Journal of Mathematical Imaging and Vision
https://doi.org/10.1007/s10851-019-00889-w

The Global Optimization Geometry of Shallow Linear Neural Networks

Zhihui Zhu[1] · Daniel Soudry[2] · Yonina C. Eldar[3] · Michael B. Wakin[4]

Received: 3 November 2018 / Accepted: 21 May 2019

Abstract

We examine the squared error loss landscape of shallow linear neural networks. We show—with significantly milder assumptions than previous works—that the corresponding optimization problems have benign geometric properties: There are no spurious local minima, and the Hessian at every saddle point has at least one negative eigenvalue. This means that at every saddle point there is a directional negative curvature which algorithms can utilize to further decrease the objective value. These geometric properties imply that many local search algorithms (such as the gradient descent which is widely utilized for training neural networks) can provably solve the training problem with global convergence.

Keywords Deep learning · Linear neural network · Optimization geometry · Strict saddle · Spurious local minima

1 Introduction

A neural network consists of a sequence of operations (a.k.a. layers), each of which performs a linear transformation of its input, followed by a point-wise activation function, such as a sigmoid function or the rectified linear unit (ReLU) [38]. Deep artificial neural networks (i.e., deep learning) have recently led to the state-of-the-art empirical performance in many areas including computer vision, machine learning, and signal processing [5,17,19,22,28,29,32].

One crucial property of neural networks is their ability to approximate nonlinear functions. It has been shown that even a shallow neural network (i.e., a network with only one hidden layer) with a point-wise activation function has the universal approximation ability [10,17]. In particular, a shallow network with a sufficient number of activations (a.k.a. neurons) can approximate continuous functions on compact subsets of $\mathbb{R}^{d_0}$ with any desired accuracy, where d_0 is the dimension of the input data.

However, the universal approximation theory does not guarantee the algorithmic learnability of those parameters which correspond to the linear transformation of the layers. Neural networks may be trained (or learned) in an unsupervised manner, a semi-supervised manner, or a supervised manner which is by far the most common scenario. With supervised learning, the neural networks are trained by minimizing a loss function in terms of the parameters to be optimized and the training examples that consist of both input objects and the corresponding outputs. A popular approach for optimizing or tuning the parameters is gradient descent with the backpropagation method efficiently computing the gradient [44].

Although gradient descent and its variants work surprisingly well for training neural networks in practice, it remains an active research area to fully understand the theoretical underpinnings of this phenomenon. In general, the training optimization problems are nonconvex and it has been shown

ZZ and MBW were supported by NSF Grant CCF–1409261, NSF CAREER Grant CCF–1149225, and the DARPA Lagrange Program under ONR/SPAWAR contract N660011824020. DS was supported by the Israel Science Foundation (Grant No. 31/1031) and by the Taub Foundation.

✉ Zhihui Zhu
zzhu29@jhu.edu

Daniel Soudry
daniel.soudry@technion.ac.il

Yonina C. Eldar
yonina.eldar@weizmann.ac.il

Michael B. Wakin
mwakin@mines.edu

1 Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, USA

2 Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel

3 Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel

4 Department of Electrical Engineering, Colorado School of Mines, Golden, USA

Published online: 31 May 2019

that even training a simple neural network is NP-complete in general [4]. There is a large and rapidly increasing literature on the optimization theory of neural networks, surveying all of which is well outside our scope. Thus, we only briefly survey the works most relevant to ours.

In seeking to better understand the optimization problems in training neural networks, one line of research attempts to analyze their *geometric landscape*. The geometric landscape of an objective function relates to questions concerning the existence of spurious local minima and the existence of negative eigenvalues of the Hessian at saddle points. If the corresponding problem has no spurious local minima and all the saddle points are *strict* (i.e., the Hessian at any saddle point has a negative eigenvalue), then a number of local search algorithms [12,18,23,24] are guaranteed to find globally minimal solutions. Baldi and Hornik [2] showed that there are no spurious local minima in training shallow linear neural networks but did not address the geometric landscape around saddle points. Kawaguchi [20] further extended the analysis in [2] and showed that the loss function for training a general linear neural network has no spurious local minima and satisfies the strict saddle property (see Definition 4 in Sect. 2) for shallow neural networks under certain conditions. Kawaguchi also proved that for general deeper networks, there exist saddle points at which the Hessian is positive semi-definite (PSD), i.e., does not have any negative eigenvalues.

With respect to nonlinear neural networks, it was shown that there are no spurious local minima for a network with one ReLU node [11,43]. However, it has also been proved that there do exist spurious local minima in the population loss of shallow neural networks with even a small number (greater than one) of ReLU activation functions [37]. Fortunately, the number of spurious local minima can be significantly reduced with an over-parameterization scheme [37]. Soudry and Hoffer [40] proved that the number of sub-optimal local minima is negligible compared to the volume of global minima for multilayer neural networks when the number of training samples N goes to infinity and the number of parameters is close to N. Haeffele and Vidal [15] provided sufficient conditions to guarantee that certain local minima (having an all-zero slice) are also global minima. The training loss of multilayer neural networks at differentiable local minima was examined in [39]. Yun et al. [45] very recently provided sufficient and necessary conditions to guarantee that certain critical points are also global minima.

A second line of research attempts to understand the reason that local search *algorithms* efficiently find a local minimum. Aside from standard Newton-like methods such as cubic regularization [33] and the trust region algorithm [6], recent work [12,18,23,24] has shown that first-order methods also efficiently avoid strict saddles. It has been shown in [23,24] that a set of first-order local search techniques (such as gradient descent) with random initialization almost surely avoid strict saddles. Noisy gradient descent [12] and a variant called perturbed gradient descent [18] have been proven to efficiently avoid strict saddles from any initialization. Other types of algorithms utilizing second-order (Hessian) information [1,7,9] can also efficiently find approximate local minima.

To guarantee that gradient descent-type algorithms (which are widely adopted in training neural networks) converge to the global solution, the behavior of the saddle points of the objective functions in training neural networks is as important as the behavior of local minima.[1] However, the former has rarely been investigated compared to the latter, even for shallow linear networks. It has been shown in [2,20,21,31] that the objective function in training shallow linear networks has no spurious local minima under certain conditions. The behavior of saddle points is considered in [20], where the strict saddle property is proved for the case where both the input objects $\boldsymbol{X} \in \mathbb{R}^{d_0 \times N}$ and the corresponding outputs $\boldsymbol{Y} \in \mathbb{R}^{d_2 \times N}$ of the training samples have full row rank, $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ has distinct eigenvalues, and $d_2 \le d_0$. While the assumption on $\boldsymbol{X}$ can be easily satisfied, the assumption involving $\boldsymbol{Y}$ implicity adds constraints on the true weights. Consider a simple case where $\boldsymbol{Y} = \boldsymbol{W}_2^\star \boldsymbol{W}_1^\star \boldsymbol{X}$, with $\boldsymbol{W}_2^\star$ and $\boldsymbol{W}_1^\star$ the underlying weights to be learned. Then, the full-rank assumption on $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{W}_2^\star \boldsymbol{W}_1^\star \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}_1^{\star\mathrm{T}} \boldsymbol{W}_2^{\star\mathrm{T}}$ at least requires $\min(d_0, d_1) \ge d_2$ and $\mathrm{rank}(\boldsymbol{W}_2^\star \boldsymbol{W}_1^\star) \ge d_2$. Recently, the strict saddle property was also shown to hold without the above conditions on $\boldsymbol{X}$, $\boldsymbol{Y}$, d_0, and d_2, but only for *degenerate* critical points, specifically those points where $\mathrm{rank}(\boldsymbol{W}_2 \boldsymbol{W}_1) < \min\{d_2, d_1, d_0\}$ [34, Theorem 8].

In this paper, we analyze the optimization geometry of the loss function in training shallow linear neural networks. In doing so, we first characterize the behavior of *all* critical points of the corresponding optimization problems with an additional regularizer [see (2)], but without requiring the conditions used in [20] except the one on the input data $\boldsymbol{X}$. In particular, we examine the loss function for training a shallow linear neural network with an additional regularizer and show that it has no spurious local minima and obeys the strict saddle property if the input $\boldsymbol{X}$ has full row rank. This benign geometry ensures that a number of local search algorithms—including gradient descent—converge to a global minimum when training a shallow linear neural network with the proposed regularizer. We note that the additional regularizer [in (2)] is utilized to shrink the set of critical points and has no effect on the global minimum of the original problem. We also observe from experiments that this additional

[1] From an optimization perspective, non-strict saddle points and local minima have similar first-/second-order information and it is hard for first-/second-order methods (like gradient descent) to distinguish between them.

Table 1 Comparison of different results on characterizing the geometric landscape of the objective function in training a shallow linear network [see (1)]. Here, ? means this point is not discussed and ✓* indicates the result covers degenerate critical points only.

Result	Regularizer	Condition	No spurious local minima	Strict saddle property
[2, Fact 4]	No	$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ are of full row rank, $d_2 \leq d_0$, $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ has d_2 distinct eigenvalues	✓	?
[20, Theorem 2.3]	No	$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ are of full row rank, $d_2 \leq d_0$, $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ has d_2 distinct eigenvalues	✓	✓
[31, Theorem 2.1]	No	$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ are of full row rank	✓	?
[21, Theorem 1]	No	$d_1 \geq \min(d_0, d_2)$	✓	?
[34, Theorem 8]	No	No	✓	✓*
[47, Theorem 3]	$\\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{W}_1^{\mathrm{T}}\\|_F^2$	$d_1 \leq \min(d_0, d_2)$ and conditions (4) and (5)	✓	✓
Theorem 2	$\\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}\\|_F^2$	$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ is of full row rank	✓	✓
Theorem 3	No	$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ is of full row rank	✓	✓

regularizer speeds up the convergence of iterative algorithms in certain cases. Building on our study of the regularized problem and on [34, Theorem 8], we then show that these benign geometric properties are preserved even *without* the additional regularizer under the same assumption on the input data. Table 1 summarizes our main result and those of related works on characterizing the geometric landscape of the loss function in training shallow linear neural networks.

Outside of the context of neural networks, such geometric analysis (characterizing the behavior of all critical points) has been recognized as a powerful tool for understanding nonconvex optimization problems in applications such as phase retrieval [36,41], dictionary learning [42], tensor factorization [12], phase synchronization [30], and low-rank matrix optimization [3,13,14,25,26,35,46,47]. A similar regularizer [see (6)] to the one used in (2) is also utilized in [13,25,35,46,47] for analyzing the optimization geometry.

The outline of this paper is as follows. Section 2 contains the formal definitions for strict saddles and the strict saddle property. Section 3 presents our main result on the geometric properties for training shallow linear neural networks. The proof of our main result is given in Sect. 4.

