Characterization of Gradient Dominance and Regularity Conditions for Neural Networks

Yi Zhou zhou.1172@osu.edu
Ohio State University
Yingbin Liang liang.889@osu.edu
Ohio State University

Abstract

The past decade has witnessed a successful application of deep learning to solving many challenging problems in machine learning and artificial intelligence. However, the loss functions of neural networks are still far from being well understood from a theoretical aspect. In this paper, we enrich the current understanding of the landscape of the square loss functions for three types of neural networks, i.e., linear networks, linear residual networks, and one-hidden-layer nonlinear networks. Specifically, when the parameter matrices are square, we establish two quadratic types of landscape properties for the square loss of these neural networks: the gradient dominance condition within the neighborhood of their full rank global minimizers and the regularity condition along certain directions and within the neighborhood of their global minimizers.

1 Introduction

The significant success of deep learning (see, e.g., [2]) has influenced many fields such as machine learning, artificial intelligence, computer vision, natural language processing, etc. Consequently, there is a rising interest in understanding the fundamental properties of deep neural networks. Among them, the landscape (also referred to as geometry) of the loss functions of neural networks is an important aspect, since it is central to determine the performance of optimization algorithms that are designed to minimize these loss functions.

This paper focuses on two important landscape properties for nonconvex optimization. The first property is referred to as *gradient dominance condition* as we describe below. Consider a global minimizer $\mathbf{x}^*$ of a generic function $f : \mathbb{R}^d \to \mathbb{R}$, and a neighborhood $\mathcal{B}_{\mathbf{x}^*}(\delta)$ around $\mathbf{x}^*$. The local gradient dominance condition with regard to $\mathbf{x}^*$ is given by, for some $\lambda > 0$

$$\forall \mathbf{x} \in \mathcal{B}_{\mathbf{x}^*}(\delta),\ f(\mathbf{x}) - f(\mathbf{x}^*) \le \lambda \left\|\nabla f(\mathbf{x})\right\|_2^2,$$

This condition has been shown to hold for a variety of nonconvex machine learning problems, e.g., phase retrieval [11] and blind deconvolution [5]. If the gradient descent algorithm iterates in the neighborhood $\mathcal{B}_{\mathbf{x}^*}(\delta)$, then the gradient dominance condition, together with a Lipschitz property of the gradient of the objective function, guarantees a linear convergence of the function value residual $f(\mathbf{x}) - f(\mathbf{x}^*)$ [4, 7]. The second property is referred to as local *regularity condition*, which is given by, for some $\alpha, \beta > 0$

$$\forall \mathbf{x} \in \mathcal{B}_{\mathbf{x}^*}(\delta),\ \langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x})\rangle \ge \alpha \left\|\nabla f(\mathbf{x})\right\|_2^2 + \beta \left\|\mathbf{x} - \mathbf{x}^*\right\|_2^2,$$

This condition can be viewed as a restricted version of the strong convexity, and it has been shown to guarantee a linear convergence of the iterate residual $\|\mathbf{x} - \mathbf{x}^*\|$ of the gradient descent algorithm [6, 1]. Problems such as phase retrieval [1], affine rank minimization [9, 8] and matrix completion [10] have been shown to satisfy the local regularity condition. This paper studies these two landscape properties for the square loss function of linear, linear residual, and one-hidden-layer nonlinear neural networks under the setting where all parameter matrices are square.

2 Preliminaries of Neural Networks

Throughout, $(\boldsymbol{X}, \boldsymbol{Y})$ denotes the input and output data matrix pair. We assume that $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times m}$, and $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{\boldsymbol{XY}}^\top \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{XY}}$ is full rank with distinct eigenvalues. The kronecker product is denoted as "$\otimes$". For a matrix $\boldsymbol{X}$, its spectral norm is denoted by $\|\boldsymbol{X}\|$. The smallest nonzero singular value is denoted by $\eta_{\min}(\boldsymbol{X})$. For a collection of matrix variables $\boldsymbol{W} := \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_l\}$, we denote $\mathrm{vec}\,(\boldsymbol{W}) := [\mathrm{vec}\,(\boldsymbol{W}_1)^\top \ \ldots \ \mathrm{vec}\,(\boldsymbol{W}_l)^\top]^\top$,

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

where $\mathrm{vec}\,(\boldsymbol{W}_k)$ denotes the vertical stack of the columns of $\boldsymbol{W}_k$. We also denote a collection of natural numbers as $[n] := \{1, \ldots, n\}$.

