k)) - L^N(\overrightarrow{W}(k+1)) \ge \sigma\eta_k\|\nabla L^N(\overrightarrow{W}(k))\|_F^2$.

Proof We will show statements (1), (2), and (3) by induction under the condition that

$$\eta_k \le \min\left\{\frac{2(1-\sigma)}{B_\delta}, \frac{\sigma(1-\alpha)\delta}{2L^N(\overrightarrow{W}(0))}\right\} \quad \text{for all } k \in \mathbb{N}, \tag{3.14}$$

hold for some $\sigma \in (0,1)$. The choice

$$\sigma = \frac{4L^N(\overrightarrow{W}(0))}{4L^N(\overrightarrow{W}(0)) + (1-\alpha)\delta B_\delta}$$

reduces (3.14) to (2.8). In the induction step for (2), we will show that if (3) holds for k, then (4) also holds for k. Below, we will always denote $W(k) = W_N(k)\cdots W_1(k)$.

Since $\overrightarrow{W}(0)$ has balancedness constant $\alpha\delta < \delta$ by assumption, (3.13) is clearly satisfied for $k = 0$. Statement (2) is trivial for $k = 0$. The bound in (3) follows from a direct combination of Proposition 3.2 with Lemma 3.1, i.e., for $j = 1,\dots,N$,

$$\begin{aligned}\left\|W_j(0)\right\|^2 &\le \left\|W(0)\right\|^{\frac{2}{N}} + (N+1)^2\delta \le \left(\frac{\sqrt{2L^N(\overrightarrow{W}(0))} + \|Y\|}{\sigma_{\min}(X)}\right)^{\frac{2}{N}} + (N+1)^2\delta \\ &= M^{\frac{2}{N}} + (N+1)^2\delta,\end{aligned}$$

using also the definition of M in (2.11).

For the induction step, we assume that (1), (2), and (3) hold for $0,1,\dots,k$ and prove that these three properties also hold for $k+1$.

Step 1: We first prove statement (2) for $k+1$. To do so, we will show that if statement (3) holds for k, then statement (4) also holds for k. This also proves (4) once the induction for (1), (2), and (3) is completed.

We consider the Taylor expansion

$$\begin{aligned}L^N\big(\overrightarrow{W}(k+1)\big) &= L^N\big(\overrightarrow{W}(k)\big) + \big\langle \nabla L^N\big(\overrightarrow{W}(k)\big), \overrightarrow{W}(k+1) - \overrightarrow{W}(k)\big\rangle \\ &\quad + \frac{1}{2}\big\langle \big(\overrightarrow{W}(k+1) - \overrightarrow{W}(k)\big)^T \nabla^2 L^N(\overrightarrow{A}_\xi), \overrightarrow{W}(k+1) - \overrightarrow{W}(k)\big\rangle,\end{aligned}$$

where

$$\nabla L^N\big(\overrightarrow{W}(k)\big) = \begin{pmatrix} \nabla_{W_1} L^N(\overrightarrow{W}(k)) \\ \vdots \\ \nabla_{W_N} L^N(\overrightarrow{W}(k)) \end{pmatrix}$$

and $\overrightarrow{A}_\xi = (A_\xi^1, \dots, A_\xi^N)$ with

$$A_\xi^i = W_i(k) + \xi\big(W_i(k+1) - W_i(k)\big) \quad \text{for some } \xi \in [0,1],\ i = 1, \dots, N.$$

Since by definition $W_j(k+1) = W_j(k) - \eta_k \nabla_{W_j} L^N(\overrightarrow{W}(k))$, this Taylor expansion can be written as

$$\begin{aligned}L^N\big(\overrightarrow{W}(k+1)\big) &= L^N\big(\overrightarrow{W}(k)\big) - \eta_k \big\langle \nabla L^N\big(\overrightarrow{W}(k)\big), \nabla L^N\big(\overrightarrow{W}(k)\big)\big\rangle_F \\ &\quad + \frac{1}{2}\eta_k^2 \big\langle \nabla L^N\big(\overrightarrow{W}(k)\big), \nabla^2 L^N(\overrightarrow{A}_\xi) \nabla L^N\big(\overrightarrow{W}(k)\big)\big\rangle_F.\end{aligned}$$

By the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}&L^N\big(\overrightarrow{W}(k)\big) - L^N\big(\overrightarrow{W}(k+1)\big) \\ &\quad \geq \eta_k \big\|\nabla L^N\big(\overrightarrow{W}(k)\big)\big\|_F^2 - \frac{1}{2}\eta_k^2 \big\|\nabla L^N\big(\overrightarrow{W}(k)\big)\big\|_F^2 \big\|\nabla^2 L^N(\overrightarrow{A}_\xi)\big\|_{F\to F} \\ &\quad \geq \left(1 - \frac{1}{2}\eta_k \big\|\nabla^2 L^N(\overrightarrow{A}_\xi)\big\|_{F\to F}\right)\eta_k \big\|\nabla L^N\big(\overrightarrow{W}(k)\big)\big\|_F^2. \qquad (3.15)\end{aligned}$$

The crucial point now is to show that $\|\nabla^2 L^N(\overrightarrow{A}_\xi)\|_{F\to F}$ is bounded by the constant B_δ defined in (2.9). By setting $\overrightarrow{\Delta} = (\Delta_1, \dots, \Delta_N)$ with $\Delta_j \in \mathbb{R}^{d_j \times d_{j-1}}$, $j = 1, \dots, N$, and writing $\nabla^2 L^N(\overrightarrow{W})(\overrightarrow{\Delta}, \overrightarrow{\Delta})$ for $\langle \overrightarrow{\Delta}, \nabla^2 L^N(\overrightarrow{W}) \overrightarrow{\Delta}\rangle$, the quadratic form $\nabla^2 L^N(\overrightarrow{W})(\overrightarrow{\Delta}, \overrightarrow{\Delta})$ defined by the Hessian can be written as

$$\begin{aligned}\nabla^2 L^N(\overrightarrow{W})(\overrightarrow{\Delta}, \overrightarrow{\Delta}) &= \sum_{j=1}^N \sum_{i=1}^N \left\langle \Delta_j, \frac{\partial^2 L^N(\overrightarrow{W})}{\partial W_i \partial W_j} \Delta_i \right\rangle \\ &= \sum_{i=1}^N \left\langle \Delta_i, \frac{\partial^2 L^N(\overrightarrow{W})}{\partial W_i^2} \Delta_i \right\rangle + \sum_{j=1}^N \sum_{\substack{i=1 \\ i\neq j}}^N \left\langle \Delta_j, \frac{\partial^2 L^N(\overrightarrow{W})}{\partial W_i \partial W_j} \Delta_i \right\rangle.\end{aligned}$$