2 Preliminaries

2.1 Notation

We use the symbols $\mathbf{I}$ and $\mathbf{0}$ to, respectively, represent the identity matrix and zero matrix with appropriate sizes. We denote the set of $r \times r$ orthonormal matrices by $\mathcal{O}_r := \{\boldsymbol{R} \in \mathbb{R}^{r\times r} : \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \mathbf{I}\}$. If a function $g(\boldsymbol{W}_1, \boldsymbol{W}_2)$ has two arguments, $\boldsymbol{W}_1 \in \mathbb{R}^{d_1\times d_0}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_2\times d_1}$, then we occasionally use the notation $g(\boldsymbol{Z})$ where we stack these two matrices into a larger one via $\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{W}_2 \\ \boldsymbol{W}_1^{\mathrm{T}}\end{bmatrix}$. For a scalar function $h(\boldsymbol{W})$ with a matrix variable $\boldsymbol{W} \in \mathbb{R}^{d_2\times d_0}$, its gradient is a $d_2 \times d_0$ matrix whose (i, j)th entry is $[\nabla h(\boldsymbol{W})]_{ij} = \frac{\partial h(\boldsymbol{W})}{\partial W_{ij}}$ for all $i \in [d_2], j \in [d_0]$. Here, $[d_2] = \{1, 2, \ldots, d_2\}$ for any $d_2 \in \mathbb{N}$ and W_{ij} is the (i, j)th entry of the matrix $\boldsymbol{W}$. Throughout the paper, the Hessian of $h(\boldsymbol{W})$ is represented by a bilinear form defined via $[\nabla^2 h(\boldsymbol{W})](\boldsymbol{A}, \boldsymbol{B}) = \sum_{i,j,k,l} \frac{\partial^2 h(\boldsymbol{W})}{\partial W_{ij}\partial W_{kl}} A_{ij} B_{kl}$ for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d_2\times d_0}$. Finally, we use $\lambda_{\min}(\cdot)$ to denote the smallest eigenvalue of a matrix.

2.2 Strict Saddle Property

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable objective function. The notions of critical points, strict saddles, and the strict saddle property are formally defined as follows.

Definition 1 (*Critical points*) $\boldsymbol{x}$ is called a critical point of h if the gradient at $\boldsymbol{x}$ vanishes, i.e., $\nabla h(\boldsymbol{x}) = \mathbf{0}$.

Definition 2 (*Strict saddles* [12]) We say a critical point $\boldsymbol{x}$ is a strict saddle if the Hessian evaluated at this point has at least one strictly negative eigenvalue, i.e., $\lambda_{\min}(\nabla^2 h(\boldsymbol{x})) < 0$.

In words, for a strict saddle, also called a ridable saddle [42], its Hessian has at least one negative eigenvalue which implies that there is a directional negative curvature that algorithms can utilize to further decrease the objective value. This property ensures that many local search algorithms can escape strict saddles by either directly exploiting the negative curvature [9] or adding noise which serves as a surrogate of the negative curvature [12,18]. On the other hand, when a saddle point has a Hessian that is positive semi-definite (PSD), it is difficult for first- and second-order methods to avoid converging to such a point. In other words, local search algorithms require exploiting higher-order (at least third-order) information in order to escape from a critical point that is neither a local minimum nor a strict saddle. We note that any local maxima are, by definition, strict saddles.

The following strict saddle property defines a set of nonconvex functions that can be efficiently minimized by a number of iterative algorithms with guaranteed convergence.

Definition 3 (*Strict saddle property* [12]) A twice differentiable function satisfies the strict saddle property if each critical point either corresponds to a local minimum or is a strict saddle.

Intuitively, the strict saddle property requires a function to have a negative curvature direction—which can be exploited by a number of iterative algorithms such as noisy gradient descent [12] and the trust region method [8] to further decrease the function value—at all critical points except for local minima.

Theorem 1 [12,24,33,42] *For a twice continuously differentiable objective function satisfying the strict saddle property, a number of iterative optimization algorithms can find a local minimum. In particular, for such functions,*

- *Gradient descent almost surely converges to a local minimum with a random initialization* [12]*;*
- *Noisy gradient descent* [12] *finds a local minimum with high probability and any initialization; and*
- *Newton-like methods such as cubic regularization* [33] *converge to a local minimum with any initialization.*

Theorem 1 ensures that many local search algorithms can be utilized to find a local minimum for strict saddle functions (i.e., ones obeying the strict saddle property). This is the main reason that significant effort has been devoted to establishing the strict saddle property for different problems [20,25,35,36,41,46].

In our analysis, we further characterize local minima as follows.

Definition 4 (*Spurious local minima*) We say a critical point $\boldsymbol{x}$ is a spurious local minimum if it is a local minimum but not a global minimum.

In other words, we separate the set of local minima into two categories: the global minima and the spurious local minima which are not global minima. Note that most local search algorithms are only guaranteed to find a local minimum, which is not necessarily a global one. Thus, to ensure the local search algorithms listed in Theorem 1 find a global minimum, in addition to the strict saddle property, the objective function is also required to have no spurious local minima.

In summary, the geometric landscape of an objective function relates to questions concerning the existence of spurious local minima and the strict saddle property. In particular, if the function has no spurious local minima and obeys the strict saddle property, then a number of iterative algorithms such as the ones listed in Theorem 1 converge to a global minimum. Our goal in the next section is to show that the objective function in training a shallow linear network with a regularizer satisfies these conditions.

3 Global Optimality in Shallow Linear Networks

In this paper, we consider the following optimization problem concerning the training of a shallow linear network:

$$\min_{\substack{\boldsymbol{W}_1\in\mathbb{R}^{d_1\times d_0}\\ \boldsymbol{W}_2\in\mathbb{R}^{d_2\times d_1}}} f(\boldsymbol{W}_1,\boldsymbol{W}_2)=\frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}-\boldsymbol{Y}\|_F^2, \tag{1}$$

where $\boldsymbol{X}\in\mathbb{R}^{d_0\times N}$ and $\boldsymbol{Y}\in\mathbb{R}^{d_2\times N}$ are the input and output training examples, and $\boldsymbol{W}_1\in\mathbb{R}^{d_1\times d_0}$ and $\boldsymbol{W}_2\in\mathbb{R}^{d_2\times d_1}$ are the model parameters (or weights) corresponding to the first and second layers, respectively. Throughout, we call d_0, d_1, and d_2 the sizes of the input layer, hidden layer, and output layer, respectively. The goal of training a neural network is to optimize the parameters $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ such that the output $\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}$ matches the desired output $\boldsymbol{Y}$.

Instead of proposing new algorithms to minimize the objective function in (1), we are interested in characterizing its geometric landscape by understanding the behavior of all of its critical points.

3.1 Main Results

We present our main theorems concerning the behavior of all of the critical points of problem (1). First, note that an ambiguity exists in the solution to problem (1) since $\boldsymbol{W}_2\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{W}_1=\boldsymbol{W}_2\boldsymbol{W}_1$ (and thus $f(\boldsymbol{W}_1,\boldsymbol{W}_2)=f(\boldsymbol{A}^{-1}\boldsymbol{W}_1,\boldsymbol{W}_2\boldsymbol{A})$) for any invertible $\boldsymbol{A}$. In other words, if $(\boldsymbol{W}_1,\boldsymbol{W}_2)$ is a critical point, then $(\boldsymbol{A}^{-1}\boldsymbol{W}_1,\boldsymbol{W}_2\boldsymbol{A})$ is also a critical point. As pointed out in [13,27,35,46,47], if $\boldsymbol{A}^{-1}\boldsymbol{W}_1$ and $\boldsymbol{A}\boldsymbol{W}_2$ are extremely unbalanced in the sense that one has very large energy and the other one has very small energy—for example $\boldsymbol{A}=t\mathbf{I}$ when t is very large or small—then it could be difficult to directly analyze the property of such critical points. We utilize an additional regularizer [see (3)] to resolve the ambiguity issue and show that the corresponding objective function has no spurious local minima and obeys the strict saddle property without requiring any of the following conditions that appear in certain works discussed in Sect. 3.2: $\boldsymbol{Y}$ is of full row rank, $d_2 \leq d_0$, $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ has d_2 distinct eigenvalues, $d_1 \leq \min(d_0, d_2)$, (4) holds, or (5) holds.

Theorem 2 *Assume that $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ is of full row rank. Then, for any $\mu > 0$, the following objective function:*

$$g(\boldsymbol{W}_1, \boldsymbol{W}_2) = \frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y}\|_F^2 + \frac{\mu}{4}\rho(\boldsymbol{W}_1, \boldsymbol{W}_2) \quad (2)$$

with

$$\rho(\boldsymbol{W}_1, \boldsymbol{W}_2) := \|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}\|_F^2 \quad (3)$$

obeys the following properties:

(i) *$g(\boldsymbol{W}_1, \boldsymbol{W}_2)$ has the same global minimum value as $f(\boldsymbol{W}_1, \boldsymbol{W}_2)$ in (1);*
(ii) *any critical point $(\boldsymbol{W}_1, \boldsymbol{W}_2)$ of g is also a critical point of f;*
(iii) *$g(\boldsymbol{W}_1, \boldsymbol{W}_2)$ has no spurious local minima and the Hessian at any saddle point has a strictly negative eigenvalue.*

The proof of Theorem 2 is given in Sect. 4.1. The main idea in proving Theorem 2 is to connect $g(\boldsymbol{W}_1, \boldsymbol{W}_2)$ in (2) with the following low-rank factorization problem:

$$\min_{\widetilde{\boldsymbol{W}}_1, \boldsymbol{W}_2} \frac{1}{2}\|\boldsymbol{W}_2\widetilde{\boldsymbol{W}}_1 - \widetilde{\boldsymbol{Y}}\|_F^2 + \frac{\mu}{4}\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \widetilde{\boldsymbol{W}}_1\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}\|_F^2,$$

where $\widetilde{\boldsymbol{W}}_1$ and $\widetilde{\boldsymbol{Y}}$ are related to $\boldsymbol{W}$ and $\boldsymbol{Y}$; see (10) in Sect. 4.1 for the formal definitions.

Theorem 2(i) states that the regularizer $\rho(\boldsymbol{W}_1, \boldsymbol{W}_2)$ in (3) has no effect on the global minimum of the original problem, i.e., the one without this regularizer. Moreover, as established in Theorem 2(ii), any critical point of g in (2) is also a critical point of f in (1), but the converse is not true. With the regularizer $\rho(\boldsymbol{W}_1, \boldsymbol{W}_2)$, which mostly plays the role of shrinking the set of critical points, we prove that g has no spurious local minima and obeys the strict saddle property.

As our results hold for any $\mu > 0$ and $g = f$ when $\mu = 0$, one may conjecture that these properties also hold for the original objective function f under the same assumptions, i.e., assuming only that $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ has full row rank. This is indeed true and is formally established in the following result.

Theorem 3 *Assume that $\boldsymbol{X}$ is of full row rank. Then, the objective function f appearing in (1) has no spurious local minima and obeys the strict saddle property.*

The proof of Theorem 3 is given in Sect. 4.2. Theorem 3 builds heavily on Theorem 2 and on [34, Theorem 8], which is also presented in Theorem 5. Specifically, as we have noted, [34, Theorem 8] characterizes the behavior of degenerate critical points. Using Theorem 2, we further prove that any *non-degenerate* critical point of f is either a global minimum or a strict saddle.

3.2 Connection to Previous Work on Shallow Linear Neural Networks

As summarized in Table 1, the results in [2,21,31] on characterizing the geometric landscape of the loss function in training shallow linear neural networks only consider the behavior of local minima, but not saddle points. The strict saddle property is proved only in [20] and partly in [34]. We first review the result in [20] concerning the optimization geometry of problem (1).