Consider a feed forward linear neural network with $l-1$ hidden layers, where each layer $k \in [l]$ is parameterized by a matrix $\boldsymbol{W}_k \in \mathbb{R}^{d\times d}$. We consider the square loss of the linear network:

$$h(\boldsymbol{W}) := \tfrac{1}{2}\left\|\boldsymbol{W}_l\boldsymbol{W}_{l-1}\ldots\boldsymbol{W}_1\boldsymbol{X}-\boldsymbol{Y}\right\|_F^2. \tag{1}$$

The linear residual network further introduces the residual structure to the linear network. That is, one adds a shortcut (identity map) for every, say, r hidden layers. Assuming we have in total l residual units. The k-th residual unit is parameterized by the parameters $\boldsymbol{A}_{kq} \in \mathbb{R}^{d\times d}, \forall q \in [r]$, and we denote $\boldsymbol{B}_k := \boldsymbol{I}+\boldsymbol{A}_{kr}\ldots\boldsymbol{A}_{k1}$. We consider the square loss of a linear residual network:

$$f(\boldsymbol{A}) := \tfrac{1}{2}\|(\boldsymbol{I}+\boldsymbol{A}_{lr}\ldots\boldsymbol{A}_{l1})\ldots(\boldsymbol{I}+\boldsymbol{A}_{kr}\ldots\boldsymbol{A}_{k1})\cdots(\boldsymbol{I}+\boldsymbol{A}_{1r}\ldots\boldsymbol{A}_{11})\boldsymbol{X}-\boldsymbol{Y}\|_F^2. \tag{2}$$

Consider a nonlinear neural network with one hidden layer, where the layer parameters satisfy $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathbb{R}^{d\times d}$ and the hidden neurons adopt a differentiable nonlinear activation function $\sigma : \mathbb{R} \to \mathbb{R}$. We consider the square loss of the nonlinear network with one hidden layer:

$$g(\boldsymbol{V}) := \tfrac{1}{2}\left\|\boldsymbol{V}_2\sigma(\boldsymbol{V}_1\boldsymbol{X})-\boldsymbol{Y}\right\|_F^2, \tag{3}$$

where σ acts on $\boldsymbol{V}_1\boldsymbol{X}$ *entrywise*. In particular, we consider a class of activation functions that satisfy the condition $\mathrm{range}(\sigma) = \mathbb{R}$. A typical example of such activation function is the class of parametric ReLU activation functions, i.e., $\sigma(x) = \max\{x, ax\}$, where $0 < a < 1$. The following theorem characterizes a useful property of the global minimizers of these networks.

Theorem 1. *Consider the global minimizers $\boldsymbol{W}^*, \boldsymbol{A}^*, \boldsymbol{V}^*$ of $h(\boldsymbol{W}), f(\boldsymbol{A}), g(\boldsymbol{V})$, respectively. Then for all $k \in [l]$, 1) $\boldsymbol{W}_k^*$ is full rank; 2) $\boldsymbol{B}_k^*$ is full rank; and 3) $\sigma(\boldsymbol{W}_1^*\boldsymbol{X})$ is full rank.*

3 Gradient Dominance Condition

We first establish the local gradient dominance condition for the loss $h(\boldsymbol{W})$ of linear networks.

Theorem 2. *Consider $h(\boldsymbol{W})$ of the linear neural network with $m = d$. Consider a global minimizer $\boldsymbol{W}^*$ and let $\tau = \frac{1}{2}\min_{k\in[l]}\eta_{\min}(\boldsymbol{W}_k^*)$. Then any point in the neighborhood of $\boldsymbol{W}^*$ defined as $\{\boldsymbol{W} : \|\boldsymbol{W}_k - \boldsymbol{W}_k^*\| < \tau,\ \forall k \in [l]\}$ satisfies*

$$h(\boldsymbol{W}) - h(\boldsymbol{W}^*) \le \lambda_h\left\|\nabla_{\mathrm{vec}(\boldsymbol{W})}h(\boldsymbol{W})\right\|_2^2, \text{ where } \lambda_h = (2l\tau^{2(l-1)}\eta_{\min}^2(\boldsymbol{X}))^{-1}. \tag{4}$$

We note that Theorem 1 guarantees that any global minimizer $\boldsymbol{W}^*$ of $h(\boldsymbol{W})$ is full rank, and hence the parameter τ defined in Theorem 2 is strictly positive. The gradient dominance condition implies a linear convergence of the function value to the global minimum via a gradient descent algorithm if the iterations stay in this τ neighborhood. In particular, a larger parameter τ (a larger minimum singular value) implies a smaller λ_h, which yields a faster convergence of the function value to the global minimum via the gradient descent algorithm. Next, we establish the local gradient dominance condition for the loss $f(\boldsymbol{A})$ of linear residual networks.