In order to compute mixed second derivatives, we introduce the notation

$$Q_i(\overrightarrow{W}, \Delta_i) = W_N \cdots W_{i+1} \Delta_i W_{i-1} \cdots W_1 X,$$

$$P_{i,j}(\overrightarrow{W}, \Delta_i, \Delta_j) = \begin{cases} W_N \cdots W_{j+1} \Delta_j W_{j-1} \cdots W_{i+1} \Delta_i W_{i-1} \cdots W_1 & \text{if } j > i, \\ W_N \cdots W_{i+1} \Delta_i W_{i-1} \cdots W_{j+1} \Delta_j W_{j-1} \cdots W_1 & \text{if } j < i, \end{cases}$$

with the understanding that $W_{i-1} \cdot W_1 = \mathrm{Id}$ for $i = 1$ and $W_N \cdots W_{i+1} = \mathrm{Id}$ for $i = N$. Using the first partial derivatives of L^N, cf. (3.2), we obtain, for $i = 1, \dots, N$,

$$\left\langle \Delta_i, \frac{\partial^2 L^N(\overrightarrow{W})}{\partial W_i^2} \Delta_i \right\rangle = \langle Q_i(\overrightarrow{W}, \Delta_i), Q_i(\overrightarrow{W}, \Delta_i) \rangle = \| Q_i(\overrightarrow{W}, \Delta_i) \|_F^2.$$

The mixed second order derivatives are given, for $i \neq j$, by

$$\left\langle \Delta_i, \frac{\partial^2 L^N(\overrightarrow{W})}{\partial W_i \partial W_j} \Delta_j \right\rangle = \langle Q_i(\overrightarrow{W}, \Delta_i), Q_j(\overrightarrow{W}, \Delta_j) \rangle + \langle L^N(\overrightarrow{W}), P_{i,j}(\overrightarrow{W}, \Delta_i, \Delta_j) \rangle.$$

This implies that

$$\begin{aligned} \nabla^2 L^N(\overrightarrow{A_\xi})(\overrightarrow{\Delta}, \overrightarrow{\Delta}) = {} & \sum_{i=1}^{N} \| Q_i(\overrightarrow{A}_\xi, \Delta_i) \|_F^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \langle Q_i(\overrightarrow{A}_\xi, \Delta_i), Q_j(\overrightarrow{A}_\xi, \Delta_j) \rangle \\ & + \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \langle A_\xi X X^T - Y X^T, P_{i,j}(\overrightarrow{A}_\xi, \Delta_i, \Delta_j) \rangle, \end{aligned}$$

where $A_\xi = A_\xi^N \cdots A_\xi^1$. The Cauchy–Schwarz inequality for the trace inner product together with $\|AB\|_F \leq \|A\| \|B\|_F$ for any matrices A, B of matching dimensions gives, for $i > j$,

$$\begin{aligned} & \left| \langle A_\xi X X^T - Y X^T, P_{i,j}(\overrightarrow{A}_\xi, \Delta_i, \Delta_j) \rangle \right| \\ & \quad = \left| \mathrm{tr}\left(\left(A_\xi X X^T - Y X^T \right)^T A_\xi^N \cdots A_\xi^{i+1} \Delta_i A_\xi^{i-1} \cdots A_\xi^{j+1} \Delta_j A_\xi^{j-1} \cdots A_\xi^1 \right) \right| \\ & \quad \leq \left\| \left(A_\xi X X^T - Y X^T \right)^T A_\xi^N \cdots A_\xi^{i+1} \Delta_i \right\|_F \left\| A_\xi^{i-1} \cdots A_\xi^{j+1} \Delta_j A_\xi^{j-1} \cdots A_\xi^1 \right\|_F \\ & \quad \leq \| A_\xi X X^T - Y X^T \| \| A_\xi^N \| \cdots \| A_\xi^{i+1} \| \|\Delta_i\|_F \| A_\xi^{i-1} \| \cdots \| A_\xi^{j+1} \| \|\Delta_j\|_F \| A_\xi^{j-1} \| \cdots \| A_\xi^1 \|, \end{aligned}$$

and similarly, for $i < j$. Another application of the Cauchy–Schwarz inequality gives

$$\left| \langle Q_i(\overrightarrow{A}_\xi, \Delta_i), Q_j(\overrightarrow{A}_\xi, \Delta_j) \rangle \right| \leq \| Q_i(\overrightarrow{A}_\xi, \Delta_i) \|_F \| Q_j(\overrightarrow{A}_\xi, \Delta_j) \|_F.$$

Consequently,

$$\left| \nabla^2 L^N(\overrightarrow{A_\xi})(\overrightarrow{\Delta}, \overrightarrow{\Delta}) \right| \leq \sum_{i,j=1}^{N} \| Q_i(\overrightarrow{A}_\xi, \Delta_i) \|_F \| Q_j(\overrightarrow{A}_\xi, \Delta_j) \|_F$$

$$
\begin{aligned}
&+\sum_{\substack{i,j=1\\ i\neq j}}^{N}\left\|A_{\xi}XX^{T}-YX^{T}\right\|\|\Delta_i\|_F\|\Delta_j\|_F\prod_{\substack{k=1\\ k\neq i,j}}^{N}\left\|A_{\xi}^{k}\right\|\\
&\leq \|X\|^{2}\sum_{i,j=1}^{N}\|\Delta_i\|_F\|\Delta_j\|_F\left(\prod_{\substack{k=1\\ k\neq i}}^{N}\left\|A_{\xi}^{k}\right\|\right)\left(\prod_{\substack{k=1\\ k\neq j}}^{N}\left\|A_{\xi}^{k}\right\|\right)\\
&+\sum_{\substack{i,j=1\\ i\neq j}}^{N}\|\Delta_i\|_F\|\Delta_j\|_F\left(\|X\|^{2}\prod_{k=1}^{N}\left\|A_{\xi}^{k}\right\|+\left\|YX^{T}\right\|\right)\prod_{\substack{k=1\\ k\neq i,j}}^{N}\left\|A_{\xi}^{k}\right\|. \qquad (3.16)
\end{aligned}
$$

Using the recursive definition of $W_i(k+1)$ and that $\xi\in[0,1]$, we further obtain, for $i=1,\dots,N$,

$$
\begin{aligned}
\left\|A_{\xi}^{i}\right\| &= \left\|W_i(k)+\xi\big(W_i(k+1)-W_i(k)\big)\right\| \leq \left\|W_i(k)\right\|+\left\|W_i(k+1)-W_i(k)\right\|\\
&= \left\|W_i(k)\right\|+\left\|\eta_k\nabla_{W_i}L^{N}\big(\overrightarrow{W}(k)\big)\right\|\\
&= \left\|W_i(k)\right\|+\eta_k\left\|W_{i+1}^{T}(k)\cdots W_{N}^{T}(k)\nabla L^{1}\big(W(k)\big)W_1(k)^{T}\cdots W_{i-1}(k)\right\|.
\end{aligned}
$$

