Spurious Local Minima are Common in Two-Layer ReLU Neural Networks

Itay Safran [1] **Ohad Shamir** [1]

Abstract

We consider the optimization problem associated with training simple ReLU neural networks of the form $\mathbf{x} \mapsto \sum_{i=1}^{k} \max\{0, \mathbf{w}_i^\top \mathbf{x}\}$ with respect to the squared loss. We provide a computer-assisted proof that even if the input distribution is standard Gaussian, even if the dimension is arbitrarily large, and even if the target values are generated by such a network, with orthonormal parameter vectors, the problem can still have spurious local minima once $6 \leq k \leq 20$. By a concentration of measure argument, this implies that in high input dimensions, *nearly all* target networks of the relevant sizes lead to spurious local minima. Moreover, we conduct experiments which show that the probability of hitting such local minima is quite high, and increasing with the network size. On the positive side, mild over-parameterization appears to drastically reduce such local minima, indicating that an over-parameterization assumption is necessary to get a positive result in this setting.

1. Introduction

One of the biggest mysteries of deep learning is why neural networks are successfully trained in practice using gradient-based methods, despite the inherent non-convexity of the associated optimization problem. For example, non-convex problems can have poor local minima, which will cause any local search method (and in particular, gradient-based ones) to fail. Thus, it is natural to ask what types of assumptions, in the context of training neural networks, might mitigate such problems. For example, recent work has shown that other non-convex learning problems, such as phase retrieval, matrix completion, dictionary learning, and tensor decomposition, do not have spurious local minima under suitable assumptions, in which case local search methods have a chance of succeeding (e.g., (Ge et al., 2015; Sun et al., 2015; Ge et al., 2016; Bhojanapalli et al., 2016)). Is it possible to prove similar positive results for neural networks?

In this paper, we focus on perhaps the simplest non-trivial ReLU neural networks, namely predictors of the form

$$\mathbf{x} \mapsto \sum_{i=1}^{k} [\mathbf{w}_i^\top \mathbf{x}]_+$$

for some $k > 1$, where $[z]_+ = \max\{0, z\}$ is the ReLU function, $\mathbf{x}$ is a vector in $\mathbb{R}^d$, and $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are parameter vectors. We consider directly optimizing the expected squared loss, where the input is standard Gaussian, and in the realizable case – namely, that the target values are generated by a network of a similar architecture:

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_k} \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{2} \left(\sum_{i=1}^{k} [\mathbf{w}_i^\top \mathbf{x}]_+ - \sum_{i=1}^{k} [\mathbf{v}_i^\top \mathbf{x}]_+ \right)^2 \right]. \tag{1}$$

Note that here, the choice $\mathbf{w}_i = \mathbf{v}_{\sigma(i)}$ (for all $i = 1, \ldots, k$ and any permutation σ) is a global minimum with zero expected loss. Several recent papers analyzed such objectives, in the hope of showing that it does not suffer from spurious local minima (see related work below for more details).

Our main contribution is to prove that unfortunately, this conjecture is false, and that Eq. (1) indeed has spurious local minima once $6 \leq k \leq 20$. Moreover, this is true even if the dimension is arbitrarily large, and even if we assume that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are orthonormal vectors. In fact, since in high dimensions randomly-chosen vectors are approximately orthogonal, and the landscape of the objective function is robust to small perturbations, we can show that spurious local minima exist for *nearly all* neural network problems as in Eq. (1), in high enough dimension (with respect to, say, a Gaussian distribution over $\mathbf{v}_1, \ldots, \mathbf{v}_k$). Moreover, we show experimentally that these local minima are not pathological, and that standard gradient descent can easily get trapped in them, with a probability which seems to increase towards 1 with the network size.

[1] Weizmann Institute of Science, Rehovot, Israel. Correspondence to: Itay Safran <itay.safran@weizmann.ac.il>, Ohad Shamir <ohad.shamir@weizmann.ac.il>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

Our proof technique is a bit unorthodox. Although it is possible to write down the gradient of Eq. (1) in closed form (without the expectation), it is not clear how to get analytical expressions for its roots, and hence characterize the stationary points of Eq. (1). As far as we know, an analytical expression for the roots might not even exist. Instead, we employed the following strategy: We ran standard gradient descent with random initialization on the objective function, until we reached a point which is both suboptimal (function value being significantly higher than 0); approximate stationary (gradient norm very close to 0); and with a strictly positive definite Hessian (with minimal eigenvalue significantly larger than 0). We use a computer to verify these conditions in a formal manner, avoiding floating-point arithmetic and the possibility of rounding errors. Relying on these numbers, we employ a Taylor expansion argument, to show that we must have arrived at a point very close to a local (non-global) minimum of Eq. (1), hence establishing the existence of such minima.