Theorem 4 [20, Theorem 2.3] *Assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of full rank with $d_2 \leq d_0$ and $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ has d_2 distinct eigenvalues. Then, the objective function f appearing in (1) has no spurious local minima and obeys the strict saddle property.*

Theorem 4 implies that the objective function in (1) has benign geometric properties if $d_2 \leq d_0$ and the training samples are such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of full row rank and $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ has d_2 distinct eigenvalues. The recent work [31] generalizes the first point of Theorem 4 (i.e., no spurious local minima) by getting rid of the assumption that $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ has d_2 distinct eigenvalues. However, the geometry of the saddle points is not characterized in [31]. In [21], it is also proved that the condition on $\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ is not necessary. In particular, when applied to (1), the result in [21] implies that the objective function in (1) has no spurious local minima when $d_1 \leq \min(d_0, d_2)$. This condition requires that the hidden layer is narrower than the input and output layers. Again, the optimization geometry around saddle points is not discussed in [21].

We now review the more recent result in [34, Theorem 8].

Theorem 5 [34, Theorem 8] *The objective function f appearing in (1) has no spurious local minima. Moreover, any critical point $\boldsymbol{Z}$ of f that is degenerate (i.e., for which $\operatorname{rank}(\boldsymbol{W}_2\boldsymbol{W}_1) < \min\{d_2, d_1, d_0\}$) is either a global minimum of f or a strict saddle.*

In cases where the global minimum of f is non-degenerate—for example when $\boldsymbol{Y} = \boldsymbol{W}_2^\star\boldsymbol{W}_1^\star\boldsymbol{X}$ for some $\boldsymbol{W}_2^\star$ and $\boldsymbol{W}_1^\star$ such that $\boldsymbol{W}_2^\star\boldsymbol{W}_1^\star$ is non-degenerate—Theorem 5 implies that all degenerate critical points are strict saddles. However, we note that the behavior of non-degenerate critical points in these cases is more important from the algorithmic point of view, since one can always check the rank of a convergent point and perturb it if it is degenerate, but this is not possible at non-degenerate convergent points. Our Theorem 3 generalizes Theorem 5 to ensure that every critical point that is not a global minimum is a strict saddle, regardless of its rank.

Next, as a direct consequence of [47, Theorem 3], the following result also establishes certain conditions under which the objective function in (1) with an additional regularizer [see (6)] has no spurious local minima and obeys the strict saddle property.

Corollary 1 [47, Theorem 3] *Suppose* $d_1 \le \min(d_0, d_2)$*. Furthermore, for any* $d_2 \times d_0$ *matrix* $\boldsymbol{A}$ *with* $\operatorname{rank}(\boldsymbol{A}) \le 4d_1$*, suppose the following holds*

$$\alpha\|\boldsymbol{A}\|_F^2 \le \operatorname{trace}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}) \le \beta\|\boldsymbol{A}\|_F^2 \tag{4}$$

for some positive α *and* β *such that* $\frac{\beta}{\alpha} \le 1.5$*. Furthermore, suppose* $\min_{\boldsymbol{W}\in\mathbb{R}^{d_2\times d_0}}\|\boldsymbol{W}\boldsymbol{X}-\boldsymbol{Y}\|_F^2$ *admits a solution* $\boldsymbol{W}^\star$ *which satisfies*

$$0 < \operatorname{rank}(\boldsymbol{W}^\star) = r^\star \le d_1. \tag{5}$$

Then, for any $0 < \mu \le \frac{\alpha}{16}$*, the following objective function*

$$\begin{aligned} h(\boldsymbol{W}_1, \boldsymbol{W}_2) =& \frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}-\boldsymbol{Y}\|_F^2 \\ &+ \frac{\mu}{4}\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{W}_1^{\mathrm{T}}\|_F^2, \end{aligned} \tag{6}$$

has no spurious local minima and the Hessian at any saddle point has a strictly negative eigenvalue with

$$\begin{aligned} &\lambda_{\min}\left(\nabla^2 h(\boldsymbol{W}_1, \boldsymbol{W}_2)\right) \le \\ &\begin{cases} -0.08\alpha\sigma_{d_1}(\boldsymbol{W}^\star), & d_1 = r^\star \\ -0.05\alpha \cdot \min\left\{\sigma_{r_c}^2(\boldsymbol{W}_2\boldsymbol{W}_1), \sigma_{r^\star}(\boldsymbol{W}^\star)\right\}, & d_1 > r^\star \\ -0.1\alpha\sigma_{r^\star}(\boldsymbol{W}^\star), & r_c = 0, \end{cases} \end{aligned} \tag{7}$$

where $r_c \le d_1$ *is the rank of* $\boldsymbol{W}_1\boldsymbol{W}_2$*,* $\lambda_{\min}(\cdot)$ *represents the smallest eigenvalue, and* $\sigma_\ell(\cdot)$ *denotes the* ℓ*th largest singular value.*

Corollary 1, following from [47, Theorem 3], utilizes a regularizer $\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{W}_1^{\mathrm{T}}\|_F^2$ which balances the energy between $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ and has the effect of shrinking the set of critical points. This allows one to show that each critical point is either a global minimum or a strict saddle. Similar to Theorem 2(i), this regularizer also has no effect on the global minimum of the original problem (1).

As we explained before, Theorem 4 implicitly requires that $\min(d_0, d_1) \ge d_2$ and $\operatorname{rank}(\boldsymbol{W}_2^\star\boldsymbol{W}_1^\star) \ge d_2$. On the other hand, Corollary 1 requires $d_1 \le \min(d_0, d_2)$ and (4). When $d_1 \le \min(d_0, d_2)$, the hidden layer is narrower than the input and output layers. Note that (4) has nothing to do with the underlying network parameters $\boldsymbol{W}_1^\star$ and $\boldsymbol{W}_2^\star$, but requires the training data matrix $\boldsymbol{X}$ to act as an isometry operator for rank-$4d_1$ matrices. To see this, we rewrite

$$\operatorname{trace}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}) = \sum_{i=1}^N \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}_i = \sum_{i=1}^N \langle \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\rangle$$

which is a sum of the rank-one measurements of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

Unlike Theorem 4, which requires that $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ is of full rank and $d_2 \le d_0$, and unlike Corollary 1, which requires (4) and $d_1 \le \min(d_0, d_2)$, Theorems 2 and 3 only necessitate that $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ is full rank and have no condition on the size of d_0, d_1, and d_2. As we explained before, suppose $\boldsymbol{Y}$ is generated as $\boldsymbol{Y} = \boldsymbol{W}_2^\star\boldsymbol{W}_1^\star\boldsymbol{X}$, where $\boldsymbol{W}_2^\star$ and $\boldsymbol{W}_1^\star$ are the underlying weights to be recovered. Then, the full-rank assumption of $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{W}_2^\star\boldsymbol{W}_1^\star\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\star\mathrm{T}}\boldsymbol{W}_2^{\star\mathrm{T}}$ at least requires $\min(d_0, d_1) \ge d_2$ and $\operatorname{rank}(\boldsymbol{W}_2^\star\boldsymbol{W}_1^\star) \ge d_2$. In other words, Theorem 4 necessitates that the hidden layer is wider than the output, while Theorems 2 and 3 work for networks where the hidden layer is narrower than the input and output layers. On the other hand, Theorems 2 and 3 allow for the hidden layer of the network to be either narrower or wider than the input and the output layers.

Finally, consider a three-layer network with $\boldsymbol{X} = \mathbf{I}$. In this case, (1) reduces to a matrix factorization problem where $f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \|\boldsymbol{W}_2\boldsymbol{W}_1 - \boldsymbol{Y}\|_F^2$ and the regularizer in (2) is the same as the one in (6). Theorem 4 requires that $\boldsymbol{Y}$ is of full row rank and has d_2 distinct singular values. For the matrix factorization problem, we know from Corollary 1 that for any $\boldsymbol{Y}$, h has benign geometry (i.e., no spurious local minima and the strict saddle property) as long as $d_1 \le \min(d_0, d_2)$. As a direct consequence of Theorems 2 and 3, this benign geometry is also preserved even when $d_1 > d_0$ or $d_1 > d_2$ for matrix factorization via minimizing

$$g(\boldsymbol{W}_1, \boldsymbol{W}_2) = \|\boldsymbol{W}_2\boldsymbol{W}_1 - \boldsymbol{Y}\|_F^2 + \frac{\mu}{4}\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{W}_1^{\mathrm{T}}\|_F^2$$

where $\mu \ge 0$ (note that one can get rid of the regularizer by setting $\mu = 0$).

3.3 Possible Extensions

As we mentioned before, an ambiguity exists in the solution to the original training problem f in (1). In particular, $\boldsymbol{W}_2\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{W}_1 = \boldsymbol{W}_2\boldsymbol{W}_1$ (and thus $f(\boldsymbol{W}_1, \boldsymbol{W}_2) = f(\boldsymbol{A}^{-1}\boldsymbol{W}_1, \boldsymbol{W}_2\boldsymbol{A})$) for any invertible $\boldsymbol{A}$. Similar to [13,27,35,46,47], we utilize a regularizer ρ in (3) to address this ambiguity issue by shrinking the set of critical points, and we characterize the behavior of every critical point of the regularized objective g in (2). However, unlike the works mentioned above that only focus on a regularized problem, we further prove that the original problem f has a similar favorable geometry to g. We believe this technique could also be applied to problems such as the low-rank matrix recovery problems in [13,27,35,46,47]; in particular, it could potentially provide an answer to an open question arising in [47] as to whether the regularizer is needed since it is empirically observed that gradient descent always efficiently finds the global minimum even without the regularizer.

It is also of interest to extend this approach to deep linear neural networks which have a similar ambiguity issue. For example, consider a four-layer neural network which transforms $\boldsymbol{X}$ into $\boldsymbol{W}_3\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}$. In this case, aside from the regularizer in (3), one can utilize an additional regularizer such as $\|\boldsymbol{W}_3^{\mathrm{T}}\boldsymbol{W}_3-\boldsymbol{W}_2\boldsymbol{W}_2^{\mathrm{T}}\|_F^2$ to make $\boldsymbol{W}_3$ and $\boldsymbol{W}_2$ balanced (i.e., $\boldsymbol{W}_3^{\mathrm{T}}\boldsymbol{W}_3=\boldsymbol{W}_2\boldsymbol{W}_2^{\mathrm{T}}$). Similar to the analysis of the shallow linear network, the first step would be to characterize all the critical points for the problem with two regularizers. One would then need to insure that the original training problem without regularizer has a similar geometry to the regularized one. Toward that goal, one would need to extend Theorem 5—which characterizes the properties of degenerate critical points—to deep linear networks. We leave a full investigation for future work.

