Deep Neural Networks with Multi-Branch Architectures Are Intrinsically Less Non-Convex

Hongyang Zhang
Carnegie Mellon University
hongyanz@cs.cmu.edu

Junru Shao
Carnegie Mellon University
junrus@cs.cmu.edu

Ruslan Salakhutdinov
Carnegie Mellon University
rsalakhu@cs.cmu.edu

Abstract

Several recently proposed architectures of neural networks such as ResNeXt, Inception, Xception, SqueezeNet and Wide ResNet are based on the designing idea of having multiple branches and have demonstrated improved performance in many applications. We show that one cause for such success is due to the fact that the multi-branch architecture is intrinsically less non-convex in terms of duality gap. The duality gap measures the degree of *intrinsic* non-convexity of an optimization problem: smaller gap in relative value implies lower degree of *intrinsic* non-convexity. The challenge is to quantitatively measure the duality gap of highly non-convex problems such as deep neural networks. In this work, we provide strong guarantees of this quantity for two classes of network architectures. For the neural networks with *arbitrary activation functions*, multi-branch architecture and a variant of hinge loss, we show that the duality gap of both population and empirical risks shrinks to zero as the number of branches increases. This result sheds light on better understanding the power of over-parametrization where increasing the number of branches tends to make the loss surface less non-convex. For the neural networks with linear activation function and ℓ_2 loss, we show that the duality gap of empirical risk is zero. Our two results work for *arbitrary depths*, while the analytical techniques might be of independent interest to non-convex optimization more broadly. Experiments on both synthetic and real-world datasets validate our results.

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

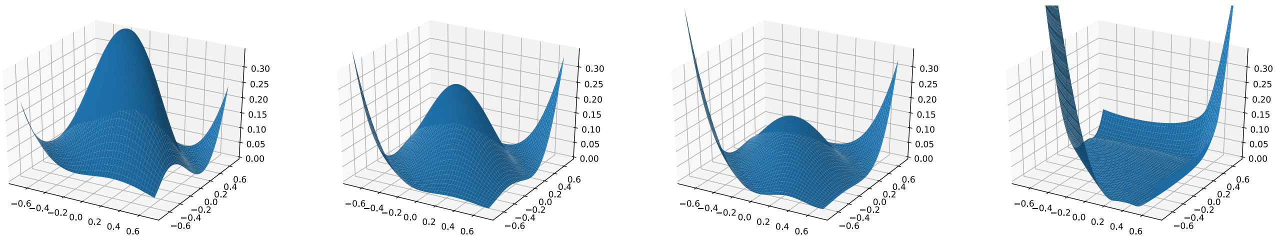
1 Introduction

Deep neural networks are a central object of study in machine learning, computer vision, and many other domains. They have substantially improved over conventional learning algorithms in many areas, including speech recognition, object detection, and natural language processing [28]. The focus of this work is to investigate the duality gap of deep neural networks. The duality gap is the discrepancy between the optimal values of primal and dual problems. While it has been well understood for convex optimization, little is known for non-convex problems. A smaller duality gap in relative value typically implies that the problem itself is less non-convex intrinsically, and thus is easier to optimize.[1] Our results establish that: *Deep neural networks with multi-branch architecture have small duality gap in relative value.*

Our study is motivated by the computational difficulties of deep neural networks due to its non-convex nature. While many works have witnessed the power of local search algorithms for deep neural networks [16], these algorithms typically converge to a suboptimal solution in the worst cases according to various empirical observations [52, 28]. It is reported that for a single-hidden-layer neural network, when the number of hidden units is small, stochastic gradient descent may get easily stuck at the poor local minima [27, 49]. Furthermore, there is significant evidence indicating that when the networks are deep enough, bad saddle points do exist [1] and might be hard to escape [15, 21, 10, 1].

Given the computational obstacles, several efforts have been devoted to designing new architectures to alleviate the above issues, including over-parametrization [17, 53, 23, 41, 2, 45] and multi-branch architectures [64, 56, 18, 61, 33, 58]. Empirically, increasing the number of hidden units of a single-hidden-layer network

[1] Throughout the paper, we discuss the duality gap w.r.t. the Lagrangian function, rather than the augmented Lagrangian function as in Chapter 11 of [48] where the duality gap is always zero.



(a) $I = 10$. (b) $I = 30$. (c) $I = 70$. (d) $I = 1,000$.

Figure 1: The loss surface of one-hidden-layer ReLU network projected onto a 2-d plane, which is spanned by three points to which the SGD algorithm converges according to three different initialization seeds. It shows that as the number of hidden neurons I increases, the landscape becomes less non-convex.

encourages the first-order methods to converge to a global solution, which probably supports the folklore that the loss surface of a wider network looks more "convex" (see Figure 1). Furthermore, several recently proposed architectures, including ResNeXt [61], Inception [56], Xception [18], SqueezeNet [33] and Wide ResNet [62] are based on having multiple branches and have demonstrated substantial improvement over many of the existing models in many applications. In this work, we show that one cause for such success is due to the fact that the loss of multi-branch network is less non-convex in terms of duality gap.

Why is duality gap a measure of *intrinsic* non-convexity? Although some highly non-convex problems such as PCA and quadratic programming may have small/zero duality gap, we argue that the duality gap is a measure of *intrinsic* non-convexity of an optimization problem. There are two reasons for such an argument. a) The optimal value of the dual problem is equal to the optimal value of the convex relaxation of the primal problem. Hereby, the convex relaxation is the problem arising by replacing the non-convex objective with its convex closure and replacing the non-convex feasible set with its closed convex hull. Therefore, the duality gap measures the discrepancy between the optimal values of primal problem and its convex relaxation (Taking convex problems as an example, the duality gap is zero in most cases). When the duality gap is small, one can solve the convex relaxation problem whose solution is guaranteed to being close to the solution of primal problem. b) We show in our main result that the duality gap is a lower bound of the discrepancy between objective and its convex relaxation[2] (see Theorem 1 for the case of $I = 1$). So a smaller duality gap implies a possibly smaller discrepancy between objective and its convex relaxation.

Our Contributions. This paper provides both theoretical and experimental results for the population and empirical risks of deep neural networks by estimating the duality gap.