On the more positive side, we show that an additional *over-parameterization* assumption appears to be very effective in mitigating these local minima issues: Namely, we use a network larger than that needed with unbounded computational power, and replace Eq. (1) with

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_k} \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\frac{1}{2}\left(\sum_{i=1}^{n}[\mathbf{w}_i^\top\mathbf{x}]_+ - \sum_{i=1}^{k}[\mathbf{v}_i^\top\mathbf{x}]_+\right)^2\right], \tag{2}$$

where $n > k$. In our experiments with k, n up to size 20, we observe that whereas $n = k$ leads to plenty of local minima, $n = k+1$ leads to much fewer local minima, whereas no local minima were encountered once $n \geq k+2$ (although those might still exist for larger values of k, n than those we tried). Thus, although Eq. (1) has local minima, we conjecture that Eq. (2) might still be proven to have no bad local minima, but this would *necessarily* require n to be sufficiently larger than k.

The paper is structured as follows: After surveying related work below, we provide our main results and proof ideas in Sec. 2. Sec. 3 provides additional experimental details about the local minima found, as well empirical evidence about the likelihood of reaching them using gradient descent. Some of the proofs are provided in Sec. 4, with the rest provided in the appendix.

1.1. Related Work

There is a large and rapidly increasing literature on the optimization theory of neural networks, surveying all of which is well outside our scope. Thus, in this subsection, we only briefly survey the works most relevant to ours.

We begin by noting that when minimizing the average loss over some arbitrary finite dataset, it is easy to construct problems where even for a single neuron ($k = 1$ in Eq. (1)), there are many spurious local minima (e.g., (Auer et al., 1996; Swirszcz et al., 2016)). Moreover, the probability of starting at a basin of such local minima is exponentially high in the dimension (Safran & Shamir, 2016). On the other hand, it is known that if the network is over-parameterized, and large enough compared to the data size, then there are no local minima (Poston et al., 1991; Livni et al., 2014; Haeffele & Vidal, 2015; Zhang et al., 2016; Soudry & Carmon, 2016; Soltanolkotabi et al., 2017; Nguyen & Hein, 2017; Boob & Lan, 2017). In any case, neither these positive nor negative results apply here, as we are interested in the expected (population) loss with respect to the Gaussian distribution, which is of course non-discrete. Also, several recent works have studied learning neural networks under a Gaussian distribution assumption (e.g., Janzamin et al. (2015); Brutzkus & Globerson (2017); Du et al. (2017); Li & Yuan (2017); Feizi et al. (2017); Zhang et al. (2017); Ge et al. (2017)), but using a network architecture different than ours, or focusing on algorithms rather than the geometry of the optimization problem. Finally, Shamir (2016) provides hardness results for training neural networks even under distributional assumptions, but these do not apply when making strong assumptions on *both* the input distribution and the network generating the data, as we do here.

For Eq. (1), a few works have shown that there are no spurious local minima, or that gradient descent will succeed in reaching a global minimum, provided the $\mathbf{v}_i$ vectors are in general position or orthogonal (Zhong et al., 2017; Soltanolkotabi et al., 2017; Tian, 2017). However, these results either apply only to $k = 1$, assume the algorithm is initialized close to a global optimum, or analyze the geometry of the problem only on some restricted subset of the parameter space.

The empirical observation that gradient-based methods may not work well on Eq. (1) has been made in Livni et al. (2014), and more recently in Ge et al. (2017). Moreover, Livni et al. (2014) empirically observed that over-parameterization seems to help. However, our focus here is to *prove* the existence of such local minima, as well as more precisely quantify their behavior as a function of the network sizes.

2. Main Result and Proof Technique

Before we begin, a small note on terminology: When referring to local minima of a function F on Euclidean space, we always mean spurious local minima (i.e., points $\mathbf{w}$ such that $\inf_{\mathbf{w}} F(\mathbf{w}) < F(\mathbf{w}) \leq F(\mathbf{w}')$ for all $\mathbf{w}'$ in some open neighborhood of $\mathbf{w}$).