4 Proof of Main Results

4.1 Proof of Theorem 2

In this section, we prove Theorem 2 by individually proving its three arguments. For Theorem 2(i), it is clear that $g(\boldsymbol{W}_1,\boldsymbol{W}_2)\geq f(\boldsymbol{W}_1,\boldsymbol{W}_2)$ for any $\boldsymbol{W}_1,\boldsymbol{W}_2$, where we repeat that

$$f(\boldsymbol{W}_1,\boldsymbol{W}_2)=\frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}-\boldsymbol{Y}\|_F^2.$$

Therefore, we need only show that g has the same objective function as f at the global minimum of f. The proof of Theorem 2(ii) mainly relies on (9) in Lemma 1 which implies that at any critical point $(\boldsymbol{W}_1,\boldsymbol{W}_2)$ of g, the regularizer ρ achieves its global minimum and hence its gradient is also zero, suggesting that $\nabla f(\boldsymbol{W}_1,\boldsymbol{W}_2)=\nabla g(\boldsymbol{W}_1,\boldsymbol{W}_2)=\mathbf{0}$. To prove Theorem 2(iii), we characterize the behavior of all of the critical points of the objective function g in (2). In particular, we show that for any critical point of g, if it is not a global minimum, then it is a strict saddle, i.e., its Hessian has at least one negative eigenvalue.

Proof of Theorem 2(i) Suppose the row rank of $\boldsymbol{X}$ is $d_0'\leq d_0$. Let

$$\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{8}$$

be a reduced SVD of $\boldsymbol{X}$, where $\boldsymbol{\Sigma}$ is a $d_0'\times d_0'$ diagonal matrix with positive diagonals. Then,

$$\begin{aligned} &f(\boldsymbol{W}_1,\boldsymbol{W}_2)\\ &\quad=\frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{U}\boldsymbol{\Sigma}-\boldsymbol{Y}\boldsymbol{V}\|_F^2+\|\boldsymbol{Y}\|_F^2-\|\boldsymbol{Y}\boldsymbol{V}\|_F^2\\ &\quad=f_1(\boldsymbol{W}_1,\boldsymbol{W}_2)+C,\end{aligned}$$

where $f_1(\boldsymbol{W}_1,\boldsymbol{W}_2)=\frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{U}\boldsymbol{\Sigma}-\boldsymbol{Y}\boldsymbol{V}\|_F^2$ and $C=\|\boldsymbol{Y}\|_F^2-\|\boldsymbol{Y}\boldsymbol{V}\|_F^2$. Denote by $(\boldsymbol{W}_1^\star,\boldsymbol{W}_2^\star)$ a global minimum of $f_1(\boldsymbol{W}_1,\boldsymbol{W}_2)$:

$$\begin{aligned}(\boldsymbol{W}_1^\star,\boldsymbol{W}_2^\star)&=\arg\min\nolimits_{\boldsymbol{W}_1,\boldsymbol{W}_2}f(\boldsymbol{W}_1,\boldsymbol{W}_2)\\ &=\arg\min\nolimits_{\boldsymbol{W}_1,\boldsymbol{W}_2}f_1(\boldsymbol{W}_1,\boldsymbol{W}_2).\end{aligned}$$

We now construct $(\widehat{\boldsymbol{W}}_1,\widehat{\boldsymbol{W}}_2)$ such that $g(\widehat{\boldsymbol{W}}_1,\widehat{\boldsymbol{W}}_2)=f(\boldsymbol{W}_1^\star,\boldsymbol{W}_2^\star)$. Toward that goal, let $\boldsymbol{W}_2^\star\boldsymbol{W}_1^\star\boldsymbol{U}\boldsymbol{\Sigma}=\boldsymbol{P}_1\boldsymbol{\Omega}\boldsymbol{Q}_1^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{W}_2^\star\boldsymbol{W}_1^\star\boldsymbol{U}\boldsymbol{\Sigma}$, where $\boldsymbol{\Omega}$ is a diagonal matrix with positive diagonals. Also, let $\widehat{\boldsymbol{W}}_1=\boldsymbol{\Omega}^{1/2}\boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}$ and $\widehat{\boldsymbol{W}}_2=\boldsymbol{P}_1\boldsymbol{\Omega}^{1/2}$. It follows that

$$\begin{aligned}&\widehat{\boldsymbol{W}}_2^{\mathrm{T}}\widehat{\boldsymbol{W}}_2-\widehat{\boldsymbol{W}}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{W}}_1^{\mathrm{T}}=\boldsymbol{\Omega}-\boldsymbol{\Omega}=\mathbf{0},\\ &\widehat{\boldsymbol{W}}_2\widehat{\boldsymbol{W}}_1\boldsymbol{U}\boldsymbol{\Sigma}=\boldsymbol{P}_1\boldsymbol{\Omega}\boldsymbol{Q}_1^{\mathrm{T}}=\boldsymbol{W}_2^\star\boldsymbol{W}_1^\star\boldsymbol{U}\boldsymbol{\Sigma},\end{aligned}$$

which implies that $f_1(\boldsymbol{W}_1^\star,\boldsymbol{W}_2^\star)=f_1(\widehat{\boldsymbol{W}}_1,\widehat{\boldsymbol{W}}_2)$ and

$$\begin{aligned}&f(\boldsymbol{W}_1^\star,\boldsymbol{W}_2^\star)=f_1(\boldsymbol{W}_1^\star,\boldsymbol{W}_2^\star)+C=f_1(\widehat{\boldsymbol{W}}_1,\widehat{\boldsymbol{W}}_2)+C\\ &\quad=f(\widehat{\boldsymbol{W}}_1,\widehat{\boldsymbol{W}}_2)=g(\widehat{\boldsymbol{W}}_1,\widehat{\boldsymbol{W}}_2)\end{aligned}$$

since $\|\widehat{\boldsymbol{W}}_2^{\mathrm{T}}\widehat{\boldsymbol{W}}_2-\widehat{\boldsymbol{W}}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{W}}_1^{\mathrm{T}}\|_F^2=0$. This further indicates that g and f have the same global optimum (since $g(\boldsymbol{W}_1,\boldsymbol{W}_2)\geq f(\boldsymbol{W}_1,\boldsymbol{W}_2)$ for any $(\boldsymbol{W}_1,\boldsymbol{W}_2)$). This proves that the regularizer in (2) has no effect on the global minimum of the original problem. □

Proof of Theorem 2(ii) We first establish the following result that characterizes all the critical points of g. □

Lemma 1 *Let $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ be an SVD of $\boldsymbol{X}$ as in (8), where $\boldsymbol{\Sigma}$ is a diagonal matrix with positive diagonals $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_{d_0}>0$. Let $\widetilde{\boldsymbol{Y}}:=\boldsymbol{Y}\boldsymbol{V}=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}=\sum_{j=1}^r\lambda_j\boldsymbol{p}_j\boldsymbol{q}_j^{\mathrm{T}}$ be a reduced SVD of $\widetilde{\boldsymbol{Y}}$, where r is the rank of $\widetilde{\boldsymbol{Y}}$. Then, any critical point $\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{W}_2\\ \boldsymbol{W}_1^{\mathrm{T}}\end{bmatrix}$ of (2) satisfies*

$$\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2=\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}. \tag{9}$$

Furthermore, any $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{W}_2 \\ \boldsymbol{W}_1^{\mathrm{T}} \end{bmatrix}$ *is a critical point of* $g(\boldsymbol{Z})$ *if and only if* $\boldsymbol{Z} \in \mathcal{C}_g$ *with*

$$\mathcal{C}_g = \left\{ \boldsymbol{Z} = \begin{bmatrix} \widetilde{\boldsymbol{W}}_2 \boldsymbol{R}^{\mathrm{T}} \\ \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{W}}_1^{\mathrm{T}} \boldsymbol{R}^{\mathrm{T}} \end{bmatrix} : \widetilde{\boldsymbol{Z}} = \begin{bmatrix} \widetilde{\boldsymbol{W}}_2 \\ \widetilde{\boldsymbol{W}}_1^{\mathrm{T}} \end{bmatrix}, \right.$$
$$\widetilde{\boldsymbol{z}}_i \in \left\{ \sqrt{\lambda_1} \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{q}_1 \end{bmatrix}, \ldots, \sqrt{\lambda_r} \begin{bmatrix} \boldsymbol{p}_r \\ \boldsymbol{q}_r \end{bmatrix}, \boldsymbol{0} \right\}, \tag{10}$$
$$\left. \widetilde{\boldsymbol{z}}_i^{\mathrm{T}} \widetilde{\boldsymbol{z}}_j = 0, \ \forall i \neq j, \boldsymbol{R} \in \mathcal{O}_{d_1} \right\},$$

where $\widetilde{\boldsymbol{z}}_i$ *denotes the* i*th column of* $\widetilde{\boldsymbol{Z}}$.

The proof of Lemma 1 is in "Appendix A." From (9), $g(\boldsymbol{Z}) = f(\boldsymbol{Z})$ at any critical point $\boldsymbol{Z}$. We compute the gradient of the regularizer $\rho(\boldsymbol{W}_1, \boldsymbol{W}_2)$ as

$$\nabla_{\boldsymbol{W}_1} \rho(\boldsymbol{W}_1, \boldsymbol{W}_2)$$
$$:= -\mu \boldsymbol{W}_2(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}})\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}},$$
$$\nabla_{\boldsymbol{W}_2} \rho(\boldsymbol{W}_1, \boldsymbol{W}_2)$$
$$:= \mu \boldsymbol{W}_2(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_2\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}})\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}.$$

Plugging (9) into the above equations gives

$$\nabla_{\boldsymbol{W}_1} \rho(\boldsymbol{W}_1, \boldsymbol{W}_2) = \nabla_{\boldsymbol{W}_2} \rho(\boldsymbol{W}_1, \boldsymbol{W}_2) = \boldsymbol{0}$$

for any critical point $(\boldsymbol{W}_1, \boldsymbol{W}_2)$ of g. This further implies that if $\boldsymbol{Z}$ is a critical point of g, then it must also be a critical point of f since $\nabla g(\boldsymbol{Z}) = \nabla f(\boldsymbol{Z}) + \nabla \rho(\boldsymbol{Z})$ and both $\nabla g(\boldsymbol{Z}) = \nabla \rho(\boldsymbol{Z}) = \boldsymbol{0}$, so that

$$\nabla f(\boldsymbol{Z}) = \boldsymbol{0}. \tag{11}$$

This proves Theorem 2(ii).

