ON THE LANDSCAPE OF SPARSE LINEAR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Network pruning, or sparse network has a long history and practical significance in modern applications. Although the loss functions of neural networks may yield bad landscape due to non-convexity, we focus on linear activation which already owes benign landscape. With no unrealistic assumption, we conclude the following statements for the squared loss objective of general sparse linear neural networks: 1) every local minimum is a global minimum for scalar output with any sparse structure, or non-intersected sparse first layer and dense other layers with orthogonal training data; 2) sparse linear networks have sub-optimal local-min for only sparse first layer due to low rank constraint, or output larger than three dimensions due to the global minimum of a sub-network. Overall, sparsity breaks the normal structure, cutting out the decreasing path in original fully-connected networks.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable empirical successes in the domains of computer vision, speech recognition, and natural language processing, sparking great interests in the theory behind their architectures and training. However, DNNs are often found to be highly overparameterized, making them computationally expensive with large amounts of memory and computational power. For example, it may take up to weeks on a modern multi-GPU server for large datasets such as ImageNet (Deng et al., 2009). Hence, DNNs are often unsuitable for smaller devices like embedded electronics, and there is a pressing demand for techniques to optimize models with reduced model size, faster inference and lower power consumption.

Sparse networks, that is, neural networks in which a large subset of the model parameters are zero, have emerged as one of the leading approaches for reducing model parameter count. It has been shown empirically that deep neural networks can achieve state-of-the-art results under high levels of sparsity (Han et al., 2015b; Gale et al., 2019; Louizos et al., 2017a). Modern sparse networks are mainly obtained from network pruning (Zhu & Gupta, 2017; Lee et al., 2018; Liu et al., 2018; Frankle & Carbin, 2018), which has been the subject of a great deal of work in recent years. However, training a sparse network with fixed sparsity patterns is difficult (Evci et al., 2019) and few theoretical understanding of general sparse networks has been provided.

Previous work has already analyze deep neural networks, showing that the non-convexity of the associated loss functions may cause complicated and strange optimization landscapes. However, the property of general sparse networks is poorly understood. Saxe et al. (2013) empirically showed that the optimization of deep linear models exhibits similar properties as deep nonlinear models, and for theoretical development, it is natural to begin with linear models before studying nonlinear models (Baldi & Lu, 2012). In addition, several works (Sun et al., 2020) have show bad minimum exists with nonlinear activation. Hence, it is natural to begin with linear activation to understand the impact of sparsity.

In this article, we go further to consider the global landscape of general sparse linear neural networks. We need to emphasize that dense deep linear networks already satisfy that every local minimum is a global minimum under mild conditions (Kawaguchi, 2016; Lu & Kawaguchi, 2017), but findings are different and complicated for sparse linear network. The goal of this paper is to study the relation between sparsity and local minima with the following contributions:

- First, we point out that every local minimum is a global minimum in scalar target case with any depths, any widths and any sparse structure. Besides, we also briefly show that

similar results hold for non-overlapping filters and orthogonal data feature when sparsity only occurs in the first layer.

- Second, we find out that sparse connections would already give sub-optimal local minima in general non-scalar case through analytic and numerical examples built on convergence analyze. The local-min may be produced from two situations: a sub-sparse linear network which owes its minimum as a local-min of the original sparse network; a rank-deficient solution between different data features due to sparse connections, while both cases verify the fact that sparsity cuts out the decreasing path in original fully-connected networks.

Overall, we hope our work contributes to a better understanding of the landscape of sparsity network on simple neural networks, and provide insights for future research.

The remainder of our paper is organized as follows. In Section 2, we derive the positive findings of shallow sparse linear network, providing similar landscape as dense linear networks. In Section 3, we give negative results based on several examples showing the existence of bad local-min for non-scalar case. In section 4, we briefly generalize the results from shallow to deep sparse linear networks. Missing proofs are in Appendix.

1.1 Related Work

There is a rapidly increasing literature on analyzing the loss surface of neural network objectives, surveying all of which is well outside our scope. Thus, we only briefly survey the works most relevant to ours.