Our basic result is the following:

Theorem 1. *Consider the optimization problem*

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_n\in\mathbb{R}^k} \mathbb{E}_{\mathbf{x}}\left[\frac{1}{2}\left(\sum_{i=1}^{n}[\mathbf{w}_i^\top\mathbf{x}]_+ - \sum_{i=1}^{k}[\mathbf{v}_i^\top\mathbf{x}]_+\right)^2\right],$$

where $\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ *and* $\mathbf{v}_1,\ldots,\mathbf{v}_k$ *are orthogonal unit vectors in* $\mathbb{R}^k$*. Then for* $n=k\in\{6,7,\ldots,20\}$*, as well as* $(k,n)\in\{(8,9),(10,11),(11,12),\ldots,(19,20)\}$*, the objective function above has spurious local minima.*

Remark 1. *For* k,n *smaller than* 6*, we were unable to find local minima using our proof technique, since gradient descent always seemed to converge to a global minimum. Also, although we have verified the theorem only up to* $k,n\le 20$*, the result strongly suggests that there are local minima for larger values as well. See Sec. 3 for some examples of the local minima found.*

The theorem assumes a fixed input dimension, and a particular choice of $\mathbf{v}_1,\ldots,\mathbf{v}_k$. However, due to the fact that gradient descent is invariant to orthonormal reparameterizations, these assumptions are not necessary and can be relaxed, as demonstrated by the following corollary:

Corollary 1. *Thm. 1 also applies if the space* $\mathbb{R}^k$ *is replaced by* $\mathbb{R}^d$ *for any* $d>k$ *(with* $\mathbf{x}$ *distributed as a standard Gaussian in that space). Moreover, if* $\mathbf{v}_1,\ldots,\mathbf{v}_k$ *are chosen i.i.d. from a Gaussian distribution* $\mathcal{N}(\mathbf{0},c\mathbf{I})$ *(for any* $c>0$*), the theorem still holds with probability at least* $1-\exp(-\Omega(d))$*.*

Remark 2. *The corollary is not specific to a Gaussian distribution over* $\mathbf{v}_1,\ldots,\mathbf{v}_k$*, and can be generalized to any distribution for which* $\mathbf{v}_1,\ldots,\mathbf{v}_k$ *are approximately orthogonal and of the same norm in high dimensions (see below for details).*

We now turn to explain how these results are derived, starting with Thm. 1. In what follows, we let $\mathbf{w}_1^n=(\mathbf{w}_1,\ldots,\mathbf{w}_n)\in\mathbb{R}^{kn}$ be the vector of parameters, and let $F(\mathbf{w}_1^n)$ be the objective function defined in Thm. 1 (assuming k,n are fixed). We will also assume that F is thrice-differentiable in a neighborhood of $\mathbf{w}_1^n$ (which will be shown to be true as part of our proofs), with a gradient $\nabla F(\cdot)$ and a Hessian $\nabla^2F(\cdot)$.

Clearly, a global minimum of F is obtained by $\mathbf{w}_i=\mathbf{v}_i$ for all $i=1,\ldots,k$ (and $\mathbf{w}_i=\mathbf{0}$ otherwise), in which case F attains a global minimum of 0. Thus, to prove Thm. 1, it is sufficient to find a point $\mathbf{w}_1^n\in\mathbb{R}^{kn}$ such that $\nabla F(\mathbf{w}_1^n)=0$, $\nabla^2F(\mathbf{w}_1^n)\succeq 0$, and $F(\mathbf{w}_1^n)>0$. The major difficulty is showing the existence of points where the first condition is fulfilled: Gradient descent allows us to find points where $\nabla F(\mathbf{w}_1^n)\approx 0$, but it is very unlikely to return a point where $\nabla F(\mathbf{w}_1^n)=0$ exactly. Instead, we use a Taylor-expansion argument (detailed below), to show that if we found a point such that $\nabla F(\mathbf{w}_1^n)$ is sufficiently close to 0, as well as $\nabla^2F(\mathbf{w}_1^n)\succ 0$ and $F(\mathbf{w}_1^n)>0$, then $\mathbf{w}_1^n$ must be close to a local minimum.