It follows from (3.4) and induction hypothesis (2) for k that

$$
\left\|\nabla L^{1}\big(W(k)\big)\right\|\leq\sqrt{2L^{N}\big(\overrightarrow{W}(k)\big)}\|X\|\leq\sqrt{2L^{N}\big(\overrightarrow{W}(0)\big)}\|X\|. \qquad (3.17)
$$

Using induction hypothesis (3) for k this gives

$$
\begin{aligned}
\left\|A_{\xi}^{i}\right\| &\leq \left\|W_i(k)\right\|+\eta_k\sqrt{2L^{N}\big(\overrightarrow{W}(0)\big)}\|X\|\left(\prod_{\substack{j=1\\ j\neq i}}^{N}\left\|W_j(k)\right\|\right)\\
&\leq K_{\delta}^{1/2}+\eta_k\sqrt{2L^{N}\big(\overrightarrow{W}(0)\big)}\|X\|K_{\delta}^{\frac{N-1}{2}}.
\end{aligned}
$$

By assumption (3.14) on the stepsize η_k and the definitions of K_δ and B_δ, we have

$$
\begin{aligned}
&\eta_k\sqrt{2L^{N}\big(\overrightarrow{W}(0)\big)}\|X\|K_{\delta}^{\frac{N-1}{2}}\\
&\quad\leq\frac{2(1-\sigma)}{B_\delta}\sqrt{2L^{N}\big(\overrightarrow{W}(0)\big)}\|X\|K_{\delta}^{\frac{N-1}{2}}\leq\frac{2\sqrt{2L^{N}(\overrightarrow{W}(0))}\|X\|K_{\delta}^{\frac{N-1}{2}}}{2eNK_{\delta}^{N-1}\|X\|^{2}}\\
&\quad\leq\frac{M\sigma_{\min}(X)\|X\|K_{\delta}^{\frac{N-1}{2}}}{eNK_{\delta}^{N-1}\|X\|^{2}}\leq\frac{(M^{\frac{2}{N}}+N^{2}\delta)^{\frac{N}{2}}K_{\delta}^{\frac{N-1}{2}}}{eNK_{\delta}^{N-1}}\leq\frac{1}{2N}K_{\delta}^{\frac{1}{2}}.
\end{aligned}
$$

In the first inequality of the last line, we used the fact that by definition of M

$$
\sqrt{2L^{N}\big(\overrightarrow{W}(0)\big)}=M\sigma_{\min}(X)-\|Y\|\leq M\sigma_{\min}(X),
$$

and in the last inequality of the last line, we used $M^{\frac{2}{N}}+N^{2}\delta\leq M^{\frac{2}{N}}+(N+1)^{2}\delta=K_\delta$.

It follows that

$$\|A_\xi^i\| \leq \left(1 + \frac{1}{2N}\right) K_\delta^{1/2}.$$

Substituting this bound into (3.16), we obtain

$$\begin{aligned}
&\left|\nabla^2 L^N(\overrightarrow{A_\xi})(\overrightarrow{\Delta}, \overrightarrow{\Delta})\right| \\
&\quad\leq \left(1 + \frac{1}{2N}\right)^{2N-2} K_\delta^{N-1} \|X\|^2 \sum_{i,j=1}^{N} \|\Delta_j\|_F \|\Delta_i\|_F \\
&\qquad + \left(\left(1 + \frac{1}{2N}\right)^{2N-2} K_\delta^{N-1} \|X\|^2 \right. \\
&\qquad \left. + \left(1 + \frac{1}{2N}\right)^{N-2} K_\delta^{N/2-1} \left\|XY^T\right\|\right) \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i\neq j}}^{N} \|\Delta_j\|_F \|\Delta_i\|_F \\
&\quad\leq e K_\delta^{N-1} \|X\|^2 \left(\sum_{j=1}^{N} \|\Delta_j\|\right)^2 + \left(e K_\delta^{N-1} \|X\|^2 + e^{1/2} K_\delta^{N/2-1} \left\|XY^T\right\|\right) \left(\sum_{j=1}^{N} \|\Delta_j\|\right)^2 \\
&\quad\leq \left[2eN K_\delta^{N-1} \|X\|^2 + \sqrt{e} N K_\delta^{N/2-1} \left\|XY^T\right\|\right] \|\overrightarrow{\Delta}\|_F^2,
\end{aligned}$$

where we have used the fact that $(1 + 1/(2N))^{2N} \leq e$ and that $\sum_{j=1}^{N} \|\Delta_{W_j}\|_F \leq \sqrt{N} \|\overrightarrow{\Delta}\|_F$. Hence, we derived that

$$\left\|\nabla^2 L^N(\overrightarrow{A_\xi})\right\|_{F\to F} \leq 2eN K_\delta^{N-1} \|X\|^2 + \sqrt{e} N K_\delta^{\frac{N}{2}-1} \left\|XY^T\right\| = B_\delta.$$

Substituting this estimate into (3.15) and using that the stepsizes satisfy (3.14) gives

$$\begin{aligned}
L^N\left(\overrightarrow{W}(k)\right) - L^N\left(\overrightarrow{W}(k+1)\right) &\geq \left(1 - \frac{1}{2}\eta_k B_\delta\right) \eta_k \left\|\nabla L^N\left(\overrightarrow{W}(k)\right)\right\|_F^2 \\
&\geq \sigma \eta_k \left\|\nabla L^N\left(\overrightarrow{W}(k)\right)\right\|_F^2 \geq 0. \qquad (3.18)
\end{aligned}$$

This shows statement (4) for k. It follows by induction hypothesis (2) for k that

$$L^N\left(\overrightarrow{W}(0)\right) \geq L^N\left(\overrightarrow{W}(k)\right) \geq L^N\left(\overrightarrow{W}(k+1)\right).$$

This shows statement (2) for $k + 1$.