Theorem 3. *Consider $f(\boldsymbol{A})$ of the linear residual neural network with $m = d$. Consider a full-rank global minimizer $\boldsymbol{A}^*$, and let $\tau = \frac{1}{2}\min_{k\in[l]}\eta_{\min}(\boldsymbol{B}_k^*)$, $\tilde{\tau} = \frac{1}{2}\min_{k\in[l],q\in[r]}\eta_{\min}(\boldsymbol{A}_{kq}^*)$, and pick $\hat{\tau}$ sufficiently small such that any point in the neighborhood of $\boldsymbol{A}^*$ defined as $\{\boldsymbol{A} : \|\boldsymbol{A}_{kq} - \boldsymbol{A}_{kq}^*\| < \hat{\tau}, \forall k \in [l], q \in [r]\}$ satisfies $\|\boldsymbol{B}_k - \boldsymbol{B}_k^*\| < \tau$ for all $k \in [l]$. Then any point in the neighborhood of $\boldsymbol{A}^*$ defined as $\{\boldsymbol{A} : \|\boldsymbol{A}_{kq} - \boldsymbol{A}_{kq}^*\| < \min\{\hat{\tau}, \tilde{\tau}\}, \forall k \in [l], q \in [r]\}$ satisfies*

$$f(\boldsymbol{A}) - f(\boldsymbol{A}^*) \le \lambda_f\left\|\nabla_{\mathrm{vec}(\boldsymbol{A})}f(\boldsymbol{A})\right\|_2^2, \text{ where } \lambda_f = \left(2lr\tilde{\tau}^{2(r-1)}\tau^{2(l-1)}\eta_{\min}^2(\boldsymbol{X})\right)^{-1}. \tag{5}$$

Compare to the gradient dominance condition obtained in [3], which is applicable to the neighborhood of the origin for the residual network with $r = 1$, the above result characterizes the gradient dominance condition in the neighborhood of any full-rank minimizer $\boldsymbol{A}^*$ and to more general residual networks with $r > 1$.

We note that the parameter λ_f in Theorem 3 depends on both τ and $\tilde{\tau}$, where τ captures the overall property of each residual unit and $\tilde{\tau}$ captures the property of individual linear unit in each residual unit. Hence, in general,

the λ_f in Theorem 3 for linear residual networks is very different from the λ_h in the gradient dominance condition in Theorem 2 for linear networks. When the shortcut depth r becomes large, the parameter λ_f involves $\tilde{\tau}$ that depends on more unparameterized variables in $\boldsymbol{A}^*$, and hence becomes more similar to the parameter λ_h of linear networks.

To further compare the λ_f in Theorem 3 and the λ_h in Theorem 2, consider a simplified setting of the linear residual network with the shortcut depth $r = 1$. Then $\lambda_f = \left(2l\tau^{2(l-1)}\eta_{\min}^2(\boldsymbol{X})\right)^{-1}$. Although it takes the same expression as λ_h in Theorem 2 for the linear network, the parameter τ is better regularized since $\boldsymbol{B}_k, k \in [l]$ are further parameterized by $\boldsymbol{I} + \boldsymbol{A}_k$. More specifically, when $\|\boldsymbol{A}_k\| < 1$, $\eta_{\min}(\boldsymbol{I} + \boldsymbol{A}_k)$ (and hence the parameter τ) is regularized away from zero by the identity map, which was also observed by [3]. Consequently, the identity shortcut leads to a smaller λ_f (due to larger τ) compared to a large λ_h when the parameters of linear networks have small spectral norm. Such a smaller λ_f is more desirable for optimization, because the function value approaches closer to the global minimum after one iteration of a gradient descent algorithm. We now establishes the local gradient dominance condition for the loss $g(\boldsymbol{V})$ of nonlinear networks.