- We study the duality gap of deep neural networks with *arbitrary* activation functions, *any* data distribution, and multi-branch architecture (see Theorem 1). The multi-branch architecture is general, which includes the classic one-hidden-layer architecture as a special case (see Figure 2). By Shapley-Folkman lemma, we show that the duality gap of both population and empirical risks shrinks to zero as the number of branches increases.
- We prove that the strong duality (a.k.a. zero duality gap) holds for the empirical risk of deep linear neural networks (see Theorem 2). To this end, we develop multiple new proof techniques, including *reduction to low-rank approximation* and *construction of dual certificate* (see Section 4), which might be of independent interest to other non-convex problems.
- We empirically study the loss surface of multi-branch neural networks. Our experiments verify our theoretical findings.

Notation. We will use bold capital letter to represent matrix and lower-case letter to represent scalar. Specifically, let $\mathbf{I}$ be the identity matrix and denote by $\mathbf{0}$ the all-zero matrix. Let $\{\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}} : i = 1, 2, ..., H\}$ be a set of network parameters, each of which represents the connection weights between the i-th and $(i+1)$-th layers of neural network. We use $\mathbf{W}_{:,t} \in \mathbb{R}^{n_1 \times 1}$ to indicate the t-th column of $\mathbf{W}$. We will use $\sigma_i(\mathbf{W})$ to represent the i-th largest singular value of matrix $\mathbf{W}$. Given skinny SVD $\mathbf{U\Sigma V}^T$ of matrix $\mathbf{W}$, we denote by $\mathsf{svd}_r(\mathbf{W}) = \mathbf{U}_{:,1:r}\mathbf{\Sigma}_{1:r,1:r}\mathbf{V}_{:,1:r}^T$ the truncated SVD of $\mathbf{W}$ to the first r singular values. For matrix norms, denote by $\|\mathbf{W}\|_{\mathcal{S}_H} = \left(\sum_i \sigma_i^H(\mathbf{W})\right)^{1/H}$ the matrix Schatten-H norm. Nuclear norm and Frobenius norm are special cases of Schatten-H norm: $\|\mathbf{W}\|_* = \|\mathbf{W}\|_{\mathcal{S}_1}$ and $\|\mathbf{W}\|_F = \|\mathbf{W}\|_{\mathcal{S}_2}$. We use $\|\mathbf{W}\|$ to represent the

[2] Note that the convex relaxation of objective is different from the convex relaxation of primal non-convex problem which requires convexification operation on both objective and constraint.

matrix operator norm, i.e., $\|\mathbf{W}\| = \sigma_1(\mathbf{W})$, and denote by $\mathsf{rank}(\mathbf{W})$ the rank of matrix $\mathbf{W}$. Denote by $\mathsf{Row}(\mathbf{W})$ the span of rows of $\mathbf{W}$. Let $\mathbf{W}^\dagger$ be the Moore-Penrose pseudo-inverse of $\mathbf{W}$.

For convex matrix function $K(\cdot)$, we denote by $K^*(\mathbf{\Lambda}) = \max_{\mathbf{M}} \langle \mathbf{\Lambda}, \mathbf{M} \rangle - K(\mathbf{M})$ the conjugate function of $K(\cdot)$ and $\partial K(\cdot)$ the sub-differential. We use $\mathsf{diag}(\sigma_1, ..., \sigma_r)$ to represent a $r \times r$ diagonal matrix with diagonal entries $\sigma_1, ..., \sigma_r$. Let $d_{\min} = \min\{d_i : i = 1, 2, ..., H-1\}$, and $[I] = \{1, 2, ..., I\}$. For any two matrices $\mathbf{A}$ and $\mathbf{B}$ of matching dimensions, we denote by $[\mathbf{A}, \mathbf{B}]$ the concatenation of $\mathbf{A}$ and $\mathbf{B}$ along the row and $[\mathbf{A}; \mathbf{B}]$ the concatenation of two matrices along the column.

2 Duality Gap of Multi-Branch Neural Networks

We first study the duality gap of neural networks in a classification setting. We show that the wider the network is, the smaller the duality gap becomes.

Network Setup. The output of our network follows from a multi-branch architecture (see Figure 2):

$$f(\mathbf{w}; \mathbf{x}) = \frac{1}{I} \sum_{i=1}^{I} f_i(\mathbf{w}_{(i)}; \mathbf{x}), \ \mathbf{w}_{(i)} \in \mathcal{W}_i,$$

where $\mathcal{W}_i$ is a convex set, $\mathbf{w}$ is the concatenation of all network parameters $\{\mathbf{w}_{(i)}\}_{i=1}^I$, $\mathbf{x} \in \mathbb{R}^{d_0}$ is the input instance, $\{\mathcal{W}_i\}_{i=1}^I$ is the parameter space, and $f_i(\mathbf{w}_{(i)}; \cdot)$ represents an $\mathbb{R}^{d_0} \to \mathbb{R}$ continuous mapping by a sub-network which is allowed to have *arbitrary* architecture such as convolutional and recurrent neural networks. As an example, $f_i(\mathbf{w}_{(i)}; \cdot)$ can be in the form of a H_i-layer feed-forward sub-network:

$$f_i(\mathbf{w}_{(i)}; \mathbf{x}) = \mathbf{w}_i^\top \psi_{H_i}(\mathbf{W}_{H_i}^{(i)} ... \psi_1(\mathbf{W}_1^{(i)} \mathbf{x})) \in \mathbb{R},$$
$$\mathbf{w}_{(i)} = [\mathbf{w}_i; \mathsf{vec}(\mathbf{W}_1^{(i)}); ...; \mathsf{vec}(\mathbf{W}_{H_i}^{(i)})] \in \mathbb{R}^{p_i}.$$

Hereby, the functions $\psi_k(\cdot), k = 1, 2, ..., H_i$ are allowed to encode *arbitrary* form of continuous element-wise non-linearity (and linearity) after each matrix multiplication, such as sigmoid, rectification, convolution, while the number of layers H_i in each sub-network can be *arbitrary* as well. When $H_i = 1$ and $d_{H_i} = 1$, i.e., each sub-network in Figure 2 represents one hidden unit, the architecture $f(\mathbf{w}; \mathbf{x})$ reduces to a one-hidden-layer network. We apply the so-called τ*-hinge loss* [4, 7] on the top of network output for label $y \in \{-1, +1\}$:

$$\ell_\tau(\mathbf{w}; \mathbf{x}, y) := \max\left(0, 1 - \frac{y \cdot f(\mathbf{w}; \mathbf{x})}{\tau}\right), \quad \tau > 0. \tag{1}$$

The τ-hinge loss has been widely applied in active learning of classifiers and margin based learning [4, 7].