Local minima is Global. Landscape of linear network date back to Baldi & Hornik (1989), proving shallow linear neural networks do not suffer from bad local minima. Kawaguchi (2016) generalize same results to deep linear neural networks, and subsequent several works (Arora et al., 2018; Du & Hu, 2019; Eftekhari, 2020) give direct algorithm-type convergence based on this benign property, though algorithm analysis is beyond the scope of this paper. However, situations are quite complicated with nonlinear activations. Multiple works (Ge et al., 2017; Safran & Shamir, 2018; Yun et al., 2018) show that spurious local minima can happen even in two-layer network with population or empirical loss, some are specific to two-layer and difficult to generalize to general multilayer case. Another line of works (Arora et al., 2018; Allen-Zhu et al., 2018; Du & Hu, 2019; Du et al., 2018; Li et al., 2018; Mei et al., 2018) understanding the landscape of neural network in overparameterized setting, discovering benign landscape w/o gradient method. Since modern sparse networks reserve few parameters compared to overparameterization, we still seek a fundamental view of sparsity in contrast. Our standpoint is that spurious local minima can happen when applied with specific sparsity even in linear networks.

Sparse networks. Sparse networks (Han et al., 2015b;a; Zhu & Gupta, 2017; Frankle & Carbin, 2018; Liu et al., 2018) have a long history, but appears heavily on the experiments, and mainly related to network pruning, which has practical importance for reducing model parameter count and deploying diverse devices. However, training sparse network (from scratch) suffers great difficulty. Frankle & Carbin (2018) recommend reusing the sparsity pattern found through pruning and train a sparse network from the same initialization as the original training ('lottery') to obtain comparable performance and avoid bad solution. Besides, for fixed sparsity patterns, Evci et al. (2019) attempt to find a decreasing objective path from 'bad' solutions to the 'good' ones in the sparse subspace but fail, showing bad local minima can be produced by pruning or sparsity, while we give more direct view of simple examples to verify this. Moreover, several recent works also give abundant methods (Molchanov et al., 2017; Louizos et al., 2017b; Lee et al., 2018; Carreira-Perpinán & Idelbayev, 2018) for choosing weights or sparse network structure while achieving similar performance. In theoretical view, Malach et al. (2020) prove that a sufficiently over-parameterized neural network with random weights contains a subnetwork with roughly the same accuracy as the target network, providing guarantee for 'good' sparse networks. Some works analyze convolutional network (Shalev-Shwartz et al., 2017; Du et al., 2018) as a specific sparse structure. Brutzkus & Globerson (2017) analyze non-overlapping and overlapping structure as we do, but with weight sharing to simulate CNN-type structure, and under teacher-student setting with population risk. We do not follow CNN-type network but in general sparse networks, though still linear, to conclude straightforward results.

2 LANDSCAPE OF SHALLOW SPARSE LINEAR NETWORKS

2.1 PRELIMINARIES AND NOTATION

We use bold-faced letters (e.g., $\mathbf{w}, \mathbf{a}$) to denote vectors, capital letters (e.g., W, A) for matrices. $W_{i,\cdot}$ as the i-th row, $W_{\cdot,j}$ as the j-th column of matrix W and $w_{ij} := W_{ij}$. Let $\mathcal{P}_X$ be the orthogonal projection to the column space of matrix X.

We consider the training samples and their labels as $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, respectively. By concatenating $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$, we form the data matrices $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{n \times d_y}$ with $\mathbf{x}_i^T, \mathbf{y}_i^T$ in each row, which may comes from unknown distribution $\mathcal{D}$. In our analysis in Section 2 and Section 3, we consider two-layer (sparse) linear neural network with squared loss:

$$\min_{W,A} L(W, A) = \frac{1}{2}\|Y - XWA\|_F^2, \tag{1}$$

where the first layer weight matrix $W \in \mathbb{R}^{d_x \times d}$ with $\mathbf{w}_j$ as j-th column of W, and the second layer weight matrix $A \in \mathbb{R}^{d \times d_y}$ with $\mathbf{a}_i^T$ as its i-th row. After weights pruning or sparsity constraint, several weights parameters become zero and would not be updated during retraining. We adopt $\mathcal{S}_j := \{k : W_{kj} = 0\}$ as pruned dimensions in j-th column of W, and $-\mathcal{S}_j := \mathcal{S}_j^c = [d_x] \backslash \mathcal{S}_j$, $[d] := \{1, \ldots, d\}$. Besides, $\mathbf{w}_{j,\mathcal{S}}$ as the sub-vector of $\mathbf{w}_j$ choosing positions in $\mathcal{S}$, $X_{\mathcal{S}}$ as the sub-matrix of X choosing columns indices in $\mathcal{S}$.