Proof of Theorem 2(iii) We show that any critical point of g is either a global minimum or a strict saddle. Toward that end, for any $\boldsymbol{Z} \in \mathcal{C}$, we first write the objective value at this point as

$$g(\boldsymbol{Z}) = \frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y}\|_F^2$$
$$= \frac{1}{2}\|\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{Y}\|_F^2$$
$$= \frac{1}{2}\|\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{Y}\boldsymbol{V}\|_F^2 + \|\boldsymbol{Y}\|_F^2 - \|\boldsymbol{Y}\boldsymbol{V}\|_F^2,$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ is a reduced SVD of $\boldsymbol{X}$ as defined in (8), and $\widetilde{\boldsymbol{W}}_2$ and $\widetilde{\boldsymbol{W}}_1$ are defined in (10). Noting that $\|\boldsymbol{Y}\|_F^2 - \|\boldsymbol{Y}\boldsymbol{V}\|_F^2$ is a constant in terms of the variables $\boldsymbol{W}_2$ and $\boldsymbol{W}_1$, we conclude that $\boldsymbol{Z}$ is a global minimum of $g(\boldsymbol{Z})$ if and only if $\widetilde{\boldsymbol{Z}}$ is a global minimum of

$$\widetilde{g}(\widetilde{\boldsymbol{Z}}) := \frac{1}{2}\|\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{Y}\boldsymbol{V}\|_F^2. \tag{12}$$

Based on this observation, the following result further characterizes the behavior of all of the critical points in Lemma 1. □

Lemma 2 *With the same setup as in Lemma 1, let* $\mathcal{C}$ *be defined in* (10)*. Then, all local minima of* (2) *belong to the following set* (*which contains all the global solutions of* (2))

$$\mathcal{X}_g = \left\{ \boldsymbol{Z} = \begin{bmatrix} \widetilde{\boldsymbol{W}}_2 \boldsymbol{R} \\ \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}\boldsymbol{R} \end{bmatrix} \in \mathcal{C} : \|\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{Y}\boldsymbol{V}\|_F^2 \right.$$
$$\left. = \min_{\boldsymbol{A}\in\mathbb{R}^{d_2\times d_1}, \boldsymbol{B}\in\mathbb{R}^{d_1\times d_0}} \|\boldsymbol{A}\boldsymbol{B} - \boldsymbol{Y}\boldsymbol{V}\|_F^2 \right\}. \tag{13}$$

Any $\boldsymbol{Z} \in \mathcal{C}_g \setminus \mathcal{X}_g$ *is a strict saddle of* $g(\boldsymbol{Z})$ *satisfying:*

- *If* $r \leq d_1$*, then*

$$\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -2\frac{\lambda_r}{1 + \sum_i \sigma_i^{-2}}; \tag{14}$$

- *If* $r > d_1$*, then*

$$\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -2\frac{\lambda_{d_1} - \lambda_{r'}}{1 + \sum_i \sigma_i^{-2}}, \tag{15}$$

where $\lambda_{r'}$ *is the largest singular value of* $\widetilde{\boldsymbol{Y}}$ *that is strictly smaller than* λ_{d_1}.

The proof of Lemma 2 is given in "Appendix B." Lemma 2 states that any critical point of g is either a global minimum or a strict saddle. This proves Theorem 2(iii) and thus we complete the proof of Theorem 2.

4.2 Proof of Theorem 3

Building on Theorem 2 and Theorem 5, we now consider the landscape of f in (1). Let $\mathcal{C}_f$ denote the set of critical points of f:

$$\mathcal{C}_f = \{\boldsymbol{Z} : \nabla f(\boldsymbol{Z}) = \boldsymbol{0}\}.$$

Our goal is to characterize the behavior of all critical points that are not global minima. In particular, we want to show that every critical point of f is either a global minimum or a strict saddle.

Let $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{W}_2 \\ \boldsymbol{W}_1^{\mathrm{T}} \end{bmatrix}$ be any critical point in $\mathcal{C}_f$. According to Theorem 5, when $\boldsymbol{W}_2\boldsymbol{W}_1$ is degenerate (i.e., $\operatorname{rank}(\boldsymbol{W}_2\boldsymbol{W}_1) <$

$\min\{d_2, d_1, d_0\}$), $\boldsymbol{Z}$ must be either a global minimum or a strict saddle. We now assume the other case that $\boldsymbol{W}_2\boldsymbol{W}_1$ is non-degenerate. For this case, we first construct a surrogate function $\overline{f}$ [see (18)] similar to the one in (12). We then connect the critical points of f to those of $\overline{f}$, which according to Lemma 2 are either global minima or strict saddles.

Let $\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{U}\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Theta}\boldsymbol{\Psi}^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{U}\boldsymbol{\Sigma}$, where $\boldsymbol{\Theta}$ is a diagonal and square matrix with positive singular values, and $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are orthonormal matrices of proper dimension. We now construct

$$\begin{aligned}\overline{\boldsymbol{W}}_2 &= \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Psi}\boldsymbol{\Theta}^{-1/2} = \boldsymbol{\Phi}\boldsymbol{\Theta}^{1/2},\\ \overline{\boldsymbol{W}}_1 &= \boldsymbol{\Theta}^{-1/2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}_2\boldsymbol{W}_1 = \boldsymbol{\Theta}^{1/2}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}.\end{aligned} \tag{16}$$

The above-constructed pair $\overline{\boldsymbol{Z}} = \begin{bmatrix}\overline{\boldsymbol{W}}_2\\ \overline{\boldsymbol{W}}_1^{\mathrm{T}}\end{bmatrix}$ satisfies

$$\overline{\boldsymbol{W}}_2^{\mathrm{T}}\overline{\boldsymbol{W}}_2 = \overline{\boldsymbol{W}}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\overline{\boldsymbol{W}}_1^{\mathrm{T}},\ \ \overline{\boldsymbol{W}}_2\overline{\boldsymbol{W}}_1 = \boldsymbol{W}_2\boldsymbol{W}_1. \tag{17}$$

Note that here $\overline{\boldsymbol{W}}_2$ (resp. $\overline{\boldsymbol{W}}_1$) have different numbers of columns (resp. rows) than $\boldsymbol{W}_2$ (resp. $\boldsymbol{W}_1$). We denote by

$$\overline{f}(\overline{\boldsymbol{W}}_1, \overline{\boldsymbol{W}}_2) = \frac{1}{2}\|\overline{\boldsymbol{W}}_2\overline{\boldsymbol{W}}_1\boldsymbol{U}\boldsymbol{\Sigma} - \boldsymbol{Y}\boldsymbol{V}\|_F^2. \tag{18}$$

Since $\boldsymbol{Z} \in \mathcal{C}_f$, we have $\nabla_{\boldsymbol{W}_1} f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \boldsymbol{0}$ and $\nabla_{\boldsymbol{W}_2} f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \boldsymbol{0}$. It follows that

$$\begin{aligned}\nabla_{\overline{\boldsymbol{W}}_2}\overline{f}(\overline{\boldsymbol{W}}_1, \overline{\boldsymbol{W}}_2) &= (\overline{\boldsymbol{W}}_2\overline{\boldsymbol{W}}_1\boldsymbol{U}\boldsymbol{\Sigma} - \boldsymbol{Y}\boldsymbol{V})(\overline{\boldsymbol{W}}_1\boldsymbol{U}\boldsymbol{\Sigma})^{\mathrm{T}}\\ &= \nabla_{\boldsymbol{W}_2} f(\boldsymbol{W}_1, \boldsymbol{W}_2)\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Theta}^{-1/2}\\ &= \boldsymbol{0}.\end{aligned}$$

And similarly, we have

$$\begin{aligned}&\nabla_{\overline{\boldsymbol{W}}_1}\overline{f}(\overline{\boldsymbol{W}}_1, \overline{\boldsymbol{W}}_2)\\ &\quad = \boldsymbol{\Theta}^{-1/2}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}\nabla_{\boldsymbol{W}_1} f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \boldsymbol{0},\end{aligned}$$

which together with the above inequation implies that $\overline{\boldsymbol{Z}}$ is also a critical point of $\overline{f}$. Due to (17) which states that $\overline{\boldsymbol{Z}}$ also satisfies (9), it follows from the same arguments used in Lemmas 1 and 2 that $\overline{\boldsymbol{Z}}$ is either a global minimum or a strict saddle of $\overline{f}$. Moreover, since $\overline{\boldsymbol{W}}_2\overline{\boldsymbol{W}}_1$ has the same rank as $\boldsymbol{W}_2\boldsymbol{W}_1$ which is assumed to be non-degenerate, we have that $\overline{\boldsymbol{Z}}$ is a global minimum of $\overline{f}$ if and only if

$$\|\overline{\boldsymbol{W}}_2\overline{\boldsymbol{W}}_1\boldsymbol{U}\boldsymbol{\Sigma} - \boldsymbol{Y}\boldsymbol{V}\|_F^2 = \min_{\substack{\boldsymbol{A}\in\mathbb{R}^{d_2\times d_1}\\ \boldsymbol{B}\in\mathbb{R}^{d_1\times d_0}}} \|\boldsymbol{A}\boldsymbol{B} - \boldsymbol{Y}\boldsymbol{V}\|_F^2,$$

where the minimum of the right-hand side is also achieved by the global minimum of f according to (13). Therefore, if $\overline{\boldsymbol{Z}}$ is a global minimum of $\overline{f}$, then $\boldsymbol{Z}$ is also a global minimum of f.

Now, we consider the other case when $\overline{\boldsymbol{Z}}$ is not a global minimum of $\overline{f}$, i.e., it is a strict saddle. In this case, there exists $\overline{\boldsymbol{\Delta}} = \begin{bmatrix}\overline{\boldsymbol{\Delta}}_2\\ \overline{\boldsymbol{\Delta}}_1^{\mathrm{T}}\end{bmatrix}$ such that

$$[\nabla^2\overline{f}(\overline{\boldsymbol{W}}_1, \overline{\boldsymbol{W}}_2)](\overline{\boldsymbol{\Delta}}, \overline{\boldsymbol{\Delta}}) < 0.$$

Now, construct

$$\begin{aligned}\boldsymbol{\Delta}_1 &= \boldsymbol{W}_1\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Psi}\boldsymbol{\Theta}^{-1/2}\overline{\boldsymbol{\Delta}}_1\\ \boldsymbol{\Delta}_2 &= \overline{\boldsymbol{\Delta}}_2\boldsymbol{\Theta}^{-1/2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}_2\end{aligned}$$

which satisfies

$$\begin{aligned}&\boldsymbol{W}_2\boldsymbol{\Delta}_1 = \overline{\boldsymbol{W}}_2\overline{\boldsymbol{W}}_1,\ \ \boldsymbol{\Delta}_{\boldsymbol{W}_2}\boldsymbol{W}_1 = \overline{\boldsymbol{\Delta}}_2\overline{\boldsymbol{W}}_1,\\ &\boldsymbol{\Delta}_2\boldsymbol{\Delta}_1 = \overline{\boldsymbol{\Delta}}_2\overline{\boldsymbol{\Delta}}_1.\end{aligned}$$

By the Hessian quadratic form given in (32) (ignoring the μ terms), we have

$$[\nabla^2 f(\boldsymbol{Z})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) = [\nabla^2\overline{f}(\overline{\boldsymbol{Z}})](\overline{\boldsymbol{\Delta}}, \overline{\boldsymbol{\Delta}}) < 0,$$

which implies that $\boldsymbol{Z}$ is a strict saddle of f. This completes the proof of Theorem 3.

5 Conclusion

We consider the optimization landscape of the objective function in training shallow linear networks. In particular, we proved that the corresponding optimization problems under a very mild condition have a simple landscape: There are no spurious local minima, and any critical point is either a local (and thus also global) minimum or a strict saddle such that the Hessian evaluated at this point has a strictly negative eigenvalue. These properties guarantee that a number of iterative optimization algorithms (especially gradient descent, which is widely used in training neural networks) converge to a global minimum from either a random initialization or an arbitrary initialization depending on the specific algorithm used. It would be of interest to prove similar geometric properties for the training problem without the mild condition on the row rank of $\boldsymbol{X}$.