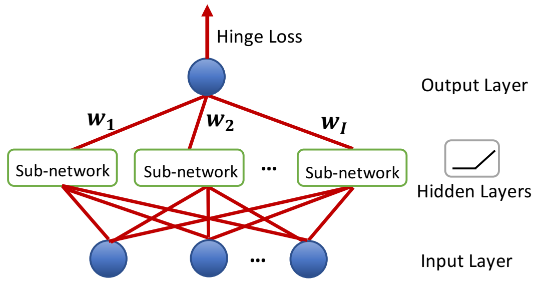


Figure 2: Multi-branch architecture, where the sub-networks are allowed to have arbitrary architectures, depths, and continuous activation functions. Hereby, I represents the number of branches. In the extreme case when the sub-network is chosen to have a single neuron, the multi-branch architecture reduces to a single-hidden-layer neural network and the I represents the network width.

When $\tau = 1$, it reduces to the classic hinge loss [43, 17, 38].

We make the following assumption on the margin parameter τ, which states that the parameter τ is sufficiently large.

Assumption 1 (Parameter τ)**.** *For sample* $(\mathbf{x}, y)$ *drawn from distribution* $\mathcal{P}$*, we have* $\tau > y \cdot f(\mathbf{w}; \mathbf{x})$ *for all* $\mathbf{w} \in \mathcal{W}_1 \times \mathcal{W}_2 \times ... \times \mathcal{W}_I$ *with probability measure* 1*.*

We further empirically observe that using smaller values of the parameter τ and other loss functions support our theoretical result as well (see experiments in Section 5). It is an interesting open question to extend our theory to more general losses in the future.

To study how close these generic neural network architectures approach the family of convex functions, we analyze the duality gap of minimizing the risk w.r.t. the loss (1) with an extra regularization constraint. The normalized duality gap is a measure of intrinsic non-convexity of a given function [13]: the gap is zero when the given function itself is convex, and is large when the loss surface is far from the convexity intrinsically. Typically, the closer the network approaches to the family of convex functions, the easier we can optimize the network.

Multi-Branch Architecture. Our analysis of multi-branch neural networks is built upon tools from non-convex geometric analysis — Shapley–Folkman lemma. Basically, the Shapley–Folkman lemma states that the sum of constrained non-convex functions is close to being convex. A neural network is an ideal target to apply this lemma to: the width of network is associated with the number of summand functions. So intuitively, the wider the neural network is, the smaller the duality gap will be. In particular, we study the following non-convex problem concerning the population risk:

$$\min_{\mathbf{w}\in\mathcal{W}_1\times\ldots\times\mathcal{W}_I} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_\tau(\mathbf{w};\mathbf{x},y)],$$
$$\text{s.t.}\ \ \frac{1}{I}\sum_{i=1}^{I} h_i(\mathbf{w}_{(i)}) \le K, \tag{2}$$

where $h_i(\cdot), i\in[I]$ are convex regularization functions, e.g., the weight decay, and K can be arbitrary such that the problem is feasible. Correspondingly, the dual problem of problem (2) is a one-dimensional convex optimization problem:[3]

$$\max_{\lambda\ge 0} \mathcal{Q}(\lambda) - \lambda K,$$
$$\mathcal{Q}(\lambda) := \inf_{\mathbf{w}\in\mathcal{W}_1\times\ldots\times\mathcal{W}_I} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_\tau(\mathbf{w};\mathbf{x},y)] + \frac{\lambda}{I}\sum_{i=1}^{I} h_i(\mathbf{w}_{(i)}). \tag{3}$$

Before proceeding, we first define some notations to be used in our main results. For $\widetilde{\mathbf{w}}\in\mathcal{W}_i$, denote by

$$\widetilde{f}_i(\widetilde{\mathbf{w}}) := \inf_{a^j,\mathbf{w}_{(i)}^j\in\mathcal{W}_i} \sum_{j=1}^{p_i+2} a^j \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1-\frac{y\cdot f_i(\mathbf{w}_{(i)}^j;\mathbf{x})}{\tau}\right)$$
$$\text{s.t.}\quad \widetilde{\mathbf{w}} = \sum_{j=1}^{p_i+2} a^j\mathbf{w}_{(i)}^j, \sum_{j=1}^{p_i+2} a^j = 1, a^j\ge 0.$$

This represents the convex relaxation of the i-th summand term $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[1-y\cdot f_i(\cdot;\mathbf{x})/\tau]$ in the objective, because the epigraph of $\widetilde{f}_i$ is exactly the convex hull of epigraph of $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[1-y\cdot f_i(\cdot;\mathbf{x})/\tau]$ by the definition of $\widetilde{f}_i$. For $\widetilde{\mathbf{w}}\in\mathcal{W}_i$, we also define

$$\widehat{f}_i(\widetilde{\mathbf{w}}) := \inf_{\mathbf{w}_{(i)}\in\mathcal{W}_i} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1-\frac{y\cdot f_i(\mathbf{w}_{(i)};\mathbf{x})}{\tau}\right)$$
$$\text{s.t.}\quad h_i(\mathbf{w}_{(i)}) \le h_i(\widetilde{\mathbf{w}}).$$

This is a "restricted" version of the i-th summand term $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[1-y\cdot f_i(\mathbf{w}_{(i)};\mathbf{x})/\tau]$ to the hard constraint $h_i(\mathbf{w}_{(i)}) \le h_i(\widetilde{\mathbf{w}})$.