We denote $p_j := d_x - |\mathcal{S}_j|$, where $|\mathcal{S}|$ is the number of $\mathcal{S}$'s elements. Then $\mathbf{w}_{j,-\mathcal{S}_j} \in \mathbb{R}^{p_j}$ as the remaining j-th column in first layer weight which leaves out pruned dimension set $\mathcal{S}_j$. Similarly, $X_{-\mathcal{S}_j} \in \mathbb{R}^{n \times p_j}$ means the remaining data matrix connected to j-th node in the first layer.

Finally, for simplification, we denote $X_{-j} = X_{-\mathcal{S}_j}$, $\mathbf{w}_{-j} = \mathbf{w}_{j,-\mathcal{S}_j}$, and $\widetilde{(\cdot)}$ as the pruned layer weight with several zero elements which **not** updated all along, if no ambiguity.

Before we begin, a small note on the sparse structure we concern: there may have unnecessary connections and nodes, such as a node with zero out-degree which can be retrieved and excluded from the final layer to the first layer, and other cases are showing in Appendix C. Thus we don't consider them in the subsequent proof and assume each data dimension has valid output connection, i.e. $\cap_{j=1}^d \mathcal{S}_j = \emptyset$.

2.2 SCALAR CASE

Suppose $d_y = 1$, then we simplify $A = (a_1, \ldots, a_d)^T$. When pruning any weight a_i in the second layer, the output of i-th node in the first layer contribute zero to final output. Hence $\mathbf{w}_i$ can also be pruned. Without loss of generality, we assume second layer parameters do not pruned. After pruning several parameters, the original problem becomes

$$\min_{\mathbf{w}_{-i}, a_i} L(\widetilde{W}, A) = \frac{1}{2}\left\| Y - \begin{pmatrix} X_{-1} & \cdots & X_{-d} \end{pmatrix} \begin{pmatrix} a_1 \mathbf{w}_{-1} \\ \vdots \\ a_d \mathbf{w}_{-d} \end{pmatrix} \right\|_F^2. \tag{2}$$

Theorem 1 *For a two-layer linear neural network with scalar output and any sparse structure, every local minimum is a global minimum.*

Proof: From Eq. 2, if a local minimum satisfies $a_i = 0$ for some $1 \leq i \leq d$, then based on the second order condition for a local minima, we have

$$\begin{pmatrix} \dfrac{\partial^2 L}{\partial a_i^2} & \dfrac{\partial^2 L}{\partial a_i \partial \mathbf{w}_{-i}^T} \\ \dfrac{\partial L}{\partial \mathbf{w}_{-i} \partial a_i} & \dfrac{\partial L}{\partial \mathbf{w}_{-i} \partial \mathbf{w}_{-i}^T} \end{pmatrix} \succeq \mathbf{0}. \tag{3}$$

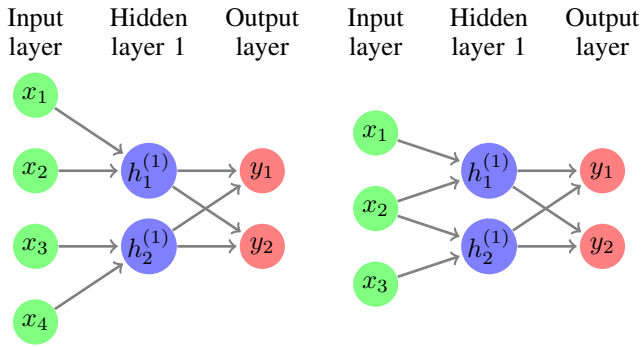

Figure 1: Sparse network without (left) / with (right) overlapping filters in the first layer.

showing that

$$\begin{pmatrix} \mathbf{w}_{-i}^T X_{-i} X_{-i}^T \mathbf{w}_{-i} & -\left(Y - \sum_{i=1}^d X_{-i}\mathbf{w}_{-i}a_i\right)^T X_{-i} \\ -X_{-i}^T\left(Y - \sum_{i=1}^d X_{-i}\mathbf{w}_{-i}a_i\right) & \mathbf{0} \end{pmatrix} \succeq \mathbf{0}. \tag{4}$$

Then $X_{-i}^T\left(Y - \sum_{i=1}^d X_{-i}\mathbf{w}_{-i}a_i\right) = 0$, which meaning the global minimum condition of $\mathbf{w}_{-i}a_i$.