Theorem 4. *Consider the loss function $g(\boldsymbol{V})$ of one-hidden-layer nonlinear neural networks with $m = d$ and with $\mathrm{range}(\sigma) = \mathbb{R}$. Consider a global minimizer $\boldsymbol{V}^*$, and let $\tau = \frac{1}{2}\eta_{\min}(\sigma(\boldsymbol{V}_1^*\boldsymbol{X}))$. Then any point in the neighborhood of $\boldsymbol{V}^*$ defined as $\{\boldsymbol{V} : \|\sigma(\boldsymbol{V}_1\boldsymbol{X}) - \sigma(\boldsymbol{V}_1^*\boldsymbol{X})\| \le \tau\}$ satisfies*

$$g(\boldsymbol{V}) - g(\boldsymbol{V}^*) \le \lambda_g \left\|\nabla_{\mathrm{vec}(\boldsymbol{V})} g(\boldsymbol{V})\right\|_2^2, \text{ where } \lambda_g = (2\tau^2)^{-1}. \tag{6}$$

We note that Theorem 1 guarantees that $\sigma(\boldsymbol{V}_1^*\boldsymbol{X})$ is full rank, and hence τ is well defined. Differently from linear networks, the gradient dominance condition for nonlinear networks holds in a nonlinear τ neighborhood that involves the activation function σ. This is naturally due to the nonlinearity of the network. Furthermore, the parameter τ depends on the nonlinear term $\sigma(\boldsymbol{V}_1\boldsymbol{X})$, whereas the τ in Theorem 2 of linear networks depends on the individual parameters $\boldsymbol{W}_k$.

4 Regularity Condition

For the linear network, define matrix $\boldsymbol{G}(\boldsymbol{W}) := [\boldsymbol{G}_1(\boldsymbol{W}), \ldots, \boldsymbol{G}_l(\boldsymbol{W})]$, where for all $k \in [l]$

$$\boldsymbol{G}_k(\boldsymbol{W}) := (\boldsymbol{W}_{k-1}\ldots\boldsymbol{W}_1\boldsymbol{X})^\top \otimes (\boldsymbol{W}_l\ldots\boldsymbol{W}_{k+1}). \tag{7}$$

The following result establishes the local regularity condition for the loss $h(\boldsymbol{W})$ of linear networks.

Theorem 5. *Consider $h(\boldsymbol{W})$ of linear neural networks with $m = d$. Further consider a global minimizer $\boldsymbol{W}^*$, and let $\zeta = 2\max_{k\in[l]}\|\boldsymbol{W}_k^*\|$. Then for any $\delta > 0$, there exists a sufficiently small $\epsilon(\delta)$ such that any point $\boldsymbol{W}$ that satisfies*

$$\|\boldsymbol{G}(\boldsymbol{W}^*)\mathrm{vec}\,(\boldsymbol{W} - \boldsymbol{W}^*)\|_2 \ge \delta\,\|\mathrm{vec}\,(\boldsymbol{W} - \boldsymbol{W}^*)\|_2 \tag{8}$$

and within the neighborhood of $\boldsymbol{W}^*$ defined as $\{\boldsymbol{W} : \|\boldsymbol{W}_k - \boldsymbol{W}_k^*\|_F < \epsilon(\delta), \forall k \in [l]\}$ satisfies

$$\langle \nabla_{\mathrm{vec}(\boldsymbol{W})} h(\boldsymbol{W}), \mathrm{vec}\,(\boldsymbol{W} - \boldsymbol{W}^*)\rangle \ge \alpha \left\|\nabla_{\mathrm{vec}(\boldsymbol{W})} h(\boldsymbol{W})\right\|_2^2 + \beta\,\|\mathrm{vec}\,(\boldsymbol{W} - \boldsymbol{W}^*)\|_2^2, \tag{9}$$

where $\alpha = \gamma/(l\zeta^{2(l-1)}\|\boldsymbol{X}\|^2)$ and $\beta = (1-\gamma)\delta^2/2$ for any $\gamma \in (0,1)$.

We note that the regularity condition has been established for various nonconvex problems [1]. There, the condition was shown to hold within the entire neighborhood of any global minimizer. In comparison, Theorem 5 guarantees the regularity condition for linear networks within a neighborhood of $\boldsymbol{W}^*$ along the directions of $\mathrm{vec}\,(\boldsymbol{W} - \boldsymbol{W}^*)$ that satisfy eq. (8). Furthermore, the parameter δ in eq. (8) determines the range of directions that satisfy eq. (8). For example, if we set $\delta = \eta_{\min}(\boldsymbol{G}(\boldsymbol{W}^*))$, then all $\boldsymbol{W}$ such that $\mathrm{vec}\,(\boldsymbol{W} - \boldsymbol{W}^*) \perp \ker \boldsymbol{G}(\boldsymbol{W}^*)$ satisfy eq. (8).