The second-order Taylor expansion of a multivariate, thrice-differentiable function F about a point $\mathbf{w}_1^n\in\mathbb{R}^{kn}$, in a direction given by a unit vector $\mathbf{u}\in\mathbb{R}^{kn}$ and using a Lagrange remainder term, is given by

$$\begin{aligned}&F\left(\mathbf{w}_1^n+t\mathbf{u}\right)\\ =&F\left(\mathbf{w}_1^n\right)+t\sum_{i_1}\frac{\partial}{\partial\mathbf{w}_{1,i_1}^n}F\left(\mathbf{w}_1^n\right)u_{i_1}\\ &+\frac{1}{2}t^2\sum_{i_1,i_2}\frac{\partial^2}{\partial\mathbf{w}_{1,i_1}^n\partial\mathbf{w}_{1,i_2}^n}F\left(\mathbf{w}_1^n\right)u_{i_1}u_{i_2}\\ &+\frac{1}{6}t^3\sum_{i_1,i_2,i_3}\frac{\partial^3}{\partial\mathbf{w}_{1,i_1}^n\partial\mathbf{w}_{1,i_2}^n\partial\mathbf{w}_{1,i_3}^n}F\left(\mathbf{w}_1^n+\xi\mathbf{u}\right)u_{i_1}u_{i_2}u_{i_3},\end{aligned}$$

for some $\xi\in(0,t)$, and where $\mathbf{w}_{1,i}^n$ denotes the i-th coordinate of $\mathbf{w}_1^n$. Denoting the remainder term as $R_{\mathbf{w}_1^n,\mathbf{u}}$, we have

$$\begin{aligned}F\left(\mathbf{w}_1^n+t\mathbf{u}\right)=&F\left(\mathbf{w}_1^n\right)+t\nabla F\left(\mathbf{w}_1^n\right)^\top\mathbf{u}\\ &+\frac{1}{2}t^2\mathbf{u}^\top\nabla^2F\left(\mathbf{w}_1^n\right)\mathbf{u}+\frac{1}{6}t^3R_{\mathbf{w}_1^n,\mathbf{u}}\,.\end{aligned}\tag{3}$$

Now, suppose that the point $\mathbf{w}_1^n$ we obtain by gradient descent satisfies $||\nabla F(\mathbf{w}_1^n)||\le\epsilon$, $\nabla^2F(\mathbf{w}_1^n)\succeq\lambda_{\min}\cdot\mathbf{I}$ and $|R_{\mathbf{w}_1^n,\mathbf{u}}|\le B$ (for some positive $\lambda_{\min},\epsilon,B$), uniformly for all unit vectors $\mathbf{u}$. By the Taylor expansion above, this implies that for all unit $\mathbf{u}$,

$$\begin{aligned}F(\mathbf{w}_1^n+t\mathbf{u})\ \ge&\ F(\mathbf{w}_1^n)-t\,||\nabla F(\mathbf{w}_1^n)||\cdot||\mathbf{u}||\\ &+\frac{t^2}{2}\lambda_{\min}\,||\mathbf{u}||^2-\frac{t^3}{6}B\\ =&\ F(\mathbf{w}_1^n)-\epsilon t+\frac{\lambda_{\min}t^2}{2}-\frac{Bt^3}{6}\\ =&\ F(\mathbf{w}_1^n)+t\left(\frac{\lambda_{\min}}{2}t-\frac{B}{6}t^2-\epsilon\right).\end{aligned}$$

An elementary calculation reveals that the term $t\left(\frac{\lambda_{\min}}{2}t-\frac{B}{6}t^2-\epsilon\right)$ is strictly positive for any t in the open interval of

$$\frac{3\lambda_{\min}\pm\sqrt{9\lambda_{\min}^2-24B\epsilon}}{2B}$$

(and in particular, in the closed interval of $\frac{3\lambda_{\min}\pm\sqrt{9\lambda_{\min}^2-25B\epsilon}}{2B}$). This implies that there is some small closed ball $\bar{B}_t$ of radius $t>0$ centered at $\mathbf{w}_1^n$ (and with boundary S), such that $F(\mathbf{w}_1^n)<\min_{\mathbf{w}_1'^n\in S}F(\mathbf{w}_1'^n)$. Moreover, since F is continuous, it is minimized over $\bar{B}_t$

at some point $\mathbf{w}_1^{*n}$. But then

$$F(\mathbf{w}_1^{*n}) = \min_{\mathbf{w}_1'^n \in B} F(\mathbf{w}_1'^n) \le F(\mathbf{w}_1^n) < \min_{\mathbf{w}_1'^n \in S} F(\mathbf{w}_1'^n), \tag{4}$$

so $\mathbf{w}_1^{*n}$ must reside in the interior of $\bar{B}_t$. Thus, it is minimal in an open neighborhood containing it, hence it is a local minimum. Overall, we have arrived at the following key lemma:

Lemma 1. *Assume that* $||\nabla F(\mathbf{w}_1^n)|| \le \epsilon$, $\left|R_{\mathbf{w}_1^n,\mathbf{u}}\right| \le B$ *for some* $\epsilon, B > 0$ *and all unit vectors* $\mathbf{u}$*, and let* $\lambda_{\min} > 0$ *denote the smallest eigenvalue of* $\nabla^2 F(\mathbf{w}_1^n)$*. If* $9\lambda_{\min}^2 - 25B\epsilon \ge 0$*, then the function* F *contains a local minimum, within a distance of at most*

$$r := \frac{3\lambda_{\min} - \sqrt{9\lambda_{\min}^2 - 25B\epsilon}}{2B}$$

from $\mathbf{w}_1^n$*.*

The only missing element is that the local minimum might be a global minimum. To rule this out, one can simply use the fact that F is a Lipschitz function, so that if $F(\mathbf{w}_1^n)$ is much larger than 0, the neighboring local minimum can't have a value of 0, and hence cannot be global:

Lemma 2. *Under the conditions of Lemma 1, if it also holds that*

$$F(\mathbf{w}_1^n) > r^2\left(\frac{1}{2} + (n^2 - n)\left(\frac{(\max_i ||\mathbf{w}_i|| + r)}{2\pi(\min_i ||\mathbf{w}_i|| - r)} + \frac{1}{2}\right) + \frac{nk \cdot \max_i ||\mathbf{v}_i||}{2\pi(\min_i ||\mathbf{w}_i|| - r)} + r\epsilon\right), \tag{5}$$

then the local minimum is non-global.

The formal proof of this lemma appears in Subsection A.3 in the appendix.

Most of the technical proof of Thm. 1 consists in rigorously verifying the conditions of Lemma 1 and Lemma 2. A major hurdle is that floating-point calculations are not guaranteed to be accurate (due to the possibility of round-off and other errors), so for a formal proof, one needs to use software that comes with guaranteed numerical accuracy. In our work, we chose to use variable precision arithmetic (VPA), a standard package of MATLAB which is based on symbolic arithmetic, and allows performing elementary numerical computations with an arbitrary number of guaranteed digits of precision. The main technical issue we faced is that some calculations are not easily done with a few elementary arithmetical operations (in particular, the standard way to compute $\lambda_{\min}$ would be via a spectral decomposition of the Hessian matrix). The bulk of the proof consists of showing how we bound the quantities relevant to Lemma 1 in an elementary manner.

Finally, we turn to discuss how Corollary 1 is proven, given Thm. 1 (see Subsection 4.3 for a more formal derivation). The proof idea is that the objective does not have any "non-trivial" structure outside the span of $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Therefore, if we take a local minima for $\mathbb{R}^k$, and pad it with $d - k$ zeros, we get a point in $\mathbb{R}^d$ for which the gradient's norm is unchanged, the Hessian has the same spectrum for any $d \ge k + 1$, and the third derivatives are still bounded. Hence, that point is a local minimum in the higher-dimensional problem as well. As to the second part of the corollary, the only property of the Gaussian distribution we need is that in high dimensions, if we sample $\mathbf{v}_1, \ldots, \mathbf{v}_k$, then we are overwhelmingly likely to get approximately orthogonal vectors with approximately the same norm. Hence, up to rotation and scaling, we get a small perturbation $\tilde{F}$ of the objective F considered in Thm. 1. Moreover, for large enough d, we can make the perturbation arbitrarily small, uniformly in some compact domain. Now, recall that we prove the existence of some local minimum $\mathbf{w}_1^{*n}$, by showing that $F(\mathbf{w}_1^n) < \min_{\mathbf{w}_1'^n \in S} F(\mathbf{w}_1'^n)$ in some small sphere S enclosing $\mathbf{w}_1^n$. If the perturbations are small enough, we also have $\tilde{F}(\mathbf{w}_1^n) < \min_{\mathbf{w}_1'^n \in S} \tilde{F}(\mathbf{w}_1'^n)$, which by arguments similar to before, imply that $\mathbf{w}_1^n$ is close to a local minimum of $\tilde{F}$.

3. Experiments

So far, we proved the *existence* of local minima for the objective function in Eq. (2). However, this does not say anything about the likelihood of gradient descent to reach them. We now turn to study this question empirically.

For each value of (k, n), where $k \in [20]$ and $n \in \{k, \ldots, 20\}$, we ran 1000 instantiations of gradient descent on the objective in Eq. (2), each starting from a different random initialization [1]. Each instantiation was ran with a fixed step size of 0.1, until reaching a candidate stationary point / local minima (the stopping criterion was that the gradient norm w.r.t. any $\mathbf{w}_i$ is at most 10^{-9}). Points obtaining objective values less than 10^{-3} were ignored as those are likely to be close to a global minimum. Interestingly, no points with value between 10^{-3} and 10^{-2} were found. For all remaining candidate points, we verified that the conditions in Lemmas 1 and 2 are met[2] to

[1] We used standard Xavier initialization: Each weight vector $\mathbf{w}_i$ was samples i.i.d. from a Gaussian distribution in $\mathbb{R}^k$, with zero mean and covariance $\frac{1}{k}\mathbf{I}$.