Otherwise, $a_i \neq 0$, then from the first-order condition for a local minima,

$$\frac{\partial L}{\partial \mathbf{w}_{-i}} = a_i X_{-i}^T\left(Y - \sum_{i=1}^d X_{-i}\mathbf{w}_{-i}a_i\right) = 0,$$

showing that $X_{-i}^T\left(Y - \sum_{i=1}^d X_{-i}\mathbf{w}_{-i}a_i\right) = 0$, which also gives the global minimum condition of $\mathbf{w}_{-i}a_i$. Hence every local minimum is a global minimum. □

2.3 Non-scalar Case with Dense Second Layer

Now we discuss non-scalar case. By the intractable and various sparse structure, we first consider pruning only the first layer while retaining the dense second layer. Then the remaining problem is formulated as below:

$$\min_{\mathbf{w}_{-i},\mathbf{a}_i} L(\widetilde{W}, A) = \frac{1}{2}\left\|Y - \sum_{i=1}^d X_{-i}\mathbf{w}_{-i}\mathbf{a}_i^T\right\|_F^2, \tag{5}$$

Intuitively, if we can separate the weight parameters into d parts, the based on linear network results, we can still guarantee no bad local-min. We show that non-overlapping first layer weight or disjoint feature extractor, as the left graph of Figure 1 depicts, and orthogonal data feature meet requirements.

Theorem 2 *For a two-layer sparse linear neural network with dense second layer, assume that X has full column rank, and $\forall\, i \neq j$, $X_{-i}^T X_{-j} = \mathbf{0}$, then every local minimum is a global minimum.*

Proof: Since $\forall\, i \neq j$, $X_{-i}^T X_{-j} = \mathbf{0}$ and X has full column rank, then X_{-i}, X_{-j} share no same columns. Besides, from our assumption $\cap_{j=1}^d \mathcal{S}_j = \emptyset$, then $\cap_{j=1}^d \mathcal{S}_j = [d_x]$, meaning that $(X_{-1}, \ldots, X_{-d})$ is X with different arrangement of columns. Hence

$$Y = \mathcal{P}_X Y + (I - \mathcal{P}_X)Y = X(X^TX)^{-1}X^TY + (I - \mathcal{P}_X)Y \triangleq \sum_{i=1}^d X_{-i}Z_i + (I - \mathcal{P}_X)Y, \tag{6}$$

where $Z_i = \left((X^TX)^{-1}X^TY\right)_{-\mathcal{S}_i,\cdot}$, then the original objective becomes

$$\min_{\mathbf{w}_{-i},\mathbf{a}_i} L(\widetilde{W}, A) = \frac{1}{2}\left\|\sum_{i=1}^d X_{-i}\left(Z_i - \mathbf{w}_{-i}\mathbf{a}_i^T\right)\right\|_F^2 = \frac{1}{2}\sum_{i=1}^d \left\|X_{-i}\left(Z_i - \mathbf{w}_{-i}\mathbf{a}_i^T\right)\right\|_F^2. \tag{7}$$

We will see that the objective has already been separated into d parts, while each part is a two-layer dense linear network with full column rank data matrix. Based on Theorem 2.2 in Lu & Kawaguchi (2017) or Eckart-Young-Mirsky theorem (Eckart & Young, 1936; Mirsky, 1960). Hence, every local minimum is a global minimum. □

Additionally, we need to explain the assumption that non-overlapping filters in the first layer involves convolution networks (Brutzkus & Globerson, 2017) if weight sharing used. Otherwise, we show a bad local minima exists when the first layer is overlapped or not orthogonal provided in Section 3.

2.4 General Case with $d_y = 2$

Previous findings imply positive results with single target dimension, or specific sparse structure and data assumption. In this subsection, we discuss arbitrarily sparse structure with target dimension $d_y = 2$.

We first prove that some connections still owe common benign landscape which can be removed.

Theorem 3 *For a sparse two-layer network, a node with full out-degree and one in-degree can be removed if we consider the remaining structure with objective under some projected data, having no influence for spurious local minima.*

Previous result simplifies the sparse structure including such hidden node with one connection to input and full connection to output. Next, we will provide another type reduction with only one connection to output when sparsity applied to both layers with $d_y = 2$.

We mention the final layer output as node 1 and node 2, and the hidden nodes which have only one connection to the output layer as $\mathcal{R}_1$ and $\mathcal{R}_2$ while the remaining full connection set as $\mathcal{R}$. Set $\mathcal{T}_1 = \cap_{j\in\mathcal{R}_1}\mathcal{S}_j, \mathcal{T}_2 = \cap_{j\in\mathcal{R}_2}\mathcal{S}_j$. We define $\mathcal{U}(\mathcal{T}) = \{j : w_{ij} \neq 0, i \in \mathcal{T}\}$ as the hidden node set connected with data feature in $\mathcal{T}$. When the hidden node set connected to $\mathcal{T}_1$ and $\mathcal{T}_2$ satisfy the condition below, we are able to simply the sparse structure into the dense layer case.