A Proof of Lemma 1

A.1 Proof of (9)

Intuitively, the regularizer ρ in (3) forces $\boldsymbol{W}_2^{\mathrm{T}}$ and $\boldsymbol{W}_1\boldsymbol{X}$ to be balanced (i.e., $\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 = \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}$). We show that with this regularizer, any critical point of g obeys (9). To establish this, first note that any critical point $\boldsymbol{Z}$ of $g(\boldsymbol{Z})$ satisfies $\nabla g(\boldsymbol{Z}) = \mathbf{0}$, i.e.,

$$\begin{aligned}\nabla_{\boldsymbol{W}_1} g(\boldsymbol{W}_1, \boldsymbol{W}_2) = \boldsymbol{W}_2^{\mathrm{T}}(\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}} \\ - \mu(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}})\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} = \mathbf{0},\end{aligned} \tag{19}$$

and

$$\begin{aligned}\nabla_{\boldsymbol{W}_2} g(\boldsymbol{W}_1, \boldsymbol{W}_2) = (\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}} \\ + \mu\boldsymbol{W}_2(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}) = \mathbf{0}.\end{aligned} \tag{20}$$

By (19), we obtain

$$\begin{aligned}&\boldsymbol{W}_2^{\mathrm{T}}(\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}} \\ &= \mu(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}})\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}.\end{aligned} \tag{21}$$

Multiplying (20) on the left by $\boldsymbol{W}_2^{\mathrm{T}}$ and plugging the result with the expression for $\boldsymbol{W}_2^{\mathrm{T}}(\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}$ in (21) gives

$$\begin{aligned}&(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}})\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}} \\ &+ \boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}) = \mathbf{0},\end{aligned}$$

which is equivalent to

$$\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 = \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}.$$

Note that $\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2$ and $\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}$ are the principal square roots (i.e., PSD square roots) of $\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2$ and $\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}$, respectively. Utilizing the result that a PSD matrix $\boldsymbol{A}$ has a unique PSD matrix $\boldsymbol{B}$ such that $\boldsymbol{B}^k = \boldsymbol{A}$ for any $k \geq 1$ [16, Theorem 7.2.6], we obtain

$$\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 = \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}$$

for any critical point $\boldsymbol{Z}$.

A.2 Proof of (10)

To show (10), we first plug (9) back into (19) and (20), simplifying the first-order optimality equation as

$$\begin{aligned}&\boldsymbol{W}_2^{\mathrm{T}}(\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}} = \mathbf{0}, \\ &(\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}} = \mathbf{0}.\end{aligned} \tag{22}$$

What remains is to find all $(\boldsymbol{W}_1, \boldsymbol{W}_2)$ that satisfy the above equation.

Let $\boldsymbol{W}_2 = \boldsymbol{L}\boldsymbol{\Pi}\boldsymbol{R}^{\mathrm{T}}$ be a full SVD of $\boldsymbol{W}_2$, where $\boldsymbol{L} \in \mathbb{R}^{d_2\times d_2}$ and $\boldsymbol{R} \in \mathbb{R}^{d_1\times d_1}$ are orthonormal matrices. Define

$$\widetilde{\boldsymbol{W}}_2 = \boldsymbol{W}_2\boldsymbol{R} = \boldsymbol{L}\boldsymbol{\Pi}, \; \widetilde{\boldsymbol{W}}_1 = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{W}_1\boldsymbol{U}\boldsymbol{\Sigma}. \tag{23}$$

Since $\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}} = \boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2$ [see (9)], we have

$$\widetilde{\boldsymbol{W}}_1\widetilde{\boldsymbol{W}}_1^{\mathrm{T}} = \widetilde{\boldsymbol{W}}_2^{\mathrm{T}}\widetilde{\boldsymbol{W}}_2 = \boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Pi}. \tag{24}$$

Noting that $\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Pi}$ is a diagonal matrix with nonnegative diagonals, it follows that $\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}$ is an orthogonal matrix, but possibly includes zero columns.

Due to (22), we have

$$\begin{aligned}&\widetilde{\boldsymbol{W}}_2^{\mathrm{T}}(\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{Y}\boldsymbol{V})\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} \\ &\quad = \boldsymbol{R}^{\mathrm{T}}(\boldsymbol{W}_2^{\mathrm{T}}(\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}) = \mathbf{0}, \\ &(\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{Y}\boldsymbol{V})\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1^{\mathrm{T}} \\ &\quad = (\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{R} = \mathbf{0},\end{aligned} \tag{25}$$

where we utilized the reduced SVD decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ in (8). Note that the diagonals of $\boldsymbol{\Sigma}$ are all positive and recall

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{V}.$$

Then, (25) gives

$$\begin{aligned}&\widetilde{\boldsymbol{W}}_2^{\mathrm{T}}(\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \widetilde{\boldsymbol{Y}}) = \mathbf{0}, \\ &(\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \widetilde{\boldsymbol{Y}})\widetilde{\boldsymbol{W}}_1^{\mathrm{T}} = \mathbf{0}.\end{aligned} \tag{26}$$

We now compute all $\widetilde{\boldsymbol{W}}_2$ and $\widetilde{\boldsymbol{W}}_1$ satisfying (26). To that end, let $\boldsymbol{\phi} \in \mathbb{R}^{d_2}$ and $\boldsymbol{\psi} \in \mathbb{R}^{d_0}$ be the ith column and the ith row of $\widetilde{\boldsymbol{W}}_2$ and $\widetilde{\boldsymbol{W}}_1$, respectively. Due to (24), we have

$$\|\boldsymbol{\phi}\|_2 = \|\boldsymbol{\psi}\|_2. \tag{27}$$

It follows from (26) that

$$\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\boldsymbol{\phi} = \|\boldsymbol{\phi}\|_2^2\boldsymbol{\psi}, \tag{28}$$

$$\widetilde{\boldsymbol{Y}}\boldsymbol{\psi} = \|\boldsymbol{\psi}\|_2^2\boldsymbol{\phi}. \tag{29}$$

Multiplying (28) by $\widetilde{\boldsymbol{Y}}$ and plugging (29) into the resulting equation gives

$$\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\boldsymbol{\phi} = \|\boldsymbol{\phi}\|_2^4\boldsymbol{\phi}, \tag{30}$$

where we used (27). Similarly, we have

$$\widetilde{\boldsymbol{Y}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}\boldsymbol{\psi} = \|\boldsymbol{\psi}\|_2^4\boldsymbol{\psi}. \tag{31}$$

Let $\widetilde{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}} = \sum_{j=1}^{r}\lambda_j\boldsymbol{p}_j\boldsymbol{q}_j^{\mathrm{T}}$ be the reduced SVD of $\widetilde{\boldsymbol{Y}}$. It follows from (30) that $\boldsymbol{\phi}$ is either a zero vector (i.e., $\boldsymbol{\phi} = \mathbf{0}$), or a left singular vector of $\widetilde{\boldsymbol{Y}}$ (i.e., $\boldsymbol{\phi} = \alpha\boldsymbol{p}_j$ for some $j \in [r]$). Plugging $\boldsymbol{\phi} = \alpha\boldsymbol{p}_j$ into (30) gives

$$\lambda_j^2 = \alpha^4.$$

Thus, $\boldsymbol{\phi} = \pm\sqrt{\lambda_j}\boldsymbol{p}_j$. If $\boldsymbol{\phi} = \mathbf{0}$, then due to (27), we have $\boldsymbol{\psi} = \mathbf{0}$. If $\boldsymbol{\phi} = \pm\sqrt{\lambda_j}\boldsymbol{p}_j$, then plugging into (28) gives

$$\boldsymbol{\psi} = \pm\sqrt{\lambda_j}\boldsymbol{q}_j.$$

Thus, we conclude that

$$(\boldsymbol{\phi}, \boldsymbol{\psi}) \in \left\{\pm\sqrt{\lambda_1}(\boldsymbol{p}_1, \boldsymbol{q}_1), \ldots, \pm\sqrt{\lambda_r}(\boldsymbol{p}_r, \boldsymbol{q}_r), (\mathbf{0}, \mathbf{0})\right\},$$

which together with (24) implies that any critical point $\boldsymbol{Z}$ belongs to (10) by absorbing the sign $\pm$ into $\boldsymbol{R}$.

We now prove the other direction $\Rightarrow$. For any $\boldsymbol{Z} \in \mathcal{C}$, we compute the gradient of g at this point and directly verify it satisfies (19) and (20), i.e., $\boldsymbol{Z}$ is a critical point of $g(\boldsymbol{Z})$. This completes the proof of Lemma 1.

B Proof of Lemma 2

Due to the fact that $\boldsymbol{Z}$ is a global minimum of $g(\boldsymbol{Z})$ if and only if $\widetilde{\boldsymbol{Z}}$ is a global minimum of $\widetilde{g}(\widetilde{\boldsymbol{Z}})$, we know any $\boldsymbol{Z} \in \mathcal{X}$ is a global minimum of $g(\boldsymbol{Z})$. The rest is to show that any $\boldsymbol{Z} \in \mathcal{C} \setminus \mathcal{X}$ is a strict saddle. For this purpose, we first compute the Hessian quadrature form $\nabla^2 g(\boldsymbol{Z})[\boldsymbol{\Delta}, \boldsymbol{\Delta}]$ for any $\boldsymbol{\Delta} = \begin{bmatrix}\boldsymbol{\Delta}_2 \\ \boldsymbol{\Delta}_1^{\mathrm{T}}\end{bmatrix}$ (with $\boldsymbol{\Delta}_1 \in \mathbb{R}^{d_1 \times d_0}$, $\boldsymbol{\Delta}_2 \in \mathbb{R}^{d_2 \times d_1}$) as

$$\begin{aligned}
&\nabla^2 g(\boldsymbol{Z})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] \\
&\quad = \|(\boldsymbol{W}_2\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2\boldsymbol{W}_1)\boldsymbol{X}\|_F^2 \\
&\qquad + 2\left\langle \boldsymbol{\Delta}_2\boldsymbol{\Delta}_1, (\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}\right\rangle \\
&\qquad + \mu\big(\langle \boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}, \boldsymbol{\Delta}_2^{\mathrm{T}}\boldsymbol{\Delta}_2 - \boldsymbol{\Delta}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Delta}_1^{\mathrm{T}}\rangle + \\
&\qquad \frac{1}{2}\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Delta}_1^{\mathrm{T}} - \boldsymbol{\Delta}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{T}\|_F^2\big) \\
&\quad = \|(\boldsymbol{W}_2\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2\boldsymbol{W}_1)\boldsymbol{X}_1\|_F^2 \\
&\qquad + 2\left\langle \boldsymbol{\Delta}_2\boldsymbol{\Delta}_1, (\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}\right\rangle + \\
&\qquad \frac{\mu}{2}\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Delta}_1^{\mathrm{T}} - \boldsymbol{\Delta}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{T}\|_F^2,
\end{aligned} \tag{32}$$

where the second equality follows because any critical point $\boldsymbol{Z}$ satisfies (9). We continue the proof by considering two cases in which we provide explicit expressions for the set $\mathcal{X}$ that contains all the global minima and construct a negative direction for g at all the points $\mathcal{C} \setminus \mathcal{X}$.