Step 2: Let us now show that statement (1) holds at iteration $k + 1$. For $j = 1, \ldots, N - 1$, we obtain

$$\begin{aligned}
&\left\| W_{j+1}^T(k+1) W_{j+1}(k+1) - W_j(k+1) W_j^T(k+1) \right\| \\
&\quad= \left\| \left(W_{j+1}(k) - \eta_k \nabla_{W_{j+1}} L^N\left(\overrightarrow{W}(k)\right)\right)^T \left(W_{j+1}(k) - \eta_k \nabla_{W_{j+1}} L^N\left(\overrightarrow{W}(k)\right)\right) \right. \\
&\qquad \left. - \left(W_j(k) - \eta_k \nabla_{W_j} L^N\left(\overrightarrow{W}(k)\right)\right)\left(W_j(k) - \eta_k \nabla_{W_j} L^N\left(\overrightarrow{W}(k)\right)\right)^T \right\|
\end{aligned}$$

$$
\begin{aligned}
&= \| W_{j+1}^T(k) W_{j+1}(k) - W_j(k) W_j^T(k) \\
&\quad + \eta_k \big(-W_{j+1}^T(k) W_{j+2}^T(k) \ldots W_N^T(k) \nabla L^1 \big(W(k) \big) W_1^T(k) \cdots W_j^T(k) \\
&\quad - W_j(k) \cdots W_1(k) \nabla^T L^1 \big(W(k) \big) W_N(k) \cdots W_{j+2}(k) W_{j+1}(k) \big) \\
&\quad + W_j(k) W_{j-1}(k) \cdots W_1(k) \nabla^T L^1 \big(W(k) \big) W_N(k) \cdots W_{j+2}(k) W_{j+1}(k) \\
&\quad + W_{j+1}^T(k) W_{j+2}^T(k) \cdots W_N^T(k) \nabla L^1 \big(W(k) \big) W_1^T(k) \cdots W_{j-1}^T(k) W_j^T(k)) \\
&\quad + \eta_k^2 \big(\nabla_{W_{j+1}}^T L^N \big(\overrightarrow{W}(k) \big) \nabla_{W_{j+1}} L^N \big(\overrightarrow{W}(k) \big) - \nabla_{W_j} L^N \big(\overrightarrow{W}(k) \big) \nabla_{W_j}^T L^N \big(\overrightarrow{W}(k) \big) \big) \| \\
&\leq \big\| W_{j+1}^T(k) W_{j+1}(k) - W_j(k) W_j^T(k) \big\| + \eta_k^2 \big(\big\| \nabla_{W_{j+1}} L^N \big(\overrightarrow{W}(k) \big) \big\|^2 + \big\| \nabla_{W_j} L^N \big(\overrightarrow{W}(k) \big) \big\|^2 \big).
\end{aligned}
$$

Applying this inequality repeatedly, we obtain

$$
\begin{aligned}
&\big\| W_{j+1}^T(k+1) W_{j+1}(k+1) - W_j(k+1) W_j^T(k+1) \big\| \\
&\quad \leq \big\| W_{j+1}^T(0) W_{j+1}(0) - W_j(0) W_j^T(0) \big\| \\
&\qquad + \sum_{\ell=0}^{k} \eta_\ell^2 \big(\big\| \nabla_{W_{j+1}} L^N \big(\overrightarrow{W}(\ell) \big) \big\|^2 + \big\| \nabla_{W_j} L^N \big(\overrightarrow{W}(\ell) \big) \big\|^2 \big) \\
&\quad \leq \alpha\delta + 2 \Big(\max_{\ell=0,\ldots,k} \eta_\ell \Big) \sum_{\ell=0}^{k} \eta_\ell \big\| \nabla L^N \big(\overrightarrow{W}(\ell) \big) \big\|_F^2,
\end{aligned}
\tag{3.19}
$$

where we have used the fact that $\overrightarrow{W}(0)$ has balancedness constant $\alpha\delta$ by assumption and that

$$
\begin{aligned}
\big\| \nabla L^N \big(\overrightarrow{W}(k) \big) \big\|_F^2 &\geq \max_{\ell=1,\ldots,N} \big\| \nabla_{W_\ell} L^N \big(\overrightarrow{W}(k) \big) \big\|^2 \\
&\geq \frac{1}{2} \big(\big\| \nabla_{W_j} L^N \big(\overrightarrow{W}(k) \big) \big\|^2 + \big\| \nabla_{W_{j+1}} L^N \big(\overrightarrow{W}(k) \big) \big\|^2 \big).
\end{aligned}
$$

Inequality (3.18) from the previous step gives

$$
\begin{aligned}
L^N \big(\overrightarrow{W}(0) \big) - L^N \big(\overrightarrow{W}(k+1) \big) &= \sum_{j=0}^{k} \big(L^N \big(\overrightarrow{W}(j) \big) - L^N \big(\overrightarrow{W}(j+1) \big) \big) \\
&\geq \sigma \sum_{j=0}^{k} \eta_k \big\| \nabla L^N \big(\overrightarrow{W}(k) \big) \big\|_F^2.
\end{aligned}
\tag{3.20}
$$

Combining inequalities (3.19) and (3.20) yields

$$
\begin{aligned}
&\big\| W_{j+1}^T(k+1) W_{j+1}(k+1) - W_j(k+1) W_j^T(k+1) \big\| \\
&\quad \leq \alpha\delta + \frac{2}{\sigma} \Big(\max_{\ell=0,\ldots,k} \eta_\ell \Big) \big(L^N \big(\overrightarrow{W}(0) \big) - L^N \big(\overrightarrow{W}(k+1) \big) \big) \\
&\quad \leq \alpha\delta + \frac{2}{\sigma} \Big(\max_{\ell=0,\ldots,k} \eta_\ell \Big) L^N \big(\overrightarrow{W}(0) \big) \leq \alpha\delta + (1-\alpha)\delta = \delta,
\end{aligned}
$$

where we have used Condition (3.14) on the stepsizes. This proves statement (1) for $k+1$.

Step 3: For the proof of statement (3) for $k+1$, we use the fact that we have already shown that (1) and (2) hold for $k+1$. It follows from Proposition 3.2 and Lemma 3.1 that

$$\left\| W_j(k+1)\right\|^2 \leq \left\| W(k+1)\right\|^{\frac{2}{N}} + (N+1)^2\delta \leq \left(\frac{\sqrt{2L^N(\overrightarrow{W}(0)) + \|Y\|}}{\sigma_{\min}(X)}\right)^{\frac{2}{N}} + (N+1)^2\delta = K_\delta.$$

This shows (3) for $k+1$ and completes the proof of the proposition. □

3.3 Convergence of gradient descent to a critical point

We will use a result from [1] to prove Theorem 2.4, which is based on the following definition.

Definition 3.5 (Strong descent conditions [1]) We say that a sequence $x_k \in \mathbb{R}^n$ satisfies the strong descent conditions (for a differentiable function $f:\mathbb{R}^n \to \mathbb{R}$) if

$$f(x_k) - f(x_{k+1}) \geq \sigma \left\|\nabla f(x_k)\right\| \|x_{k+1} - x_k\| \tag{3.21}$$

$$\text{and } f(x_{k+1}) = f(x_k) \quad \Longrightarrow \quad x_{k+1} = x_k \tag{3.22}$$

hold for some $\sigma > 0$ and for all k larger than some K.