For all $\boldsymbol{W}$ that satisfy the regularity condition, it can be shown that one gradient descent iteration yields an update that is closer to the global minimizer $\boldsymbol{W}^*$ [1]. Hence, $\boldsymbol{W}^*$ serves as an attractive point along the directions of $\mathrm{vec}\,(\boldsymbol{W} - \boldsymbol{W}^*)$ that satisfy eq. (8). Furthermore, the value of δ in eq. (8) affects the parameter β in the regularity condition. Larger δ results larger β, which further implies that one gradient descent iteration at the point $\boldsymbol{W}$ yields an update that is closer to the global minimizer $\boldsymbol{W}^*$ [1].

For the linear residual network, define $\boldsymbol{Q}(\boldsymbol{A}) := [\boldsymbol{Q}_{11}(\boldsymbol{A}), \ldots, \boldsymbol{Q}_{lr}(\boldsymbol{A})]$, where for all $k \in [l], q \in [r]$

$$\boldsymbol{Q}_{kq}(\boldsymbol{A}) := \left[(\boldsymbol{B}_{k-1}\ldots\boldsymbol{B}_1\boldsymbol{X})^\top \otimes (\boldsymbol{B}_l\ldots\boldsymbol{B}_{k+1})\right]\left[\left(\boldsymbol{A}_{k(q-1)}\ldots\boldsymbol{A}_{k1}\right)^\top \otimes \left(\boldsymbol{A}_{kr}\ldots\boldsymbol{A}_{k(q+1)}\right)\right].$$

We then establish the local regularity condition for the loss $f(\boldsymbol{A})$ of linear residual networks.

Theorem 6. *Consider $f(\boldsymbol{A})$ of linear residual neural networks with $m = d$. Further consider a global minimizer $\boldsymbol{A}^*$, and let $\zeta = 2\max_{k\in[l]} \|\boldsymbol{B}_k^*\|$ and $\tilde{\zeta} = 2\max_{k\in[l],q\in[r]} \|\boldsymbol{A}_{kq}^*\|$. Then, for any constant $\delta > 0$, there exists a sufficiently small $\epsilon(\delta)$ such that any point $\boldsymbol{A}$ that satisfies*

$$\|\boldsymbol{Q}(\boldsymbol{A}^*)\mathrm{vec}\,(\boldsymbol{A}-\boldsymbol{A}^*)\|_2 \geq \delta \|\mathrm{vec}\,(\boldsymbol{A}-\boldsymbol{A}^*)\|_2$$

and within the neighborhood defined as $\{\boldsymbol{A} : \|\boldsymbol{A}_{kq} - \boldsymbol{A}_{kq}^*\|_F < \epsilon(\delta), \forall k \in [l], q \in [r]\}$, satisfies

$$\left\langle \nabla_{\mathrm{vec}(\boldsymbol{A})} f(\boldsymbol{A}), \mathrm{vec}\,(\boldsymbol{A}-\boldsymbol{A}^*)\right\rangle \geq \alpha \left\|\nabla_{\mathrm{vec}(\boldsymbol{A})} f(\boldsymbol{A})\right\|_2^2 + \beta \|\mathrm{vec}\,(\boldsymbol{A}-\boldsymbol{A}^*)\|_2^2, \tag{10}$$

where $\alpha = \gamma/(lr\tilde{\zeta}^{2(r-1)}\zeta^{2(l-1)}\|\boldsymbol{X}\|^2)$ and $\beta = (1-\gamma)\delta^2/2$ with any $\gamma \in (0,1)$.

Similarly to the regularity condition for linear networks, the regularity condition for linear residual networks holds along the directions of $\mathrm{vec}\,(\boldsymbol{A}-\boldsymbol{A}^*)$ that depends on $\boldsymbol{Q}(\boldsymbol{A}^*)$. However, the parametrization of $\boldsymbol{Q}(\boldsymbol{A}^*)$ is different from that of $\boldsymbol{G}(\boldsymbol{W}^*)$ of linear networks. To illustrate, consider a simplified setting where the shortcut depth is $r = 1$. Then, $\boldsymbol{Q}_k(\boldsymbol{A}^*) = \left(\boldsymbol{B}_{k-1}^* \ldots \boldsymbol{B}_1^* \boldsymbol{X}\right)^\top \otimes \left(\boldsymbol{B}_l^* \ldots \boldsymbol{B}_{k+1}^*\right)$. Although it takes a similar form as $\boldsymbol{G}(\boldsymbol{W}^*)$ of the linear network, the reparameterization $\boldsymbol{B}_k^* = \boldsymbol{I} + \boldsymbol{A}_k^*$ keeps $\eta_{\min}(\boldsymbol{B}_k^*)$ away from zero when $\|\boldsymbol{A}_k^*\|$ is small. This enlarges $\eta_{\min}(\boldsymbol{Q}_k(\boldsymbol{A}^*))$ so that the direction constraint can be satisfied along a wider range of directions. In this way, $\boldsymbol{A}^*$ attracts the optimization iteration to converge along a wider range of directions in the neighborhood of the origin.