[2] Since running our algorithm for all suspicious points found on all architectures is time consuming, we instead identified points that are equivalent up to permutations on the order of neurons and of the data coordinates, since the objective is invariant under such permutations. By bounding the maximal Euclidean distance between these points and using the Lipschitzness of the objective and its Hessian (see Thm. 4 and Lemma 7), this allowed us to run the algorithm on a single representative from a family of equivalent points and speed up the running time drastically. Also,

conclude that these points are indeed close to spurious local minima (in all cases, the distance turned out to be less than $2 \cdot 10^{-6}$). Our verification process included verifying thrice-differentiability in the enclosing balls containing the minima by asserting they contain no singular points, hence the objective is an analytical expression when restricted to these balls where differentiability follows.

In Tables 1 and 2, we summarize the percentage of instantiations which were verified to converge close to a spurious local minimum, as a function of k, n. We note that among candidate points found, only a tiny fraction could not be verified to be local minima (this only occured for network sizes $(k,n) \in \{(15,16),(17,18),(20,20)\}$, and consist only $0.1\%, 2.4\%, 0.9\%$ of the instantiations respectively). In the tables, we also provide the minimal eigenvalue of the Hessian of the objective, and the objective value (or equivalently, the optimization error) at the points found, averaged over the instantiations[3]. Note that since the minimal eigenvalue is strictly positive and varies slightly inside the enclosing ball, this indicates that these are in fact strict local minima. As the tables demonstrate, the probability of converging to a spurious local minimum increases rapidly with k, n, and suggests that it eventually becomes overwhelming as long as $n \approx k$. However, on a positive note, mild over-parameterization seems to remedy this, as no local minima were found for $n \geq k+2$ where $n \leq 20$, and local minima for $n = k+1$ are much more scarce than for $n = k$. We leave the investigation of local minima for larger values of k, n to future work.

In Fig. 1, we show the distribution of the objective values obtained in the points found, over the 1000 instantiations of several architectures. The figure clearly indicates that apart from a higher chance of converging to local minima, larger architectures also tend to have worse values attained on these minima.

Finally, in examples 1 and 2 below, we present some specific local minima found for $n = k = 6$ and $k = 8, n = 9$, and discuss their properties. We note that these are the smallest networks (with $n = k$ and $n \neq k$ respectively) for which we were able to find such points.

Table 1. Spurious local minima found for $n = k$

k	n	% of runs converging to local minima	Average minimal eigenvalue	Average objective value
6	6	0.3%	0.0047	0.025
7	7	5.5%	0.014	0.023
8	8	12.6%	0.021	0.021
9	9	21.8%	0.027	0.02
10	10	34.6%	0.03	0.022
11	11	45.5%	0.034	0.022
12	12	58.5%	0.035	0.021
13	13	73%	0.037	0.022
14	14	73.6%	0.038	0.023
15	15	80.3%	0.038	0.024
16	16	85.1%	0.038	0.027
17	17	89.7%	0.039	0.027
18	18	90%	0.039	0.029
19	19	93.4%	0.038	0.031
20	20	94%	0.038	0.033

Table 2. Spurious local minima found for $n \neq k$

k	n	% of runs converging to local minima	Average minimal eigenvalue	Average objective value
8	9	0.1%	0.0059	0.021
10	11	0.1%	0.0057	0.018
11	12	0.1%	0.0056	0.017
12	13	0.3%	0.0054	0.016
13	14	1.5%	0.0015	0.038
14	15	5.5%	0.002	0.033
15	16	10.1%	0.004	0.032
16	17	18%	0.0055	0.031
17	18	20.9%	0.007	0.031
18	19	36.9%	0.0064	0.028
19	20	49.1%	0.0077	0.027