The next theorem is essentially an extension of the Lojasiewicz theorem to discrete variants of gradient flows.

Theorem 3.6 [1, *Theorem* 3.2] *Let* $f:\mathbb{R}^n \to \mathbb{R}$ *be an analytic cost function. Let the sequence* $\{x_k\}_{k=1,2,\ldots}$ *satisfy the strong descent conditions* (*Definition* 3.5). *Then, either* $\lim_{k\to\infty}\|x_k\| = +\infty$, *or there exists a single point* $x^* \in \mathbb{R}$ *such that*

$$\lim_{k\to\infty} x_k = x^*$$

Now, we are ready to prove Theorem 2.4.

Proof By point (4) of Proposition 3.4 and since $\overrightarrow{W}(k+1) - \overrightarrow{W}(k) = \eta_k \nabla L^N(\overrightarrow{W}(k))$ for all $k \in \mathbb{N}_0$, we have

$$L^N\big(\overrightarrow{W}(k)\big) - L^N\big(\overrightarrow{W}(k+1)\big) \geq \sigma \left\|\nabla L^N\big(\overrightarrow{W}(k)\big)\right\|_F \left\|\overrightarrow{W}(k+1) - \overrightarrow{W}(k)\right\|_F, \tag{3.23}$$

which means that the first part (3.21) of the strong descent condition holds. This implies then that also the second part (3.22) of the strong descent condition holds, since if $L^N(\overrightarrow{W}(k+1)) = L^N(\overrightarrow{W}(k))$, it follows that

$$\left\|\nabla L^N\big(\overrightarrow{W}(k)\big)\right\|_F \left\|\overrightarrow{W}(k+1) - \overrightarrow{W}(k)\right\|_F = 0,$$

hence $\overrightarrow{W}(k+1) = \overrightarrow{W}(k)$ or $\nabla L^N(\overrightarrow{W}(k)) = 0$, but the latter again implies $\overrightarrow{W}(k+1) = \overrightarrow{W}(k)$. Thus, indeed $\overrightarrow{W}(k+1) = \overrightarrow{W}(k)$ if $L^N(\overrightarrow{W}(k+1)) = L^N(\overrightarrow{W}(k))$.

Since by Proposition 3.4, the sequence $(\overrightarrow{W}(k))_{k\in\mathbb{N}_0}$ is bounded and L^N is analytic, it follows from Theorem 3.6 that there exists $\overrightarrow{W}^*$ such that

$$\lim_{k\to\infty} \overrightarrow{W}(k) = \overrightarrow{W}^*.$$

It remains to show that $\overrightarrow{W}^*$ is a critical point of L^N. Since $\nabla L^N(\overrightarrow{W})$ is continuous in $\overrightarrow{W}$, it follows that $\nabla L^N(\overrightarrow{W}^*) = \lim_{k\to\infty} \nabla L^N(\overrightarrow{W}(k))$ and that

$$\left\|\nabla L^N\left(\overrightarrow{W}^*\right)\right\|_F = \lim_{k\to\infty}\left\|\nabla L^N\left(\overrightarrow{W}(k)\right)\right\|_F =: c.$$

In order to show that $\overrightarrow{W}^*$ is a critical point, it suffices to show that $c = 0$. A repeated application of point (4) of Proposition 3.4 gives

$$L^N\left(\overrightarrow{W}(0)\right) - L^N\left(\overrightarrow{W}(k+1)\right) \geq \sigma \sum_{j=0}^{k} \eta_j \left\|\nabla L^N\left(\overrightarrow{W}(j)\right)\right\|_F^2 \quad \text{for any } k \in \mathbb{N},$$

hence, taking the limit,

$$L^N\left(\overrightarrow{W}(0)\right) \geq \sigma \sum_{k=0}^{\infty} \eta_k \left\|\nabla L^N\left(\overrightarrow{W}(k)\right)\right\|_F^2.$$

Assume now that $c \neq 0$. Then, $c > 0$, and there exists $k_0 \in \mathbb{N}$ such that

$$\left\|\nabla L^N\left(\overrightarrow{W}(k)\right)\right\|_F \geq \frac{c}{2} \quad \forall k \geq k_0.$$

But then

$$L^N\left(\overrightarrow{W}(0)\right) \geq \sigma \sum_{k=k_0}^{\infty} \eta_k \left\|\nabla L^N\left(\overrightarrow{W}(k)\right)\right\|_F^2 \geq \frac{c^2}{4}\sigma \sum_{k=k_0}^{\infty} \eta_k,$$

which by $\sigma > 0$ contradicts our assumption that $\sum_{k=0}^{\infty} \eta_k = \infty$. Thus, indeed $c = 0$, and $\overrightarrow{W}^*$ is a critical point of L^N. □

4 Convergence to a global minimum for almost all initializations

Let us now transfer [5, Theorem 6.12] to our situation of the gradient descent method by showing Theorem 2.6. Our proof is based on the following abstract theorem, which basically states that gradient descent schemes avoid strict saddle points for almost all initializations. The case of constant stepsizes (condition (1)) was shown in [20, Proposition 1], while the one for stepsizes converging to zero was proven in [23, Theorem 5.1]. We call a critical point z^* of a twice continuously differentiable function f a strict saddle point if the Hessian $\nabla^2 f(z^*)$ has at least one negative eigenvalue. Intuitively, this means that there is a direction (indicated by the eigenvector corresponding to a negative eigenvalue) in which the function decreases like a square function. Such decay is fast enough in order to direct almost all trajectories away from the saddle point (and towards such directions of decrease). This intuition is made rigorous in the following theorem.

Theorem 4.1 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a twice continuously differentiable function and consider the gradient descent scheme*

$$z(k+1) = z(k) - \eta_k \nabla f\big(z(k)\big),$$

where (η_k) satisfies one of the following conditions.

(1) *The sequence (η_k) is constant, i.e., $\eta_k = \eta$ for some $\eta > 0$ for all $k \in \mathbb{N}$.*

(2) *It holds*

$$\eta_k \geq C\frac{1}{k} \quad \text{for some } C > 0 \quad \text{and} \quad \lim_{k\to\infty} \eta_k = 0.$$

Then, the set of initializations $z(0) \in \mathbb{R}^p$, such that $(z(k))_k$ converges to a strict saddle point of f, has measure zero.

Now, we are ready to prove Theorem 2.6 by exploiting the analysis of the strict saddle points of L^N that has been performed in [5], extending [18, 26].