Case i: $r \le d_1$. In this case, $\min \widetilde{g}(\widetilde{\boldsymbol{Z}}) = 0$ and $\widetilde{g}(\widetilde{\boldsymbol{Z}})$ achieves its global minimum 0 if and only if $\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 = \boldsymbol{Y}\boldsymbol{V}$. Thus, we rewrite $\mathcal{X}$ as

$$\mathcal{X} = \left\{\boldsymbol{Z} = \begin{bmatrix}\widetilde{\boldsymbol{W}}_2\boldsymbol{R} \\ \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}\boldsymbol{R}\end{bmatrix} \in \mathcal{C} : \widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 = \boldsymbol{Y}\boldsymbol{V}\right\},$$

which further implies that

$$\mathcal{C} \setminus \mathcal{X} = \left\{\boldsymbol{Z} = \begin{bmatrix}\widetilde{\boldsymbol{W}}_2\boldsymbol{R} \\ \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}\boldsymbol{R}\end{bmatrix} \in \mathcal{C} : \boldsymbol{Y}\boldsymbol{V} - \widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 = \sum_{i\in\Omega}\lambda_i\boldsymbol{p}_i\boldsymbol{q}_i^{\mathrm{T}}, \Omega \subset [r]\right\}.$$

Thus, for any $\boldsymbol{Z} \in \mathcal{C} \setminus \mathcal{X}$, the corresponding $\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1$ is a low-rank approximation to $\boldsymbol{Y}\boldsymbol{V}$.

Let $k \in \Omega$. We have

$$\boldsymbol{p}_k^{\mathrm{T}}\widetilde{\boldsymbol{W}}_2 = \mathbf{0}, \ \widetilde{\boldsymbol{W}}_1\boldsymbol{q}_k = \mathbf{0}. \tag{33}$$

In words, $\boldsymbol{p}_k$ and $\boldsymbol{q}_k$ are orthogonal to $\widetilde{\boldsymbol{W}}_2$ and $\widetilde{\boldsymbol{W}}_1$, respectively. Let $\boldsymbol{\alpha} \in \mathbb{R}^{d_1}$ be the eigenvector associated with the smallest eigenvalue of $\widetilde{\boldsymbol{Z}}^{\mathrm{T}}\widetilde{\boldsymbol{Z}}$. Note that such $\boldsymbol{\alpha}$ simultaneously lives in the null spaces of $\widetilde{\boldsymbol{W}}_2$ and $\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}$ since $\widetilde{\boldsymbol{Z}}$ is rank deficient, indicating

$$0 = \boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{Z}}^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_2^{\mathrm{T}}\widetilde{\boldsymbol{W}}_2\boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}\boldsymbol{\alpha},$$

which further implies

$$\widetilde{\boldsymbol{W}}_2\boldsymbol{\alpha} = \mathbf{0}, \ \widetilde{\boldsymbol{W}}_1^{\mathrm{T}}\boldsymbol{\alpha} = \mathbf{0}. \tag{34}$$

With this property, we construct $\boldsymbol{\Delta}$ by setting $\boldsymbol{\Delta}_2 = \boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R}$ and $\boldsymbol{\Delta}_1 = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}$.

Now, we show that $\boldsymbol{Z}$ is a strict saddle by arguing that $g(\boldsymbol{Z})$ has a strictly negative curvature along the constructed direction $\boldsymbol{\Delta}$, i.e., $[\nabla^2 g(\boldsymbol{Z})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) < 0$. For this purpose, we compute the three terms in (32) as follows:

$$\|(\boldsymbol{W}_2\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2\boldsymbol{W}_1)\boldsymbol{X}_1\|_F^2 = 0 \tag{35}$$

since $\boldsymbol{W}_2\boldsymbol{\Delta}_1 = \boldsymbol{W}_2\boldsymbol{R}^{\mathrm{T}}\boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}} = \widetilde{\boldsymbol{W}}_2\boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}} = \mathbf{0}$ and $\boldsymbol{\Delta}_2\boldsymbol{W}_1 = \boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{W}_1 = \boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1 = \mathbf{0}$ by utilizing (34);

$$\|\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_2^{\mathrm{T}}\boldsymbol{W}_2 - \boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Delta}_1^{\mathrm{T}} - \boldsymbol{\Delta}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_1^{T}\|_F^2 = 0$$

since it follows from (33) that $\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{\Delta}_2 = \boldsymbol{R}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_2^{\mathrm{T}}\boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R} = \mathbf{0}$ and

$$\begin{aligned}
\boldsymbol{W}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Delta}_1^{\mathrm{T}} &= \boldsymbol{R}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{q}_k\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R} \\
&= \boldsymbol{R}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1\boldsymbol{q}_k\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R} = \mathbf{0};
\end{aligned}$$

and

$$\begin{aligned}
&\left\langle \boldsymbol{\Delta}_2\boldsymbol{\Delta}_1, (\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}\right\rangle \\
&\quad = \left\langle \boldsymbol{p}_k\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}, (\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{Y}\boldsymbol{V})\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right\rangle \\
&\quad = \left\langle \boldsymbol{p}_k\boldsymbol{q}_k^{\mathrm{T}}, \widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1\right\rangle - \left\langle \boldsymbol{p}_k\boldsymbol{q}_k^{\mathrm{T}}, \boldsymbol{Y}\boldsymbol{V}\right\rangle = -\lambda_k,
\end{aligned}$$

where the last equality utilizes (33). Thus, we have

$$\nabla^2 g(\boldsymbol{Z})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] = -2\lambda_k \le -2\lambda_r.$$

We finally obtain (14) by noting that

$$\begin{aligned}\|\boldsymbol{\Delta}\|_F^2 &= \|\boldsymbol{\Delta}_1\|_F^2 + \|\boldsymbol{\Delta}_2\|_F^2\\ &= \|\boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R}\|_F^2 + \|\boldsymbol{R}^{\mathrm{T}}\boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\|_F^2\\ &= 1 + \|\boldsymbol{\Sigma}^{-1}\boldsymbol{q}_k\|_F^2 \le 1 + \|\boldsymbol{\Sigma}^{-1}\|_F^2\|\boldsymbol{q}_k\|_F^2\\ &= 1 + \|\boldsymbol{\Sigma}^{-1}\|_F^2,\end{aligned}$$

where the inequality follows from the Cauchy–Schwartz inequality $|\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}| \le \|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2$.
Case ii: $r > d_1$. In this case, minimizing $\widetilde{g}(\widetilde{\boldsymbol{Z}})$ in (12) is equivalent to finding a low-rank approximation to $\boldsymbol{YV}$. Let Γ denote the indices of the singular vectors $\{\boldsymbol{p}_j\}$ and $\{\boldsymbol{q}_j\}$ that are included in $\widetilde{\boldsymbol{Z}}$, that is,

$$\{\widetilde{z}_i, i \in [d_1]\} = \left\{\boldsymbol{0}, \sqrt{\lambda_j}\begin{bmatrix}\boldsymbol{p}_j\\ \boldsymbol{q}_j\end{bmatrix}, j \in \Gamma\right\}.$$

Then, for any $\widetilde{\boldsymbol{Z}}$, we have

$$\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{YV} = \sum_{i\neq\lambda}\lambda_i\boldsymbol{p}_i\boldsymbol{q}_i$$

and

$$\widetilde{g}(\widetilde{\boldsymbol{Z}}) = \frac{1}{2}\|\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{YV}\|_F^2 = \frac{1}{2}\left(\sum_{i\neq\Lambda}\lambda_i^2\right),$$

which implies that $\widetilde{\boldsymbol{Z}}$ is a global minimum of $\widetilde{g}(\widetilde{\boldsymbol{Z}})$ if

$$\|\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{YV}\|_F^2 = \sum_{i>d_1}\lambda_i^2.$$

To simply the following analysis, we assume $\lambda_{d_1} > \lambda_{d_1+1}$, but the argument is similar in the case of repeated eigenvalues at λ_{d_1} (i.e., $\lambda_{d_1} = \lambda_{d_1+1} = \cdots$). In this case, we know for any $\boldsymbol{Z} \in \mathcal{C} \setminus \mathcal{X}$ that is not a global minimum, there exists $\Omega \subset [r]$ which contains $k \in \Omega, k \le d_1$ such that

$$\boldsymbol{YV} - \widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 = \sum_{i\in\Omega}\lambda_i\boldsymbol{p}_i\boldsymbol{q}_i^{\mathrm{T}}.$$

Similar to Case i, we have

$$\boldsymbol{p}_k^{\mathrm{T}}\widetilde{\boldsymbol{W}}_2 = \boldsymbol{0}, \ \ \widetilde{\boldsymbol{W}}_1\boldsymbol{q}_k = \boldsymbol{0}. \tag{36}$$

Let $\boldsymbol{\alpha} \in \mathbb{R}^{d_1}$ be the eigenvector associated with the smallest eigenvalue of $\widetilde{\boldsymbol{Z}}^{\mathrm{T}}\widetilde{\boldsymbol{Z}}$. By the form of $\widetilde{\boldsymbol{Z}}$ in (10), we have

$$\|\widetilde{\boldsymbol{W}}_2\boldsymbol{\alpha}\|_2^2 = \|\widetilde{\boldsymbol{W}}_1^{\mathrm{T}}\boldsymbol{\alpha}\|_2^2 \le \lambda_{d_1+1}, \tag{37}$$

where the inequality attains equality when $d_1 + 1 \in \Omega$. As in Case i, we construct $\boldsymbol{\Delta}$ by setting $\boldsymbol{\Delta}_2 = \boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R}$ and $\boldsymbol{\Delta}_1 = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}$. We now show that $\boldsymbol{Z}$ is a strict saddle by arguing that $g(\boldsymbol{Z})$ has a strictly negative curvature along the constructed direction $\boldsymbol{\Delta}$ (i.e., $[\nabla^2 g(\boldsymbol{Z})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) < 0$) by computing the three terms in (32) as follows:

$$\begin{aligned}&\|(\boldsymbol{W}_2\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2\boldsymbol{W}_1)\boldsymbol{X}_1\|_F^2\\ &\quad= \left\|\widetilde{\boldsymbol{W}}_2\boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1\boldsymbol{V}^{\mathrm{T}}\right\|_F^2\\ &\quad= \|\widetilde{\boldsymbol{W}}_2\boldsymbol{\alpha}\|_F^2 + \left\|+\boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1\right\|_F^2 + 2\left\langle\widetilde{\boldsymbol{W}}_2\boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}, \boldsymbol{p}_k\boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{W}}_1\right\rangle\\ &\quad\le 2\lambda_{d_1+1},\end{aligned}$$

where the last line follows from (36) and (37);

$$\|(\boldsymbol{W}_2\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2\boldsymbol{W}_1)\boldsymbol{X}_1\|_F^2 = 0$$

holds with a similar argument as in (35); and

$$\begin{aligned}&\left\langle\boldsymbol{\Delta}_2\boldsymbol{\Delta}_1, (\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^{\mathrm{T}}\right\rangle\\ &\quad= \left\langle\boldsymbol{p}_k\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}, (\widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1 - \boldsymbol{YV})\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right\rangle\\ &\quad= \left\langle\boldsymbol{p}_k\boldsymbol{q}_k^{\mathrm{T}}, \widetilde{\boldsymbol{W}}_2\widetilde{\boldsymbol{W}}_1\right\rangle - \left\langle\boldsymbol{p}_k\boldsymbol{q}_k^{\mathrm{T}}, \boldsymbol{YV}\right\rangle\\ &\quad= -\lambda_k \le -\lambda_{d_1},\end{aligned}$$

where the last equality used (36) and the fact that $k \le d_1$. Thus, we have

$$\nabla^2 g(\boldsymbol{Z})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] \le -2(\lambda_{d_1} - \lambda_{d_1+1}),$$

completing the proof of Lemma 2.