Theorem 4 *For a sparse two-layer linear network with $d_y = 2$, if $\mathcal{U}(\mathcal{T}_1 \cap \mathcal{T}_2) \cap \mathcal{U}(\mathcal{T}_1 \setminus \mathcal{T}_2) = \emptyset$ and $\mathcal{U}(\mathcal{T}_1 \cap \mathcal{T}_2) \cap \mathcal{U}(\mathcal{T}_2 \setminus \mathcal{T}_1) = \emptyset$, then there is a sub-network with dense second layer optimized with some projected training data, sharing same local minimum for the remaining parameters.*

The formal proof of the theorem can be found in Appendix E. Besides, from the proof of Thm 4, the objective is converted into two objectives with weight sharing in the first layer even the assumption doesn't meet. Weight sharing structure has shown in some related works (Shalev-Shwartz et al., 2017; Brutzkus & Globerson, 2017), so we do not give detailed description and leave it as future work.

Now for a sparse two-layer linear network with $d_y = 2$, we focus on the case which has dense second layer. If one hidden node only has one in-degree, then based on Thm 3, we can remove this node and optimized with some projected data. Therefore, each hidden node should have at least two in-degree, since one hidden node is obvious no bad local minima, then the least sparse structure has two hidden node with totally eight connections (e.g., two constructions in Figure 1). We will show the existence of bad local-minima in this case in Section 3.

2.5 General Case with $d_y \geq 3$

We close this section by discovering that a sub-network with its global minima may introduce a spurious local minima of the original sparse network when $d_y \geq 3$, .

Theorem 5 *There exists a spurious local minima that is a global minimum of sub-network in two-layer sparse network when output dimension $d_y \geq 3$.*

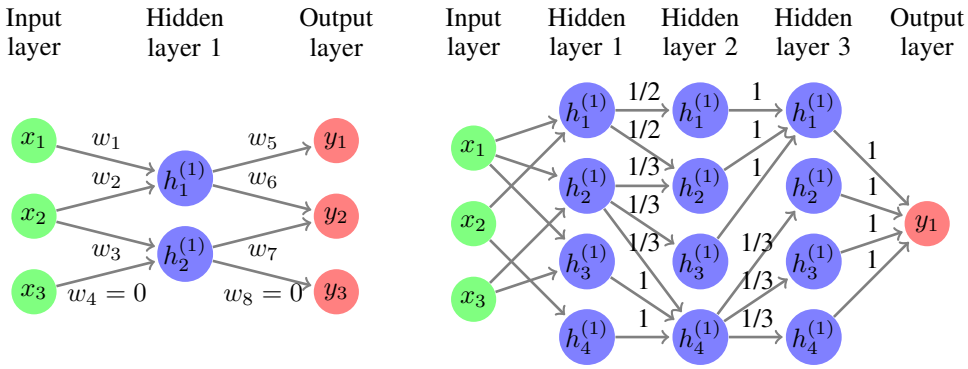

Figure 2: Left: spurious local minimum exists in sparse linear network with $d_y = 3$ shown in above, that is the global minimum of a sub-network. Right: one simple weight assignment for obtaining the global minimum in deep sparse linear network with scalar output.

Proof: We consider the sparse structure in Figure 2 with only eight connections. The objective is

$$\begin{aligned}\min_{w_i} L(w_1,\dots,w_8) &= \frac{1}{2}\left\| Y - X\begin{pmatrix} w_1 & 0 \\ w_2 & w_3 \\ 0 & w_4\end{pmatrix}\begin{pmatrix} w_5 & w_6 & 0 \\ 0 & w_7 & w_8\end{pmatrix}\right\|_F^2 \\ &= \frac{1}{2}\left\| Y - X\begin{pmatrix} w_1w_5 & w_1w_6 & 0 \\ w_2w_5 & w_2w_6 + w_3w_7 & w_3w_8 \\ 0 & w_4w_7 & w_4w_8\end{pmatrix}\right\|_F^2 .\end{aligned} \tag{8}$$