Proof Due to definitions (2.12), (2.13) of the constants $\delta_{\mathcal{B}}$, $L_{\mathcal{B}}$, and $M_{\mathcal{B}}$ together with condition (2.14) on the stepsizes η_k, the conditions of Theorem 2.4 are satisfied for each initialization $\overrightarrow{W}(0) \in \mathcal{B}$. Hence, $\overrightarrow{W}(k)$ converges to a critical point of L^N for all $\overrightarrow{W}(0) \in \mathcal{B}$. By Theorem 4.1, the convergence of gradient descent with initial values in $\mathcal{B}$ and with stepsizes η_k to a strict saddle point occurs only for a subset of $\mathcal{B}$ that has measure zero.

The rest of the proof is the same as the corresponding reasoning in the proof of [5, Theorem 6.12]. Let us repeat only the main aspects from [5]. Recall that $q = \operatorname{rank}(Q)$ (cf. (2.6)), $r = \min_{j=0,\dots,N} d_j$ and denote by $\overrightarrow{W} = (W_1, \dots, W_N)$ the limit of $\overrightarrow{W}(l)$, $W = W_N \cdots W_1$ and $k = \operatorname{rank}(W)$. Then, $k \leq r$, and W is a critical point of L^1 restricted to manifold $\mathcal{M}_k$ of rank k matrices [5, Proposition 6.8(a)]. Then, [5, Proposition 6.6(1)] implies that $k \leq q$. If W is not a global minimizer of L^1 restricted to $\mathcal{M}_k$, then W is a strict saddle point of L^N by [5, Proposition 6.9]. As argued above, the set of initializations converging to such a point has measure zero, showing part (a). (Note that for $N \geq 3$ and $k < \min\{r, q\}$ a global minimizer of L^1 restricted to $\mathcal{M}_k$ may correspond to a non-strict saddle point $\overrightarrow{W}$ of L^N, see [5, Proposition 6.10].) If $N = 2$, then by [5, Proposition 6.11] any critical point $\overrightarrow{W} = (W_1, W_2)$ such that $W = W_2 W_1$ is a global minimum of L^1 restricted to $\mathcal{M}_k$ for some $k < \bar{r}$ is a strict saddle point of L^2, which shows part (b) of the theorem. □

5 Numerical experiments

In this section, we illustrate our theoretical results with numerical experiments. In particular, we test convergence of gradient descent for various choices of constant and decreasing stepsizes and with $N = 2$, $N = 3$ and $N = 5$ layers.

The sample size is chosen as $m = 3 \cdot d$ with $d = 70$. For our experiments, we generate our dataset $X \in \mathbb{R}^{d_x \times m}$ randomly with entries drawn from a mean zero Gaussian distribution with variance $\sigma^2 = 1/d$, where $d_x = d$. The data matrix $Y \in \mathbb{R}^{d \times m}$ is a random matrix of rank $r = 2$, which is generated as described below. We initialize the weight matrices $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ such that $\overrightarrow{W}(0) = (W_1, \dots, W_N)$ is balanced, i.e., has balancedness constant 0 so that $\alpha = 0$ in Theorems 2.4 and 2.6, in the following way. The rank parameter is chosen as $r = 2$ and the dimensions d_j as

$$d_0 = d, \qquad d_1 = r, \qquad d_j = \operatorname{round}\left(r + (j-1)\frac{d-r}{N-1}\right), \quad j = 2,\dots,N,$$

where round(z) rounds a real number z to the nearest integer. We randomly generate orthogonal matrices $U_1 \in \mathbb{R}^{d\times d}$, $V_j \in \mathbb{R}^{d_j\times d_j}$, $j = 1,\dots,N$, according to the uniform distribution on the corresponding unitary groups and let $U_j \in \mathbb{R}^{d_j\times d_1}$, $j = 2,\dots,N$ be the matrix composed of the first d_1 columns of V_{j-1}. We then set

$$W_j = V_j I_{d_j,d_1} U_j^T,$$

where for any $n_1, n_2 \in \mathbb{N}$ the matrix $I_{n_1,n_2} \in \mathbb{R}^{n_1\times n_2}$ is a rectangular diagonal matrix with ones on the diagonal. By orthogonality and construction of U_{j+1}, it follows that for all $j = 1,\dots,N-1$, we have

$$\begin{aligned} W_{j+1}^T W_{j+1} &= U_{j+1} I_{d_1,d_{j+1}} V_{j+1}^T V_{j+1} I_{d_{j+1},d_1} U_{j+1}^T = U_{j+1} U_{j+1}^T = V_j I_{d_j,d_1} U_j^T U_j I_{d_1,d_j} V_j^T \\ &= W_j W_j^T \end{aligned}$$

so that the tuple $(W_1,\dots,W_N)$ is balanced. The random matrix $Y \in \mathbb{R}^{d\times m}$ of rank 2 is generated as $Y = \widetilde{W_N}\cdots\widetilde{W_1}X$ with matrices $\widetilde{W_j}$ generated in the same way as the matrices W_j. We decided to choose a matrix Y of rank 2 so that the global minimizer of L^1 is also of rank 2 and convergence to it means that L^N converges to zero, which is simple to check.

In our first set of experiments, we use a constant stepsize, i.e., $\eta_k = \eta$. Using $\alpha = 0$, the sufficient condition in Theorem 2.4 reads

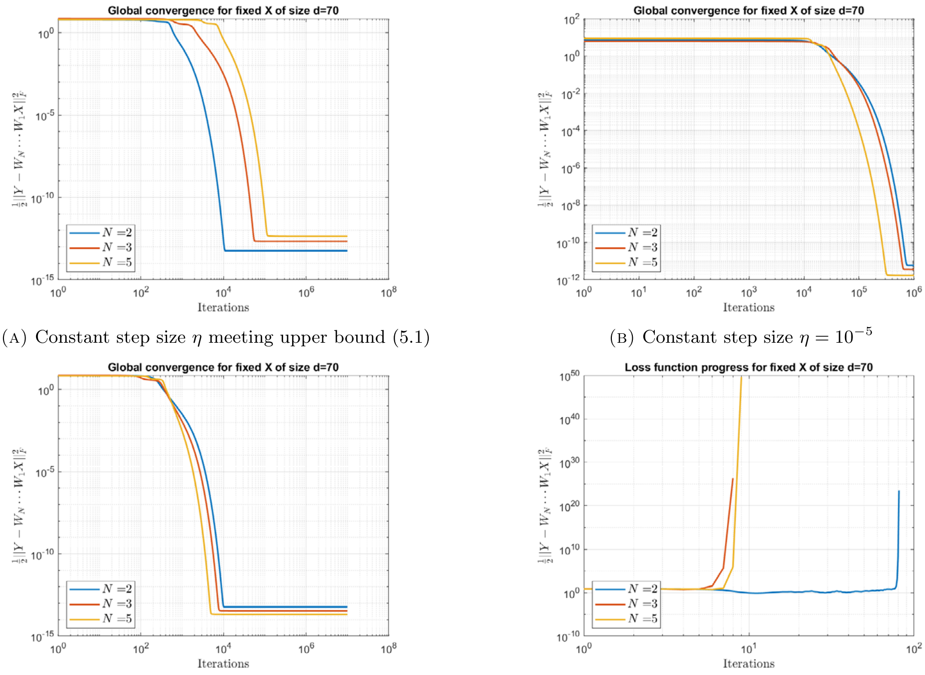
$$\eta \leq \frac{2\delta}{4L^N(\overrightarrow{W}(0)) + \delta B_\delta}, \tag{5.1}$$