Example 1. *Out of 1000 gradient descent instantiations for $n = k = 6$, three converged close to a local minimum. All three were verified to be essentially identical (after permuting the neurons and up to an Euclidean distance of $1.2 \cdot 10^{-8}$), and have the following form:*

$$\begin{bmatrix} -0.602 & 0.308 & 0.308 & 0.308 & 0.308 & 0.308 \\ 0.225 & 0.987 & -0.050 & -0.050 & -0.050 & -0.050 \\ 0.225 & -0.050 & 0.987 & -0.050 & -0.050 & -0.050 \\ 0.225 & -0.050 & -0.050 & 0.987 & -0.050 & -0.050 \\ 0.225 & -0.050 & -0.050 & -0.050 & 0.987 & -0.050 \\ 0.225 & -0.050 & -0.050 & -0.050 & -0.050 & 0.987 \end{bmatrix},$$

where the parameter vector of each of the 6 *neurons corresponds to a column of the matrix denoted* $\mathbf{w}_1^6$. *The Hessian of the objective at* $\mathbf{w}_1^6$, $\nabla^2 F\left(\mathbf{w}_1^6\right)$, *was confirmed to have minimal eigenvalue* $\lambda_{\min}\left(\nabla^2 F\left(\mathbf{w}_1^6\right)\right) \geq 0.004699$. *This implied that all three suspicious points found for* $n = k = 6$ *are of distance at most* $r = 1.12 \cdot 10^{-7}$ *from a local minimum with objective value at least* 0.02508.

Example 2. *Out of 1000 gradient descent initializations for $k = 8, n = 9$, one converged to a local minimum. The*

the objective was tested to be thrice-differentiable in all enclosing balls of radii returned by the algorithm. Specifically, we ensured that no two such balls intersect (which results in two identical neurons, where the objective is not thrice-differentiable) and that no ball contains the origin (which results in a neuron with weight $\mathbf{0}$, where again the objective is not thrice-differentiable).

[3] Since all points are extremely close to a local minimum, the objective at the minimum is essentially the same, up to a deviation on order less than $1.1 \cdot 10^{-9}$. Also, the minimal eigenvalues vary by at most $5.7 \cdot 10^{-4}$.

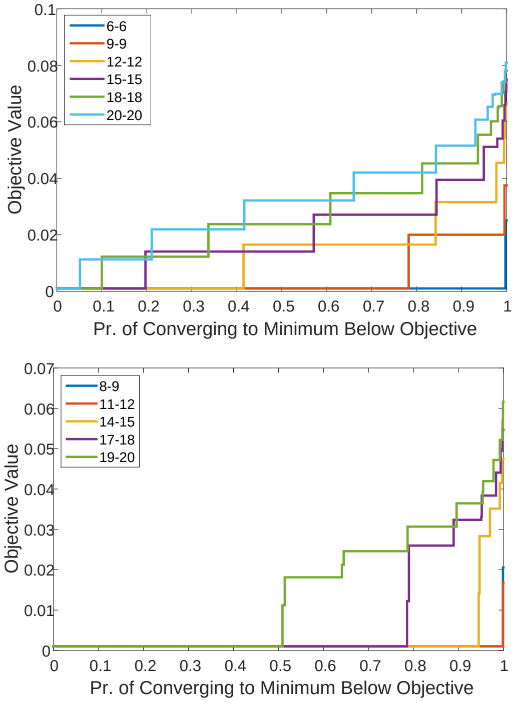

Figure 1. The empirical probability of converging to a minimum with objective value smaller than a given quantity, out of the 1000 runs. Different lines correspond to different choices of (k, n). Best viewed in color.

point found, denoted $\mathbf{w}_1^9$*, is given below:*

$$\begin{bmatrix} 0.99 & -0.03 & \dots & -0.03 & -0.03 & 0.13 & 0.07 \\ -0.03 & 0.99 & \dots & -0.03 & -0.03 & 0.13 & 0.07 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ -0.03 & -0.03 & \dots & 0.99 & -0.03 & 0.13 & 0.07 \\ -0.03 & -0.03 & \dots & -0.03 & 0.99 & 0.13 & 0.07 \\ 0.23 & 0.23 & \dots & 0.23 & 0.23 & -0.19 & -0.49 \end{bmatrix},$$

where the parameter vector of each of the 9 *neurons corresponds to a column of* $\mathbf{w}_1^9$*. The Hessian of the objective at* $\mathbf{w}_1^9$, $\nabla^2 F\left(\mathbf{w}_1^9\right)$*, was confirmed to have minimal eigenvalue* $\lambda_{\min}\left(\nabla^2 F\left(\mathbf{w}_1^9\right)\right) \geq 0.005944$*. This implied that* $\mathbf{w}_1^9$ *is of distance at most* $r = 7.8 \cdot 10^{-8}$ *from a local minimum with objective value at least* 0.02056.