Consider $X = I_3$, $Y = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 10 & 0 \\ 0 & 0 & 4\end{pmatrix}$, which satisfies X, Y have full column rank that is the common assumption in previous work. Then $\begin{pmatrix} w_1 & 0 \\ w_2 & w_3 \\ 0 & w_4\end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 2 \\ 0 & 0\end{pmatrix}$, $\begin{pmatrix} w_5 & w_6 & 0 \\ 0 & w_7 & w_8\end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0\end{pmatrix}$ satisfies $\nabla L = \mathbf{0}$, and $L(w_1,\dots,w_8) = 8$. Besides, for any small disturbance $\delta_i, i = 1\dots,8$,

$$\begin{aligned} 2L(w_1+\delta_1,\dots,w_8+\delta_8) &= (\delta_1+\delta_5+\delta_1\delta_5)^2 + (2\delta_1+\delta_6+\delta_1\delta_6)^2 + (\delta_2+2\delta_5+\delta_2\delta_5)^2 \\ &\quad + (2\delta_2+2\delta_6+\delta_2\delta_6+3\delta_3+2\delta_7+\delta_3\delta_7)^2 \\ &\quad + (2+\delta_3)^2\delta_8^2 + (3+\delta_7)^2\delta_4^2 + (\delta_4\delta_8-4)^2 \\ &\geq (2+\delta_3)^2\delta_8^2 + (3+\delta_7)^2\delta_4^2 + (\delta_4\delta_8-4)^2 \\ &\geq 2\left[(2+\delta_3)(3+\delta_7)-4\right]|\delta_4\delta_8| + 16 \end{aligned} \tag{9}$$

Since perturbation δ_i are small, we have $(2+\delta_3)(3+\delta_7)-4 > 0$, hence $L(w_1+\delta_1,\dots,w_8+\delta_8) \geq 8$, verifying the local minimum.

However, when $\begin{pmatrix} w_1 & 0 \\ w_2 & w_3 \\ 0 & w_4\end{pmatrix} = \begin{pmatrix} \sqrt{10}/5 & 0 \\ \sqrt{10} & 0 \\ 0 & 2\end{pmatrix}$, $\begin{pmatrix} w_5 & w_6 & 0 \\ 0 & w_7 & w_8\end{pmatrix} = \begin{pmatrix} \sqrt{10}/5 & \sqrt{10} & 0 \\ 0 & 0 & 2\end{pmatrix}$, $L(w_1,\dots,w_8) = 0.18 < 8$. Hence, bad local minimum exists. □

We underline that the bad local minimum is produced from the sub-network when w_4=w_8=0. Since we encounter no bad local minimum in dense linear network, hence sparse connections indeed destroy the benign landscape because sparsity obstructs the decreasing path as Evci et al. (2019) mentioned from experiments.

3 Bad Local Minimum with Sparse First Layer

Now we turn back to dense second layer case in Section 2.3 with d_y=2 and assume X has full column rank. We give an algorithm to check the existence of spurious local minima when $\exists\, i \neq j$, $X_{-i}^T X_{-j} \neq \mathbf{0}$.

Algorithm 1 Sparse-2-Opt (Z_1, Z_2, D): Obtain the solution of two-layer sparse linear neural network with two hidden neurons.

1: **Input:** Target matrix Z_1, Z_2 and covariance diagonal matrix D.
2: Initialize $\mathbf{w}_2, d_2, \mathbf{a}_2$;
3: **while** *not converge* **do**
4: $\quad \mathbf{w}_1, d_1, \mathbf{a}_1 = SVD(Z_1 + D(Z_2 - d_2\mathbf{w}_2\mathbf{a}_2^T), 1)$;
5: $\quad \mathbf{w}_2, d_2, \mathbf{a}_2 = SVD(Z_2 + D^T(Z_1 - d_1\mathbf{w}_1\mathbf{a}_1^T), 1)$;
6: **end while**
7: $\mathbf{w}_1 = d_1\mathbf{w}_1, \mathbf{w}_2 = d_2\mathbf{w}_2$.
8: Verifying the minimal eigenvalue of Hessian matrix $\lambda_1(\nabla^2 L) \approx 0$, and the third smallest eigenvalue $\lambda_3(\nabla^2 L)$ is large away from zero, return the solution, otherwise do again from another initialization.
9: **Output:** $\mathbf{w}_1, \mathbf{a}_1, \mathbf{w}_2, \mathbf{a}_2$.