References

1. Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., Ma, T.: Finding approximate local minima faster than gradient descent. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pp. 1195–1199. ACM (2017)
2. Baldi, P., Hornik, K.: Neural networks and principal component analysis: learning from examples without local minima. Neural Netw. **2**(1), 53–58 (1989)
3. Bhojanapalli, S., Neyshabur, B., Srebro, N.: Global optimality of local search for low rank matrix recovery. In: Advances in Neural Information Processing Systems, pp. 3873–3881 (2016)
4. Blum, A., Rivest, R.L.: Training a 3-node neural network is NP-complete. In: Advances in Neural Information Processing Systems (NIPS), pp. 494–501 (1989)
5. Borgerding, M., Schniter, P., Rangan, S.: Amp-inspired deep networks for sparse linear inverse problems. IEEE Trans. Signal Process. **65**(16), 4293–4308 (2017)
6. Byrd, R.H., Gilbert, J.C., Nocedal, J.: A trust region method based on interior point techniques for nonlinear programming. Math. Program. **89**(1), 149–185 (2000)

7. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Accelerated methods for non-convex optimization. arXiv preprint arXiv:1611.00756 (2016)
8. Conn, A.R., Gould, N.I., Toint, P.L.: Trust Region Methods. SIAM, Philadelphia (2000)
9. Curtis, F.E., Robinson, D.P.: Exploiting negative curvature in deterministic and stochastic optimization. arXiv preprint arXiv:1703.00412 (2017)
10. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**(4), 303–314 (1989)
11. Du, S.S., Lee, J.D., Tian, Y.: When is a convolutional filter easy to learn? arXiv preprint arXiv:1709.06129 (2017)
12. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points—online stochastic gradient for tensor decomposition. In: Conference on Learning Theory, pp. 797–842 (2015)
13. Ge, R., Jin, C., Zheng, Y.: No spurious local minima in nonconvex low rank problems: a unified geometric analysis. In: International Conference on Machine Learning, pp. 1233–1242 (2017)
14. Ge, R., Lee, J.D., Ma, T.: Matrix completion has no spurious local minimum. In: Advances in Neural Information Processing Systems, pp. 2973–2981 (2016)
15. Haeffele, B.D., Vidal, R.: Global optimality in neural network training. In: Conference on Computer Vision and Pattern Recognition, pp. 7331–7339 (2017)
16. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2012)
17. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
18. Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., Jordan, M.I.: How to escape saddle points efficiently. In: International Conference on Machine Learning, pp. 1724–1732 (2017)
19. Kamilov, U.S., Mansour, H.: Learning optimal nonlinearities for iterative thresholding algorithms. IEEE Signal Process. Lett. **23**(5), 747–751 (2016)
20. Kawaguchi, K.: Deep learning without poor local minima. In: Advances in Neural Information Processing Systems, pp. 586–594 (2016)
21. Laurent, T., von Brecht, J.: Deep linear neural networks with arbitrary loss: all local minima are global. arXiv preprint arXiv:1712.01473 (2017)
22. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
23. Lee, J.D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M.I., Recht, B.: First-order methods almost always avoid saddle points. arXiv preprint arXiv:1710.07406 (2017)
24. Lee, J.D., Simchowitz, M., Jordan, M.I., Recht, B.: Gradient descent only converges to minimizers. In: Conference on Learning Theory, pp. 1246–1257 (2016)
25. Li, Q., Zhu, Z., Tang, G.: The non-convex geometry of low-rank matrix optimization. Inf. Inference: J. IMA **8**, 51–96 (2019)
26. Li, X., Wang, Z., Lu, J., Arora, R., Haupt, J., Liu, H., Zhao, T.: Symmetry, saddle points, and global geometry of nonconvex matrix factorization. arXiv preprint arXiv:1612.09296 (2016)
27. Li, X., Zhu, Z., So, A.M.C., Vidal, R.: Nonconvex robust low-rank matrix recovery. arXiv preprint arXiv:1809.09237 (2018)
28. Li, Y., Zhang, Y., Huang, X., Ma, J.: Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. **99**, 1–16 (2018)
29. Li, Y., Zhang, Y., Huang, X., Yuille, A.L.: Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. ISPRS J. Photogramm. Remote Sens. **146**, 182–196 (2018)
30. Liu, H., Yue, M.C., Man-Cho So, A.: On the estimation performance and convergence rate of the generalized power method for phase synchronization. SIAM J. Optim. **27**(4), 2426–2446 (2017)
31. Lu, H., Kawaguchi, K.: Depth creates no bad local minima. arXiv preprint arXiv:1702.08580 (2017)
32. Mousavi, A., Patel, A.B., Baraniuk, R.G.: A deep learning approach to structured signal recovery. In: 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1336–1343 (2015)
33. Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Math. Program. **108**(1), 177–205 (2006)
34. Nouiehed, M., Razaviyayn, M.: Learning deep models: critical points and local openness. arXiv preprint arXiv:1803.02968 (2018)
35. Park, D., Kyrillidis, A., Carmanis, C., Sanghavi, S.: Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In: Artificial Intelligence and Statistics, pp. 65–74 (2017)
36. Qu, Q., Zhang, Y., Eldar, Y.C., Wright, J.: Convolutional phase retrieval via gradient descent. arXiv preprint arXiv:1712.00716 (2017)
37. Safran, I., Shamir, O.: Spurious local minima are common in two-layer relu neural networks. arXiv preprint arXiv:1712.08968 (2017)
38. Schalkoff, R.J.: Artificial Neural Networks, vol. 1. McGraw-Hill, New York (1997)
39. Soudry, D., Carmon, Y.: No bad local minima: Data independent training error guarantees for multilayer neural networks. arXiv preprint arXiv:1605.08361 (2016)
40. Soudry, D., Hoffer, E.: Exponentially vanishing sub-optimal local minima in multilayer neural networks. arXiv preprint arXiv:1702.05777 (2017)
41. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. Found. Comput. Math. **18**, 1–68 (2018)
42. Sun, J., Qu, Q., Wright, J.: Complete dictionary recovery over the sphere I: overview and the geometric picture. IEEE Trans. Inf. Theory **63**(2), 853–884 (2017)
43. Tian, Y.: An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In: International Conference on Machine Learning, pp. 3404–3413 (2017)
44. Werbos, P.: Beyond regression: new fools for prediction and analysis in the behavioral sciences. Ph.D. thesis, Harvard University (1974)
45. Yun, C., Sra, S., Jadbabaie, A.: Global optimality conditions for deep neural networks. arXiv preprint arXiv:1707.02444 (2017)
46. Zhu, Z., Li, Q., Tang, G., Wakin, M.B.: The global optimization geometry of low-rank matrix optimization. arXiv preprint arXiv:1703.01256 (2017)
47. Zhu, Z., Li, Q., Tang, G., Wakin, M.B.: Global optimality in low-rank matrix optimization. IEEE Trans. Signal Process. **66**(13), 3614–3628 (2018)

Zhihui Zhu is a Postdoctoral Fellow in the Mathematical Institute for Data Science at the Johns Hopkins University. He received his B.Eng. degree in communications engineering in 2012 from Zhejiang University of Technology (Jianxing Honors College), and his Ph.D. degree in electrical engineering in 2017 at the Colorado School of Mines, where his research was awarded a Graduate Research Award. His research interests include exploiting inherent structures and applying optimization methods with guaranteed performance for signal processing, machine learning, and data analysis.

Daniel Soudry Since October 2017, Daniel Soudry is an assistant professor (Taub Fellow) in the Department of Electrical Engineering at the Technion, working in the areas of machine learning and theoretical neuroscience. Before that, he did his post-doc (as a Gruss Lipper fellow) working with Prof. Liam Paninski in the Department of Statistics, the Center for Theoretical Neuroscience the Grossman Center for Statistics of the Mind at Columbia University. He did his Ph.D. in the Department of Electrical Engineering at the Technion, Israel Institute of technology, under the guidance of Prof. Ron Meir. He received his B.Sc. degree in Electrical Engineering and Physics from the Technion.

Yonina C. Eldar received the B.Sc. degree in Physics in 1995 and the B.Sc. degree in Electrical Engineering in 1996 both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in Electrical Engineering and Computer Science in 2002 from the Massachusetts Institute of Technology (MIT), Cambridge. She is currently a Professor in the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. She was previously a Professor in the Department of Electrical Engineering at the Technion, where she held the Edwards Chair in Engineering. She is also a Visiting Professor at MIT, a Visiting Scientist at the Broad Institute, and an Adjunct Professor at Duke University and was a Visiting Professor at Stanford. She is a member of the Israel Academy of Sciences and Humanities (elected 2017), an IEEE Fellow, and a EURASIP Fellow. Her research interests are in the broad areas of statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology and optics. Dr. Eldar has received numerous awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), and the IEEE Kiyo Tomiyasu Award (2016). She was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. She received the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair at the Technion, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion's Award for Excellence in Teaching (twice). She received several best paper awards and best demo awards together with her research students and colleagues including the SIAM outstanding Paper Prize and the IET Circuits, Devices and Systems Premium Award, and was selected as one of the 50 most influential women in Israel. She was a member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education. She is the Editor in Chief of Foundations and Trends in Signal Processing, a member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. In the past, she was a Signal Processing Society Distinguished Lecturer, member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and served as an associate editor for the IEEE Transactions On Signal Processing, the EURASIP Journal of Signal Processing, the SIAM Journal on Matrix Analysis and Applications, and the SIAM Journal on Imaging Sciences. She was Co-Chair and Technical Co-Chair of several international conferences and workshops.

Michael B. Wakin is a Professor of Electrical Engineering at the Colorado School of Mines. Dr. Wakin received a B.S. in electrical engineering and a B.A. in mathematics in 2000 (summa cum laude), an M.S. in electrical engineering in 2002, and a Ph.D. in electrical engineering in 2007, all from Rice University. He was an NSF Mathematical Sciences Postdoctoral Research Fellow at Caltech from 2006–2007, an Assistant Professor at the University of Michigan from 2007–2008, and a Ben L. Fryrear Associate Professor at Mines from 2015–2017. His research interests include signal and data processing using sparse, low-rank, and manifold-based models. In 2007, Dr. Wakin shared the Hershel M. Rich Invention Award from Rice University for the design of a single-pixel camera based on compressive sensing. In 2008, Dr. Wakin received the DARPA Young Faculty Award for his research in compressive multi-signal processing for environments such as sensor and camera networks. In 2012, Dr. Wakin received the NSF CAREER Award for research into dimensionality reduction techniques for structured data sets. In 2014, Dr. Wakin received the Excellence in Research Award for his research as a junior faculty member at Mines. Dr. Wakin is a recipient of the Best Paper Award from the IEEE Signal Processing Society. He has served as an Associate Editor for IEEE Signal Processing Letters and is currently an Associate Editor for IEEE Transactions on Signal Processing.