Our main results for multi-branch neural networks are as follows:

Theorem 1. *Denote by* $\inf(\mathbf{P})$ *the minimum of primal problem* (2) *and* $\sup(\mathbf{D})$ *the maximum of dual problem* (3). *Let* $\Delta_i := \sup_{\mathbf{w}\in\mathcal{W}_i}\left\{\widehat{f}_i(\mathbf{w}) - \widetilde{f}_i(\mathbf{w})\right\} \ge 0$ *and* $\Delta_{worst} := \max_{i\in[I]}\Delta_i$. *Suppose* $\mathcal{W}_i$*'s are compact and both* $f_i(\mathbf{w}_{(i)};\mathbf{x})$ *and* $h_i(\mathbf{w}_{(i)})$ *are continuous w.r.t.* $\mathbf{w}_{(i)}$. *If there exists at least one feasible solution of problem* $(\mathbf{P})$*, then under Assumption 1 the duality gap w.r.t. problems* (2) *and* (3) *can be bounded by*

$$0 \le \frac{\inf(\mathbf{P}) - \sup(\mathbf{D})}{\Delta_{worst}} \le \frac{2}{I}.$$

Remark 1. *Note that* Δ_i *measures the divergence between the function value of* $\widehat{f}_i$ *and its convex relaxation* $\widetilde{f}_i$. *The constant* Δ_{worst} *is the maximal divergence among all sub-networks, which grows slowly with the increase of* I. *This is because* Δ_{worst} *only measures the divergence of* one branch. *The normalized duality gap* $(\inf(\mathbf{P}) - \sup(\mathbf{D}))/\Delta_{worst}$ *has been widely used before to measure the degree of non-convexity of optimization problems [13, 57, 14, 24, 22]. Such a normalization avoids trivialities in characterizing the degree of non-convexity: scaling the objective function by any constant does not change the value of normalized duality gap.*

Remark 2. *Even though Theorem 1 is in the form of population risk, the conclusion still holds for the* empirical loss *as well. This can be achieved by setting the marginal distribution* $\mathcal{P}_\mathbf{x}$ *as the uniform distribution on a finite set and* $\mathcal{P}_y$ *as the corresponding labels uniformly distributed on the same finite set.*

Remark 3. *Setting* K *in problem* (2) *infinitely large implies that Theorem 1 holds for unconstrained deep neural networks as well.*

Inspiration for Architecture Designs. Theorem 1 shows that the duality gap of deep network shrinks when the width I is large; when $I\to+\infty$, surprisingly, deep network is as easy as a convex optimization, as the gap is zero. An intuitive explanation is that the large number of randomly initialized hidden units represent all possible features. Thus the optimization problem involves just training the top layer of the network, which is convex. Our result encourages a class of network architectures with multiple branches and supports some of the most successful architectures in practice, such as Inception [56], Xception [18], ResNeXt [61], SqueezeNet [33], Wide ResNet [62], Shake-Shake regularization [25] — all of which benefit from the split-transform-merge behaviour as shown in Figure 2.

Related Works. While many efforts have been devoted to studying the local minima or saddle points of deep neural networks [42, 68, 54, 36, 60, 59], little is known about the duality gap of deep networks. In particular, Choromanska et al. [20, 19] showed that the number of poor local minima cannot be too large. Kawaguchi [35] improved over the results of [20, 19] by assuming that the activation functions are independent Bernoulli variables and the input data are drawn from Gaussian distribution. Xie et al. [60] and Haeffele et al. [30] studied the local minima of regularized network, but they require either the network is shallow, or

[3] Although problem (3) is convex, it does not necessarily mean the problem can be solved easily. This is because computing $\mathcal{Q}(\lambda)$ is a hard problem. So rather than trying to solve the convex dual problem, our goal is to study the duality gap in order to understand the degree of non-convexity of the problem.

the network weights are rank-deficient. Ge et al. [27] showed that every local minimum is globally optimal by modifying the activation function. Zhang et al. [67] and Aslan et al. [3] reduced the non-linear activation to the linear case by kernelization and relaxed the non-convex problem to a convex one. However, no formal guarantee was provided for the tightness of the relaxation. Theorem 1, on the other hand, bounds the duality gap of deep neural networks with mild assumptions.

Another line of research studies the convexity behaviour of neural networks when the number of hidden neurons goes to the infinity. In particular, Bach [5] proved that a single-hidden-layer network is as easy as a convex optimization by using classical non-Euclidean regularization tools. Bengio et al. [12] showed a similar phenomenon for multi-layer networks with an incremental algorithm. In comparison, Theorem 1 not only captures the convexification phenomenon when $I \to +\infty$, but also goes beyond the result as it characterizes the convergence rate of convexity of neural networks in terms of duality gap. Furthermore, the conclusion in Theorem 1 holds for the population risk, which was unknown before.

3 Strong Duality of Linear Neural Networks

In this section, we show that the duality gap is zero if the activation function is linear. Deep linear neural network has received significant attention in recent years [51, 35, 67, 44, 8, 28, 31, 9] because of its simple formulation[4] and its connection to non-linear neural networks.

Network Setup. We discuss the strong duality of regularized deep linear neural networks of the form

$$\begin{aligned}(\mathbf{W}_1^*, ..., \mathbf{W}_H^*) = \operatorname*{argmin}_{\mathbf{W}_1,...,\mathbf{W}_H} \frac{1}{2}\|\mathbf{Y} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 \\ + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right],\end{aligned} \tag{4}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d_0\times n}$ is the given instance matrix, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{d_H\times n}$ is the given label matrix, and $\mathbf{W}_i \in \mathbb{R}^{d_i\times d_{i-1}}, i \in [I]$ represents the weight matrix in each linear layer. We mention that (a) while the linear operation is simple matrix multiplications in problem (4), it can be easily extended to other linear operators, e.g., the convolutional operator or the linear operator with the bias term, by properly involving a group of kernels in the variable $\mathbf{W}_i$ [30]. (b) The regularization terms in problem (4) are of common interest, e.g., see [30]. When $H = 2$, our regularization terms reduce to $\frac{1}{2}\|\mathbf{W}_i\|_F^2$, which is well known as the weight-decay or Tikhonov regularization. (c) The regularization parameter γ is the same for each layer since we have no further information on the preference of layers.

Our analysis leads to the following guarantees for the deep linear neural networks.