For the nonlinear network, define matrix

$$\boldsymbol{H} = \left[(\boldsymbol{I}\otimes\boldsymbol{V}_2^*)\sigma'(\mathrm{diag}(\mathrm{vec}\,(\boldsymbol{V}_1^*\boldsymbol{X})))(\boldsymbol{X}^\top\otimes\boldsymbol{I}),\ \sigma(\boldsymbol{V}_1^*\boldsymbol{X})^\top\otimes\boldsymbol{I}\right]$$

We now establish the local regularity condition for the loss $g(\boldsymbol{V})$ of nonlinear networks.

Theorem 7. *Consider $g(\boldsymbol{V})$ of one-hidden-layer nonlinear neural networks with $m = d$ and $\mathrm{range}(\sigma) = \mathbb{R}$. Further consider a global minimizer $\boldsymbol{V}^*$ of $g(\boldsymbol{V})$, and let $\zeta = 2\max\{\|\sigma(\boldsymbol{V}_1^*\boldsymbol{X})\|, \|\boldsymbol{V}_2^*\|, \|\sigma'(\boldsymbol{V}_1^*\boldsymbol{X})\|_\infty\}$. Then there exists a sufficiently small $\epsilon(\delta)$ such that any point $\boldsymbol{V}$ that satisfies*

$$\|\boldsymbol{H}(\boldsymbol{V}^*)\mathrm{vec}\,(\boldsymbol{V}-\boldsymbol{V}^*)\|_2 \geq \delta\|\mathrm{vec}\,(\boldsymbol{V}-\boldsymbol{V}^*)\|_2$$

and within the neighborhood of $\boldsymbol{V}^*$ defined as $\{\boldsymbol{V} : \|\boldsymbol{V}_k - \boldsymbol{V}_k^*\|_F < \epsilon(\delta), \forall k \in [2]\}$ satisfies

$$\langle \nabla_{\mathrm{vec}(\boldsymbol{V})} g(\boldsymbol{V}), \mathrm{vec}\,(\boldsymbol{V}-\boldsymbol{V}^*)\rangle \geq \alpha \left\|\nabla_{\mathrm{vec}(\boldsymbol{V})} g(\boldsymbol{V})\right\|_2^2 + \beta\|\mathrm{vec}\,(\boldsymbol{V}-\boldsymbol{V}^*)\|_2^2, \tag{11}$$

where $\alpha = \gamma/\max\{\|\boldsymbol{X}\|^2\zeta^4, \zeta^2\}$ and $\beta = (1-\gamma)\delta^2/2$ for any $\gamma \in (0,1)$.

Thus, nonlinear neural networks with one hidden layer also have an amenable landscape near the global minimizers that attracts gradient iterates to converge along the directions restricted by $\boldsymbol{H}(\boldsymbol{V}^*)$.

5 Conclusion

In this paper, we establish the gradient dominance condition and the regularity condition for three types of neural networks in the neighborhood of their global minimizers under certain conditions. It is interesting to exploit these conditions in the convergence analysis of optimization algorithms applied to deep learning networks.

References

[1] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, April 2015.

[2] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[3] M. Hardt and T. Ma. Identity matters in deep learning. *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[4] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*, pages 795–811, 2016.

[5] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Arxiv: 1606.04933*, 2016. URL http://arxiv.org/abs/1606.04933.

[6] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2014.

[7] S. Reddi, A. Hefny, S. Sra, B. Póczós, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *Proc. International Conference on Machine Learning (ICML)*, pages 314–323, 2016.

[8] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proc. International Conference on Machine Learning (ICML)*, pages 964–973, 2016.

[9] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, pages 109–117, 2015.

[10] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *ArXiv: 1605.07051*, 2016.

[11] Y. Zhou, H. Zhang, and Y. Liang. Geometrical properties and accelerated gradient solvers of non-convex phase retrieval. In *Proc. Annual Allerton Conference on Communication, Control, and Computing*. 2016.