with B_δ in (2.9). We choose

$$\delta = \frac{M^{\frac{2}{N}}}{N^3}.$$

This slightly differs from the choice of δ suggested by Remark 2.5(b), but corresponds to the choice of δ that we would obtain at this point using the bound given by Remark 3.3 (instead of Proposition 3.2) allowing us to set $K_\delta = M^{\frac{2}{N}} + N^2$ (instead of $K_\delta = M^{\frac{2}{N}} + (N+1)^2$) in our results.

In Fig. 1, $L^N(\overrightarrow{W}(k))$ is plotted versus the iteration number. For the plot 1a, the stepsize is chosen to exactly meet the upper bound in (5.1) (with $\delta = M^{2/N}/N^3$), resulting for this experiment in the values $\eta = 7.73\cdot 10^{-4}$, $\eta = 1.29\cdot 10^{-4}$ and $\eta = 3.91\cdot 10^{-5}$ for depth 2, 3 and 5, respectively. For the plot 1b, the stepsize η is chosen somewhat smaller than the upper bound in (5.1), while for plots 1c and 1d the bound (5.1) is not satisfied. Since we observe convergence in plot 1c, this suggests that the bound of Theorem 2.4 may not be entirely sharp. However, increasing the stepsize beyond a certain value leads to divergence as suggested by plot 2d so that some bound on the stepsize is necessary (see also [8, Lemma A.1] for a necessary condition in a special case).



(A) Constant step size η meeting upper bound (5.1)

(B) Constant step size $\eta = 10^{-5}$

(C) Constant step size $\eta = 10^{-3}$

(D) Constant step size $\eta = 0.05$

Figure 1 Progress of loss function L^N for training linear networks via gradient descent for various values of the constant stepsize

In our second set of experiments, we use a sequence of stepsizes η_k that converges to zero at various speeds. For some decay rate $\gamma \geq 0$ and some constants a_1, a_2, we set

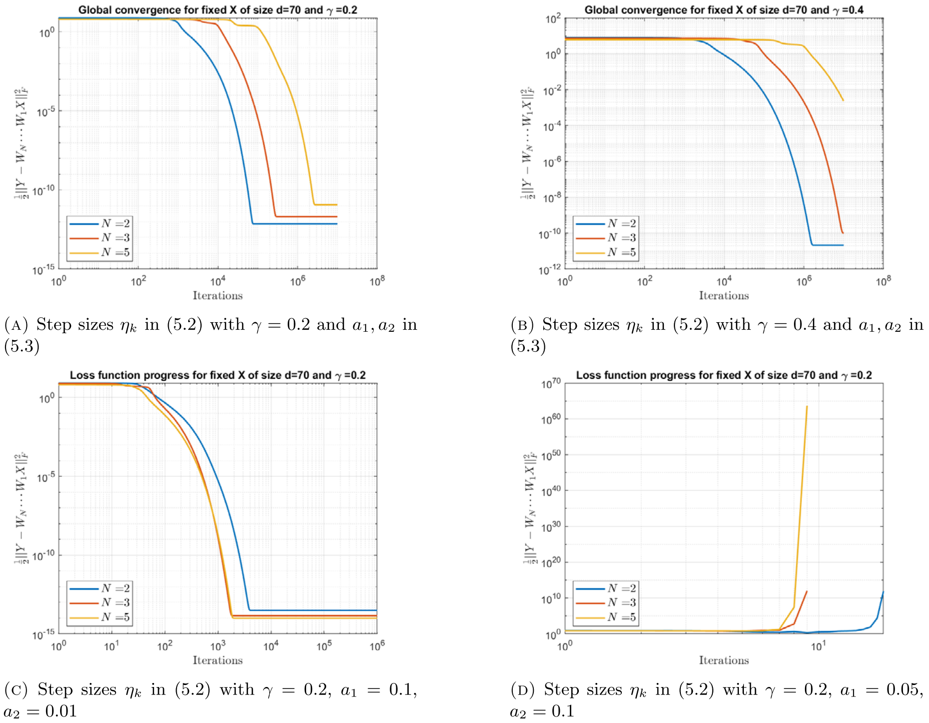
$$\eta_k = \min\left\{a_1, \frac{a_2}{(k+1)^\gamma}\right\}\gamma \geq 0, \quad \text{for all } k \in \mathbb{N}. \tag{5.2}$$

The upper bound of Theorem 2.4 is satisfied for (see also the beginning of the proof of Proposition 3.4)

$$a_1 = a_2 = \frac{2(1-\sigma)}{B_\delta}, \qquad \sigma = \frac{4L^N(\overrightarrow{W}(0))}{4L^N(\overrightarrow{W}(0)) + \delta B_\delta}. \tag{5.3}$$

Again, we choose $\delta = \frac{1}{N^3} M^{\frac{2}{N}}$, which corresponds to the choice of δ using the bound given in Remark 3.3 when testing with these values for a_1 and a_2.

The plots in Fig. 2 illustrate the convergence behavior for various choices of the constants a_1, a_2 and decay rate γ in (5.2), for $N = 2, 3, 5$. Plot 2a and 2b show convergence for the choices a_1, a_2 in (5.2) and for $\gamma = 0.2$ and $\gamma = 0.4$, respectively, leading to stepsizes satisfying the condition of Theorem 2.4. In these experiments, the resulting values of $a_1 = a_2$ are $a_1 = 7.73 \cdot 10^{-4}$ for $N = 2$, $a_1 = 1.29 \cdot 10^{-4}$ for $N = 3$ and $a_1 = 3.91 \cdot 10^{-5}$ for $N = 5$. Comparing the two plots, as well as with the plots for constant stepsize in Fig. 1, shows that fast decay of the step size leads to slower convergence of gradient descent, as expected. Note that we observe that larger values of γ are possible but will further slow down convergence, so we decided to omit the corresponding experiments here.