Theorem 2. *Denote by $\widetilde{\mathbf{Y}} := \mathbf{Y}\mathbf{X}^\dagger\mathbf{X} \in \mathbb{R}^{d_H\times n}$ and $d_{\min} := \min\{d_1, ..., d_{H-1}\} \le \min\{d_0, d_H, n\}$. Let $0 \le \gamma < \sigma_{\min}(\widetilde{\mathbf{Y}})$ and $H \ge 2$, where $\sigma_{\min}(\widetilde{\mathbf{Y}})$ stands for the minimal non-zero singular value of $\widetilde{\mathbf{Y}}$. Then the strong duality holds for deep linear neural network (4). In other words, the optimum of problem (4) is the same as its convex dual problem*

$$\begin{aligned}\mathbf{\Lambda}^* = \operatorname*{argmax}_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} -\frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}\|_{d_{\min}}^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2, \\ \text{s.t.} \quad \|\mathbf{\Lambda}\| \le \gamma,\end{aligned} \tag{5}$$

where $\|\cdot\|_{d_{\min}}^2 = \sum_{i=1}^{d_{\min}} \sigma_i^2(\cdot)$ is a convex function. Moreover, the optimal solutions of primal problem (4) can be obtained from the dual problem (5) in the following way: let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$ be the skinny SVD of matrix $\mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$, then $\mathbf{W}_i^* = [\mathbf{\Sigma}^{1/H}, \mathbf{0}; \mathbf{0}, \mathbf{0}] \in \mathbb{R}^{d_i\times d_{i-1}}$ for $i = 2, 3, ..., H-1$, $\mathbf{W}_H^* = [\mathbf{U}\mathbf{\Sigma}^{1/H}, \mathbf{0}] \in \mathbb{R}^{d_H\times d_{H-2}}$ and $\mathbf{W}_1^* = [\mathbf{\Sigma}^{1/H}\mathbf{V}^T; \mathbf{0}]\mathbf{X}^\dagger \in \mathbb{R}^{d_1\times d_0}$ is a globally optimal solution to problem (4).

The regularization parameter γ cannot be too large in order to avoid underfitting. Our result provides a suggested upper bound $\sigma_{\min}(\widetilde{\mathbf{Y}})$ for the regularization parameter, where oftentimes $\sigma_{\min}(\widetilde{\mathbf{Y}})$ characterizes the level of random noise. When $\gamma = 0$, our analysis reduces to the *un-regularized deep linear neural network*, a model which has been widely studied in [35, 44, 8, 28].

Theorem 2 implies the followig result on the landscape of deep linear neural networks: the regularized deep learning can be converted into an equivalent convex problem by dual. To the best of our knowledge, this is the first result on the strong duality of linear neural networks. We note that the strong duality rarely happens in the non-convex optimization: matrix completion [6], Fantope [47], and quadratic optimization with two quadratic constraints [11] are among the few paradigms that enjoy the strong duality. For deep networks, the effectiveness of convex relaxation has been observed empirically in [3, 67], but much remains unknown for the theoretical guarantees of the relaxation. Our work shows strong duality of regularized deep linear neural networks and provides an alternative approach to overcome the computational obstacles due

[4] Although the expressive power of deep linear neural networks and three-layer linear neural networks are the same, the analysis of landscapes of two models are significantly different, as pointed out by [28, 35, 44].

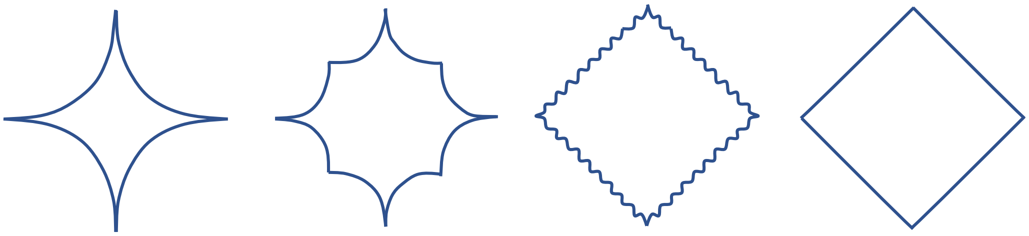


Figure 3: Visualization of Shapley-Folkman lemma. **The first figure:** an $\ell_{1/2}$ ball. **The second and third figures:** the averaged Minkowski sum of two and ten $\ell_{1/2}$ balls. **The fourth figure:** the convex hull of $\ell_{1/2}$ ball (the Minkowski average of infinitely many $\ell_{1/2}$ balls). It show that with the number of $\ell_{1/2}$ balls to be averaged increasing, the Minkowski average tends to be more convex.

to the non-convexity: one can apply convex solvers, e.g., the Douglas–Rachford algorithm,[5] for problem (5) and then conduct singular value decomposition to compute the weights $\{\mathbf{W}_i^*\}_{i=1}^H$ from $\mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$. In addition, our result inherits the benefits of convex analysis. The vast majority results on deep learning study the generalization error or expressive power by analyzing its complicated non-convex form [46, 66, 63]. In contrast, with strong duality one can investigate various properties of deep linear networks with much simpler convex form.

Related Works. The goal of convexified linear neural networks is to relax the non-convex form of deep learning to the computable convex formulations [67, 3]. While several efforts have been devoted to investigating the effectiveness of such convex surrogates, e.g., by analyzing the generalization error after the relaxation [67], little is known whether the relaxation is tight to its original problem. Our result, on the other hand, provides theoretical guarantees for the tightness of convex relaxation of deep linear networks, a phenomenon observed empirically in [3, 67].

We mention another related line of research — no bad local minima. On one hand, although recent works have shown the absence of spurious local minimum for deep linear neural networks [50, 35, 44], many of them typically lack theoretical analysis of regularization term. Specifically, Kawaguchi [35] showed that *un-regularized* deep linear neural networks have no spurious local minimum. Lu and Kawaguchi [44] proved that depth creates no bad local minimum for *un-regularized* deep linear neural networks. In contrast, our optimization problem is more general by taking the regularization term into account. On the other hand, even the "local=global" argument holds for the deep linear neural networks, it is still hard to escape bad saddle points [1]. In particular, Kawaguchi [35] proved that for linear networks deeper than three layers, there exist bad saddle points at which the Hessian does not have any negative eigenvalue. Therefore, the state-of-the-art algorithms designed to escape the saddle points might not be applicable [34, 26]. Our result provides an alternative approach to solve deep linear network by convex programming.

4 Our Techniques and Proof Sketches

In this section, we present our techniques and proof sketches of Theorems 1 and 2.