From Section 2.3, a two-layer linear network with sparse first layer is formulated as follows:

$$\min_{\tilde{W},A} \frac{1}{2} \left\| \sum_{i=1}^{d} X_{-i} \left(Z_i - \mathbf{w}_{-i}\mathbf{a}_i^T \right) \right\|_F^2, \tag{10}$$

Notice that we have no rank constraint for Z_is. Suppose singular value decomposition of $X_{-i} = U_i D_i V_i^T$ with $U_i \in \mathbb{R}^{n\times p_i}, D_i \in \mathbb{R}^{p_i\times p_i}, V_i \in \mathbb{R}^{p_i\times p_i}$, and take $D_i V_i Z_i$ and $D_i V_i \mathbf{w}_i$ as new target and variable, with a slight abuse of notation, then the problem becomes

$$\min_{\tilde{W},A} \frac{1}{2} \left\| \sum_{i=1}^{d} U_i \left(Z_i - \mathbf{w}_i\mathbf{a}_i^T \right) \right\|_F^2. \tag{11}$$

In the following, we show $d = 2$ is enough to give counter examples. Similarly, using the singular value decomposition of $U_1^T U_2 = \bar{U} D \bar{V}^T$ with a rectangle diagonal matrix $D \in \mathbb{R}^{p_1\times p_2}$. Notice U_1, U_2 are column orthogonal matrices, thus $D_{ii} \le 1$, and $|\{i : D_{ii} = 1\}|$ equals to the overlapping columns between X_{-1} and X_{-2}. Finally, the objective becomes

$$L(\mathbf{w}_1, \mathbf{w}_2, \mathbf{a}_1, \mathbf{a}_2) = \frac{1}{2}\|Z_1 - \mathbf{w}_1\mathbf{a}_1^T\|_F^2 + \frac{1}{2}\|Z_2 - \mathbf{w}_2\mathbf{a}_2^T\|_F^2 + tr[(Z_1 - \mathbf{w}_1\mathbf{a}_1^T)^T D (Z_2 - \mathbf{w}_2\mathbf{a}_2^T)]. \tag{12}$$

If we fix $\mathbf{w}_2, \mathbf{a}_2$, we can see $\mathbf{w}_1$ and $\mathbf{a}_1$ is the best rank-1 approximation of $Z_1 + D(Z_2 - \mathbf{w}_2\mathbf{a}_2^T)$, since $\mathbf{w}_1, \mathbf{a}_1$ are the solution of

$$\arg\min_{\mathbf{w}_1,\mathbf{a}_1} \|Z_1 + D(Z_2 - \mathbf{w}_2\mathbf{a}_2^T) - \mathbf{w}_1\mathbf{a}_1^T\|_F^2. \tag{13}$$

Similarly, $\mathbf{w}_2$ and $\mathbf{a}_2$ are the best rank-1 approximation of $Z_2 + D^T(Z_1 - \mathbf{w}_1\mathbf{a}_1^T)$. Empirically, we use alternating update method to find the solution in Algorithm 1 for two blocks, where $SVD(\cdot, k)$ is classical method getting largest k-th singular values and corresponding singular vectors.

Since each update do not increase the loss, leading to the convergence of sequence $\mathbf{w}_1, \mathbf{a}_1, \mathbf{w}_2, \mathbf{a}_2$. Once the algorithm converges, then the first-order condition is satisfied and two eigenvectors with zero eigenvalue of Hessian matrix are decided. Moreover, we can also prove the convergent solution is indeed a local minima (detail see Appendix B). Otherwise, we examine a local minimum using gradient descent or other optimization method started with noise, if necessary.

Based on Algorithm 1, we find several cases with bad local minima including overlapping case ($\exists i,\ D_{ii} = 1$). The results are shown in Table 1. We observe distinct gap between the local minima since our choice of elements in Z_is is small. In non-overlapping setting, the algorithm reaches the local min quickly and shows several different examples. As for overlapping setting, a simple construction is leaving out the repeated feature away with zero items in Z_is, though we also show bad-min applied with overlapping data feature in Row 4 in Table 1.

It is interesting to note that for $d = 2$, only at most two local minimum are found, and we can easily broaden the alternating update method into general d case in Appendix D, that will also verify

Table 1: Examples found by algorithm with spurious local minimum. All experiments run 600 iterations, except last one with 1000 iterations.