(A) Step sizes η_k in (5.2) with $\gamma = 0.2$ and a_1, a_2 in (5.3)

(B) Step sizes η_k in (5.2) with $\gamma = 0.4$ and a_1, a_2 in (5.3)

(C) Step sizes η_k in (5.2) with $\gamma = 0.2$, $a_1 = 0.1$, $a_2 = 0.01$

(D) Step sizes η_k in (5.2) with $\gamma = 0.2$, $a_1 = 0.05$, $a_2 = 0.1$

Figure 2 Gradient descent with decreasing stepsizes η_k as in (5.2)

Plot 2c shows convergence for a decay rate of $y = 0.2$ even though the constants a_1 and a_2 are such that η_k does not satisfy the bound of Theorem 2.4 for all k, while further increasing the value of a_2 leads to divergence as illustrated in Plot 2d.

6 Conclusion

In this article, we analysed convergence properties of GD for learning linear neural networks. We established the boundedness of GD iterates and proved its convergence to a critical point of the square loss under suitable conditions on the stepsizes. We then extended the convergence results towards a global minimum in [5] from gradient flow to gradient descent. Our work provides precise conditions that ensure convergence for both constant and decreasing stepsizes. Moreover, our maximal allowed stepsize does not vanish exponentially with the number of layers, and we also showed numerically that violating the bound for our stepsizes may result in divergence. We believe that our findings will contribute to the analysis of nonlinear neural networks. Extending the insights of this study from gradient descent to stochastic gradient descent is reserved for future work.

Acknowledgements
All authors acknowledge funding by DAAD (German Foreign Exchange Service) through the project *Understanding stochastic gradient descent in deep learning* (grant no: 57417829).

Funding
This work was supported by DAAD (German Foreign Exchange Service) through the project *Understanding stochastic gradient descent in deep learning* (grant no: 57417829).

Data availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author contributions
All authors read and approved the final manuscript.

Author details
[1]Department of Mathematics, Ludwig-Maximilians-Universität München, Theresienstr. 39, 80333, München, Germany.
[2]Chair for Mathematics of Information Processing, RWTH Aachen University, Pontdriesch 10, 52062, Aachen, Germany.
[3]Munich Center for Machine Learning, Ludwig-Maximilians-Universität München, München, Germany.

Received: 19 September 2022 Accepted: 19 December 2023 Published online: 18 July 2024

References
1. Absil, P.-A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM J. Optim. **16**(2), 531–547 (2005)
2. Arora, S., Cohen, N., Golowich, N., Hu, W.: A convergence analysis of gradient descent for deep linear neural networks. In: International Conference on Learning Representations (2019)
3. Arora, S., Cohen, N., Hazan, E.: On the optimization of deep networks: implicit acceleration by overparameterization. In: International Conference on Machine Learning (2018)
4. Arora, S., Cohen, N., Hu, W., Luo, Y.: Implicit regularization in deep matrix factorization. In: Advances in Neural Information Processing Systems, pp. 7413–7424 (2019)
5. Bah, B., Rauhut, H., Terstiege, U., Westdickenberg, M.: Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. Inf. Inference **11**(1), 307–353, (2022).
6. Bartlett, P., Helmbold, D., Long, P.: Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In: International Conference on Machine Learning, pp. 521–530. PMLR (2018)
7. Chitour, Y., Liao, Z., Couillet, R.: A geometric approach of gradient descent algorithms in neural networks (2018). Preprint. arXiv:1811.03568
8. Chou, H., Gieshoff, C., Maly, J., Rauhut, H.: Gradient descent for deep matrix factorization: dynamics and implicit bias towards low rank (2020). Preprint. arXiv:2011.13772
9. Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J.D.: Stochastic subgradient method converges on tame functions. Found. Comput. Math. **20**(1), 119–154 (2020)
10. Du, S., Hu, W.: Width provably matters in optimization for deep linear neural networks. In: International Conference on Machine Learning, pp. 1655–1664. PMLR (2019)
11. Du, S.S., Hu, W., Lee, J.D.: Algorithmic regularization in learning deep homogeneous models: layers are automatically balanced. In: ICML 2018 Workshop on Nonconvex Optimization (2018)
12. Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient descent provably optimizes over-parameterized neural networks. In: International Conference on Learning Representations (2019)
13. Elkabetz, O., Cohen, N.: Continuous vs. discrete optimization of deep neural networks. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
14. Geyer, K., Kyrillidis, A., Kalev, A.: Low-rank regularization and solution uniqueness in over-parameterized matrix sensing. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, pp. 930–940 (2020)
15. Gunasekar, S., Woodworth, B.E., Bhojanapalli, S., Neyshabur, B., Srebro, N.: Implicit regularization in matrix factorization. In: Advances in Neural Information Processing Systems, pp. 6151–6159 (2017)
16. Hu, W., Xiao, L., Pennington, J.: Provable benefit of orthogonal initialization in optimizing deep linear networks (2020). arXiv preprint. arXiv:2001.05992
17. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 795–811. Springer, Berlin (2016)
18. Kawaguchi, K.: Deep learning without poor local minima. Adv. Neural Inf. Process. Syst. **29**, 586–594 (2016)
19. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: generalization gap and sharp minima. In: International Conference on Learning Representations (2017)
20. Lee, J.D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M.I., Recht, B.: First-order methods almost always avoid strict saddle points. Math. Program. **176**(1), 311–337 (2019)
21. Mertikopoulos, P., Hallak, N., Kavis, A., Cevher, V.: On the almost sure convergence of stochastic gradient descent in non-convex problems. Adv. Neural Inf. Process. Syst. **33**, 1117–1128 (2020)
22. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Springer, Berlin (2014)
23. Panageas, I., Piliouras, G., Wang, X.: First-order methods almost always avoid saddle points: the case of vanishing step-sizes. In: Conference on Neural Information Processing Systems (2019)
24. Razin, N., Cohen, N.: Implicit regularization in deep learning may not be explainable by norms. In: Conference on Neural Information Processing Systems (2020)
25. Shamir, O.: Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In: Conference on Learning Theory, pp. 2691–2713. PMLR (2019)

26. Trager, M., Kohn, K., Bruna, J.: Pure and spurious critical points: a geometric study of linear networks. In: International Conference on Learning Representations (2020)
27. Wu, L., Wang, Q., Ma, C.: Global convergence of gradient descent for deep linear residual networks (2019). arXiv preprint. arXiv:1911.00645
28. Yun, C., Krishnan, S., Mobahi, H.: A unifying view on implicit bias in training linear neural networks. In: International Conference on Learning Representations (2021)
29. Yun, C., Sra, S., Jadbabaie, A.: Global optimality conditions for deep neural networks. In: International Conference on Learning Representations (2018)
30. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: International Conference on Learning Representations (2017)
31. Zou, D., Long, P.M., Gu, Q.: On the global convergence of training deep linear resnets. In: International Conference on Learning Representations (2020)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.