(a) Shapley-Folkman Lemma. The proof of Theorem 1 is built upon the Shapley-Folkman lemma [22, 55, 24, 13], which characterizes a convexification phenomenon concerning the average of multiple sets and is analogous to the central limit theorem in the probability theory. Consider the averaged Minkowski sum of I sets $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_I$ given by $\{I^{-1}\sum_{j\in[I]} a_j : a_j \in \mathcal{A}_j\}$. Intuitively, the lemma states that $\rho(I^{-1}\sum_{j\in[I]} \mathcal{A}_j) \to 0$ as $I \to +\infty$, where $\rho(\cdot)$ is a metric of the non-convexity of a set (see Figure 3 for visualization). We apply this lemma to the optimization formulation of deep neural networks. Denote by *augmented epigraph* the set $\{(h(\mathbf{w}), \ell(\mathbf{w})) : \text{all possible choices of } \mathbf{w}\}$, where h is the constraint and ℓ is the objective function in the optimization problem. The key observation is that the augmented epigraph of neural network loss with multi-branch architecture can be expressed as the Minkowski average of augmented epigraphs of all branches. Thus we obtain a natural connection between an optimization problem and its corresponding augmented epigraph. Applying Shapley-Folkman lemma to the augmented epigraph leads to a characteristic of non-convexity of the deep neural network.

(b) Variational Form. The proof of Theorem 2 is built upon techniques (b), (c), and (d). In particular, problem (4) is highly non-convex due to its multi-linear form over the optimized variables $\{\mathbf{W}_i\}_{i=1}^H$. Fortu-

[5] Grussler et al. [29] provided a fast algorithm to compute the proximal operators of $\frac{1}{2}\|\cdot\|_{d_{\min}}^2$. Hence, the Douglas–Rachford algorithm can find the global solution up to an ϵ error in function value in time $\mathsf{poly}(1/\epsilon)$ [32].

nately, we are able to analyze the problem by grouping $\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1\mathbf{X}$ together and converting the original non-convex problem in terms of the separate variables $\{\mathbf{W}_i\}_{i=1}^H$ to a convex optimization with respect to the new grouping variable $\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1\mathbf{X}$. This typically requires us to represent the objective function of (4) as a convex function of $\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1$. To this end, we prove that $\|\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1\mathbf{X}\|_* = \min_{\mathbf{W}_1,...,\mathbf{W}_H} \frac{1}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^H \|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right]$. So the objective function in problem (4) has an equivalent form

$$\begin{aligned}\min_{\mathbf{W}_1,...,\mathbf{W}_H} \frac{1}{2}\|\mathbf{Y} - \mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2 \\ + \gamma\|\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_*.\end{aligned} \tag{6}$$

This observation enables us to represent the optimization problem as a convex function of the output of a neural network. Therefore, we can analyze the non-convex problem by applying powerful tools from convex analysis.

(c) Reduction to Low-Rank Approximation. Our results of strong duality concerning problem (6) are inspired by the problem of low-rank matrix approximation:

$$\min_{\mathbf{W}_1,...,\mathbf{W}_H} \frac{1}{2}\|\mathbf{Y} - \mathbf{\Lambda}^* - \mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2. \tag{7}$$

We know that all local solutions of (7) are globally optimal [35, 44, 6]. To analyze the more general regularized problem (4), our main idea is to reduce problem (6) to the form of (7) by Lagrangian function. In other words, the Lagrangian function of problem (6) should be of the form (7) for a fixed Lagrangian variable $\mathbf{\Lambda}^*$, which we will construct later in subsection (d). While some prior works attempted to apply a similar reduction, their conclusions either depended on unrealistic conditions on local solutions, e.g., all local solutions are rank-deficient [30, 29], or their conclusions relied on strong assumptions on the objective functions, e.g., that the objective functions are twice-differentiable [30], which do not apply to the non-smooth problem (6). Instead, our results bypass these obstacles by formulating the strong duality of problem (6) as the existence of a dual certificate $\mathbf{\Lambda}^*$ satisfying certain dual conditions. Roughly, the dual conditions state that the optimal solution $(\mathbf{W}_1^*, \mathbf{W}_2^*, ..., \mathbf{W}_H^*)$ of problem (6) is locally optimal to problem (7). On one hand, by the above-mentioned properties of problem (7), $(\mathbf{W}_1^*, ..., \mathbf{W}_H^*)$ globally minimizes the Lagrangian function when $\mathbf{\Lambda}$ is fixed to $\mathbf{\Lambda}^*$. On the other hand, by the convexity of nuclear norm, for the fixed $(\mathbf{W}_1^*, ..., \mathbf{W}_H^*)$ the Lagrangian variable $\mathbf{\Lambda}^*$ globally optimize the Lagrangian function. Thus $(\mathbf{W}_1^*, ..., \mathbf{W}_H^*, \mathbf{\Lambda}^*)$ is a primal-dual saddle point of the Lagrangian function of problem (6). The desired strong duality is a straightforward result from this argument.

(d) Dual Certificate. The remaining proof is to construct a dual certificate $\mathbf{\Lambda}^*$ such that the dual conditions hold true. The challenge is that the dual conditions impose several constraints simultaneously on the dual certificate, making it hard to find a desired certificate. This is why progress on the dual certificate has focused on convex programming. To resolve the issue, we carefully choose the certificate as an appropriate scaling of subgradient of nuclear norm around a low-rank solution, where the nuclear norm follows from our regularization term in technique (b). Although the nuclear norm has infinitely many subgradients, we prove that our construction of dual certificate obeys all desired dual conditions. Putting techniques (b), (c), and (d) together, our proof of strong duality is completed.