Z_1	Z_2	D	$\lambda_3(\nabla^2 L)$	$\lambda_1(\nabla^2 L)$	$\|\nabla L\|_2$	Objective L
$\begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 0.9 & 0 \\ 0 & 0.8 \end{pmatrix}$	$3.6 \cdot 10^{-2}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.3694461462878
			$5.5 \cdot 10^{-2}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.3081376225475
$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.9 \end{pmatrix}$	$2.1 \cdot 10^{-1}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.5143043518476
			$1.4 \cdot 10^{-1}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.6781647585271
$\begin{pmatrix} -2 & 0 \\ 0 & -1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ -2 & 2 \end{pmatrix}$	$\begin{pmatrix} 0.8 & 0 \\ 0 & 0.1 \end{pmatrix}$	$3.4 \cdot 10^{-1}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.5373672988360
			$1.7 \cdot 10^{-1}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.6805528480352
$\begin{pmatrix} -1 & 0 \\ 1 & 1 \\ -1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 1 & 1 \\ -2 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & 0.8 \end{pmatrix}$	$1.7 \cdot 10^{-1}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.8980944246693
			$2.5 \cdot 10^{-1}$	$0 \sim 10^{-14}$	$< 10^{-14}$	0.4712847600704

similar observation: at most d local minimum produced by a sparse-first-layer network with hidden d nodes, which leaves as future work. Overall, sparsity breaks the original matrix structure, leading to additional low rank constraint in this case, and still cuts out the decreasing path in original fully-connected network.

Besides, a descent algorithm still will diverge to infinity. For instance, example in Appendix A shows that there is a sequence diverging to infinity while the function values are decreasing and converging.

4 Landscape of Deep Sparse Linear Networks

In this section, we briefly extend Thm 1 and Thm 2 to deep sparse linear networks and leave the proof in Appendix F. The intuition is that deep linear networks have similar landscape property in shallow case as Lu & Kawaguchi (2017); Eftekhari (2020) showed. However, understanding the landscape of a random deep linear network is still complicated.

Theorem 6 *For a deep sparse linear neural network with scalar output ($d_y = 1$) and any sparse structure, every local minimum is a global minimum.*

How to obtain global minimum in scalar case: One way is setting the first-layer weight as the global minimum solution in two-layer case with $a_i = 1$, then the remaining layer uniformly distribute the output of each node to the next layer. For example, if one node has k output connections, then each connection assigns weight $1/k$, hence the sum of each layer output remains the best solution to approximate target Y (see the right graph of Figure 2 for example).

Theorem 7 *For a deep sparse linear neural network with sparse first layer and dense other layers, assume that X, Y have full column rank, and $\forall\ i \neq j$, $X_{-i}^T X_{-j} = \mathbf{0}$. Besides, $d_i \geq \min\{d_1, d_y\}, \forall i \geq 1$, then every local minimum is a global minimum.*

5 Discussion

We discuss the landscape of sparse linear network, showing positive results that no spurious local minimum in scalar case, or separated first layer with orthogonal data. However, we discover bad local minima when previous conditions are violated in general sparse two-layer linear network, one is generated from low rank constraint, another is produced from sub-sparse structure. Both case shows that sparsity cuts out the decreasing path in original fully-connected network. Since dense linear networks possess benign landscape, we conclude sparsity or network pruning destroys the favourable solutions. However, some heuristic algorithms combining training and pruning still work well in practice, leading mystery of modern network pruning methods and sparse network design. Other interesting questions for future research include understanding the gap between global minima and spurious local minima, or showing similar performance of bad-min, particularly, combining with pruning algorithms.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Pierre Baldi and Zhiqin Lu. Complex-valued autoencoders. *Neural Networks*, 33:136–147, 2012.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

Miguel A Carreira-Perpinán and Yerlan Idelbayev. learning-compression algorithms for neural net pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8532–8541, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Simon S Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. *arXiv preprint arXiv:1901.08572*, 2019.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. *arXiv preprint arXiv:2002.09852*, 2020.

Utku Evci, Fabian Pedregosa, Aidan Gomez, and Erich Elsen. The difficulty of training sparse neural networks. *arXiv preprint arXiv:1906.10732*, 2019.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015a.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pp. 1135–1143, 2015b.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pp. 586–594, 2016.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

Dawei Li, Tian Ding, and Ruoyu Sun. On the benefit of width for neural networks: Disappearance of bad basins. *arXiv*, pp. arXiv–1812, 2018.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. In *Advances in neural information processing systems*, pp. 3288–3298, 2017a.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017b.

Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. *arXiv preprint arXiv:2002.00585*, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. *arXiv preprint arXiv:1701.05369*, 2017.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pp. 4433–4441. PMLR, 2018.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Weight sharing is crucial to succesful optimization. *arXiv preprint arXiv:1706.00687*, 2017.

Ruoyu Sun, Dawei Li, Shiyu Liang, Tian Ding, and Rayadurgam Srikant. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. *arXiv preprint arXiv:1802.03487*, 2018.

Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.