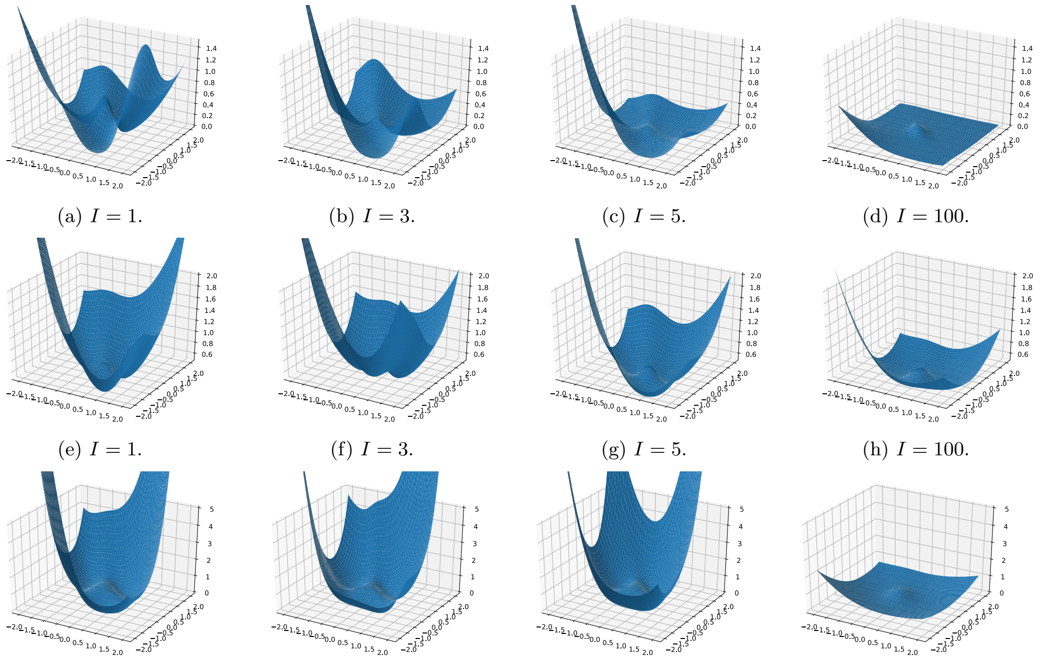
5 Experiments

In this section, we verify our theoretical contributions by the experimental validation.

5.1 Visualization of Loss Landscape

Experiments on Synthetic Datasets. We first show that over-parametrization results in a less non-convex loss surface for a synthetic dataset. The dataset consists of $1,000$ examples in $\mathbb{R}^{10}$ whose labels are generated by an underlying one-hidden-layer ReLU network $f(\mathbf{x}) = \sum_{i=1}^I \mathbf{w}_{i,2}^*[\mathbf{W}_{i,1}^*\mathbf{x}]_+$ with 11 hidden neurons [49]. We make use of the visualization technique employed by [40] to plot the landscape, where we project the high-dimensional hinge loss ($\tau = 1$) landscape onto a 2-d plane spanned by three points. These points are found by running the SGD algorithm with three different initializations until the algorithm converges. As shown in Figure 1, the landscape exhibits strong non-convexity with lots of local minima in the under-parameterized case $I = 10$. But as I increases, the landscape becomes more convex. In the extreme case, when there are $1,000$ hidden neurons in the network, no non-convexity can be observed on the landscape.

Experiments on MNIST and CIFAR-10. We next verify the phenomenon of over-parametrization on MNIST [39] and CIFAR-10 [37] datasets. For both datasets, we follow the standard preprocessing step that each pixel is normalized by subtracting its mean and dividing by its standard deviation. We do not apply data augmentation. For MNIST, we consider a single-hidden-layer network defined as: $f(\mathbf{x}) = \sum_{i=1}^I \mathbf{W}_{i,2}[\mathbf{W}_{i,1}\mathbf{x}]_+$, where $\mathbf{W}_{i,1} \in \mathbb{R}^{h\times d}$, $\mathbf{W}_{i,2} \in \mathbb{R}^{10\times h}$, d is the input dimension, h is the number of hidden neurons, and I is



(a) $I = 1$. (b) $I = 3$. (c) $I = 5$. (d) $I = 100$.

(e) $I = 1$. (f) $I = 3$. (g) $I = 5$. (h) $I = 100$.

(i) $I = 1$. (j) $I = 3$. (k) $I = 5$. (l) $I = 50,000$.

Figure 4: **Top Row:** Landscape of one-hidden-layer network on MNIST. **Middle Row:** Landscape of one-hidden-layer network on CIFAR-10. **Bottom Row:** Landscape of three-hidden-layer, multi-branch network on CIFAR-10 dataset. From left to right, the landscape looks less non-convex.

the number of branches, with $d = 784$ and $h = 8$. For CIFAR-10, in addition to considering the exact same one-hidden-layer architecture, we also test a deeper network containing 3 hidden layers of size 8-8-8, with ReLU activations and $d = 3,072$. We apply 10-class hinge loss on the top of the output of considered networks.

Figure 4 shows the changes of landscapes when I increases from 1 to 100 for MNIST, and from 1 to 50,000 for CIFAR-10, respectively. When there is only one branch, the landscapes have strong non-convexity with many local minima. As the number of branches I increases, the landscape becomes more convex. When $I = 100$ for 1-hidden-layer networks on MNIST and CIFAR-10, and $I = 50,000$ for 3-hidden-layer network on CIFAR-10, the landscape is almost convex.

5.2 Frequency of Hitting Global Minimum

To further analyze the non-convexity of loss surfaces, we consider various one-hidden-layer networks, where each network was trained 100 times using different initialization seeds under the setting discussed in our synthetic experiments of Section 5.1. Since we have the ground-truth global minimum, we record the frequency that SGD hits the global minimum up to a small error 1×10^{-5} after 100,000 iterations. Table 1 shows that increasing the number of hidden neurons results in higher hitting rate of global optimality. This further verifies that the loss surface of one-hidden-layer neural network becomes less non-convex as the width increases.

Table 1: Frequency of hitting global minimum by SGD with 100 different initialization seeds.

# Hidden Neurons	Hitting Rate
12	21 / 100
13	24 / 100
14	24 / 100
15	29 / 100
16	30 / 100
17	32 / 100
18	35 / 100
19	52 / 100
20	64 / 100
21	75 / 100

6 Conclusions

In this work, we study the duality gap for two classes of network architectures. For the neural network with *arbitrary activation functions*, multi-branch architecture and τ-hinge loss, we show that the duality gap of both population and empirical risks shrinks to zero as the number of branches increases. For the neural network with linear activation function and ℓ_2 loss, we show that the duality gap is zero. Our two results work for *arbitrary depths*, while the analytical techniques might be of independent interest to non-convex optimization more broadly, e.g., the optimization problems in adversarial defense and attack [65].

References

[1] A. Anandkumar and R. Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Annual Conference on Learning Theory*, pages 81–102, 2016.

[2] S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, 2018.

[3] Ö. Aslan, X. Zhang, and D. Schuurmans. Convex deep learning via normalized kernels. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2014.

[4] P. Awasthi, M.-F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Annual Conference on Learning Theory*, pages 152–192, 2016.

[5] F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[6] M.-F. Balcan, Y. Liang, D. P. Woodruff, and H. Zhang. Matrix completion and related problems via strong duality. In *Innovations in Theoretical Computer Science*, 2018.

[7] M.-F. F. Balcan and H. Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. In *Advances in Neural Information Processing Systems*, pages 4799–4808, 2017.

[8] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

[9] P. Baldi and Z. Lu. Complex-valued autoencoders. *Neural Networks*, 33:136–147, 2012.

[10] P. Bartlett and S. Ben-David. Hardness results for neural network approximation problems. In *European Conference on Computational Learning Theory*, pages 50–62, 1999.

[11] A. Beck and Y. C. Eldar. Strong duality in nonconvex quadratic optimization with two quadratic constraints. *SIAM Journal on Optimization*, 17(3):844–860, 2006.

[12] Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. In *Advances in neural information processing systems*, pages 123–130, 2006.

[13] D. P. Bertsekas and N. R. Sandell. Estimates of the duality gap for large-scale separable nonconvex optimization problems. In *IEEE Conference on Decision and Control*, volume 21, pages 782–785, 1982.

[14] Y. Bi and A. Tang. Refined Shapely-Folkman lemma and its application in duality gap estimation. *arXiv preprint arXiv:1610.05416*, 2016.

[15] A. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[16] A. Brutzkus and A. Globerson. Globally optimal gradient descent for a ConvNet with Gaussian inputs. In *International Conference on Machine Learning*, 2017.

[17] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018.

[18] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[19] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–200, 2015.

[20] A. Choromanska, Y. LeCun, and G. B. Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Annual Conference on Learning Theory*, pages 1756–1760, 2015.

[21] B. DasGupta, H. T. Siegelmann, and E. Sontag. On the complexity of training neural networks with continuous activation functions. *IEEE Transactions on Neural Networks*, 6(6):1490–1504, 1995.

[22] A. d'Aspremont and I. Colin. An approximate Shapley-Folkman theorem. *arXiv preprint arXiv:1712.08559*, 2017.

[23] S. S. Du and J. D. Lee. On the power of overparametrization in neural networks with quadratic activation. In *International Conference on Machine Learning*, 2018.

[24] E. X. Fang, H. Liu, and M. Wang. Blessing of massive scale: Spatial graphical model estimation with a total cardinality constraint. 2015.

[25] X. Gastaldi. Shake-Shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

[26] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Annual Conference on Learning Theory*, pages 797–842, 2015.

[27] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2017.

[28] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[29] C. Grussler, A. Rantzer, and P. Giselsson. Low-rank optimization with convex constraints. *arXiv:1606.01793*, 2016.

[30] B. D. Haeffele and R. Vidal. Global optimality in neural network training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339, 2017.

[31] M. Hardt and T. Ma. Identity matters in deep learning. In *International Conference on Learning Representations*, 2017.

[32] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[33] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and$<$ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 2016.

[34] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. *arXiv:1703.00887*, 2017.

[35] K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[36] K. Kawaguchi, B. Xie, and L. Song. Deep semi-random features for nonlinear function approximation. In *AAAI Conference on Artificial Intelligence*, 2018.

[37] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[38] T. Laurent and J. von Brecht. The multilinear structure of ReLU networks. *arXiv preprint arXiv:1712.10132*, 2017.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] H. Li, Z. Xu, G. Taylor, and T. Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

[41] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Annual Conference on Learning Theory*, 2017.

[42] S. Liang, R. Sun, J. D. Lee, and R. Srikant. Adding one neuron can eliminate all bad local minima. *arXiv preprint arXiv:1805.08671*, 2018.

[43] S. Liang, R. Sun, Y. Li, and R. Srikant. Understanding the loss surface of neural networks for binary classification. In *International Conference on Machine Learning*, 2018.

[44] H. Lu and K. Kawaguchi. Depth creates no bad local minima. *arXiv:1702.08580*, 2017.

[45] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

[46] B. Neyshabur, R. Salakhutdinov, and N. Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

[47] M. L. Overton and R. S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.

[48] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[49] I. Safran and O. Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, 2017.

[50] A. M. Saxe. *Deep linear neural networks: A theory of learning in the brain and mind*. PhD thesis, Stanford University, 2015.

[51] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120*, 2013.

[52] S. Shalev-Shwartz, O. Shamir, and S. Shammah. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, 2017.

[53] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

[54] D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv:1605.08361*, 2016.

[55] R. M. Starr. Quasi-equilibria in markets with nonconvex preferences. *Econometrica: Journal of the Econometric Society*, pages 25–38, 1969.

[56] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[57] M. Udell and S. Boyd. Bounding duality gap for separable problems with linear constraints. *Computational Optimization and Applications*, 64(2):355–378, 2016.

[58] A. Veit, M. J. Wilber, and S. Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, pages 550–558, 2016.

[59] R. Vidal, J. Bruna, R. Giryes, and S. Soatto. Mathematics of deep learning. *arXiv preprint arXiv:1712.04741*, 2017.

[60] B. Xie, Y. Liang, and L. Song. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pages 1216–1224, 2017.

[61] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5987–5995, 2017.

[62] S. Zagoruyko and N. Komodakis. Wide residual networks. In *British Machine Vision Conference*, pages 87.1–87.12, 2016.

[63] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2016.

[64] H. Zhang, S. Xu, J. Jiao, P. Xie, R. Salakhutdinov, and E. P. Xing. Stackelberg gan: Towards provable minimax equilibrium via multi-generator architectures. *arXiv preprint arXiv:1811.08010*, 2018.

[65] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

[66] Y. Zhang, J. Lee, M. Wainwright, and M. Jordan. On the learnability of fully-connected neural networks. In *Artificial Intelligence and Statistics*, pages 83–91, 2017.

[67] Y. Zhang, P. Liang, and M. J. Wainwright. Convexified convolutional neural networks. In *International Conference on Machine Learning*, 2017.

[68] P. Zhou and J. Feng. Empirical risk landscape analysis for understanding deep neural networks. In *International Conference on Learning Representations*, 2018.