It is interesting to note that the points found in examples 1 and 2, as well as all other local minima detected, have a nice symmetric structure: We see that most of the trained neurons are very close to the target neurons in most of the dimensions. Also, many of the entries appear to be the same. Surprisingly, although such constructions might seem brittle, these are indeed strict local minima. Moreover, the probability of converging to such points becomes very large as the network size increases as demonstrated by our experiments.

4. Proofs of Thm. 1 and Corollary 1

In this section, we provide a formal proof of Corollary 1, as well as an outline of the proof of Thm. 1. We also provide closed-form expressions for the objective and its derivatives. Missing parts of the proofs are provided in the appendix.

In the proofs, we use bold-faced letters (e.g., $\mathbf{w}$) to denote vectors, barred bold-faced letters (e.g., $\bar{\mathbf{w}}$) to denote vectors normalized to unit Euclidean norm, and capital letters to generally denote matrices. Given a natural number k, we let $[k]$ be shorthand for $\{1, \dots, k\}$. Given a matrix M, $||M||_{\text{sp}}$ denotes its spectral norm.

4.1. Proof of Thm. 1

To prove Thm. 1 for some (k, n), it is enough to consider some particular choice of orthogonal $\mathbf{v}_1, \dots, \mathbf{v}_k$, since any other choice amounts to rotating or reflecting the same objective function (which of course does not change the existence or non-existence of its local minima). In particular, we chose these vectors to simply be the standard basis vectors in $\mathbb{R}^k$.

As we show in Subsection 4.2 below, the objective function in Eq. (2) can be written in an explicit form (without the expectation term), as well as its gradients and Hessians. We first ran standard gradient descent, starting from random initialization and using a fixed step size of 0.1, till we reached a point $\mathbf{w}_1^n$, such that the gradient norm w.r.t. any $\mathbf{w}_i$ is at most 10^{-9}. Given this point, we use Lemma 1 and Lemma 2 to prove that it is close to a local minimum. Specifically, we built code which does the following:

1. Provide a rigorous upper bound on the norm of the gradient at a given point $\mathbf{w}_1^n$ (since we have a closed-form expression for the gradient, this only requires elementary calculations).

2. Provide a rigorous lower bound on the minimal eigenvalue of $\nabla^2 F(\mathbf{w}_1^n)$: This is the technically most demanding part, and the derivation of the algorithm is presented in Subsection A.1 in the appendix.

3. Provide a rigorous upper bound B on the remainder term $R_{\mathbf{w}_1^n, \mathbf{u}}$ (see Subsection A.2 in the appendix for the relevant calculations).

4. Provide a rigorous Lipschitz bound on the objective $F\left(\mathbf{w}_1^n\right)$, establishing Lemma 2 (see Subsection A.3 in the appendix for the relevant calculations).

We used MATLAB (version 2017b) to perform all floating-point computations, and its associated MATLAB VPA package to perform the exact symbolic computations. The code we used is freely available at https://github.com/ItaySafran/OneLayerGDconvergence.git. For any candidate local minimum, the verification took from less than a minute up to a few hours, depending on the size of k, n, when running on Intel Xeon E5 processors (ranging from E5-2430 to E5-2660).

4.2. Closed-form Expressions for $F, \nabla F$ and $\nabla^2 F$

For convenience, we will now state closed-form expressions (without an expectation) for the objective function F in Eq. (2), its gradient and its Hessian. These are also the expressions used in the code we built to verify the conditions of Lemma 1 and Lemma 2. First, we have that

$$\begin{aligned} F\left(\mathbf{w}_1^n\right) &= \frac{1}{2}\sum_{i,j=1}^{n} f\left(\mathbf{w}_i, \mathbf{w}_j\right) \\ &\quad - \sum_{\substack{i\in[n]\\ j\in[k]}} f\left(\mathbf{w}_i, \mathbf{v}_j\right) + \frac{1}{2}\sum_{i,j=1}^{k} f\left(\mathbf{v}_i, \mathbf{v}_j\right), \end{aligned} \tag{6}$$

where

$$\begin{aligned} f\left(\mathbf{w}, \mathbf{v}\right) &:= \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left[\mathbf{w}^\top\mathbf{x}\right]_+\left[\mathbf{v}^\top\mathbf{x}\right]_+\right] \\ &= \frac{1}{2\pi}\left|\left|\mathbf{w}\right|\right|\left|\left|\mathbf{v}\right|\right|\left(\sin\left(\theta_{\mathbf{w},\mathbf{v}}\right)\right. \\ &\quad \left. + \left(\pi - \theta_{\mathbf{w},\mathbf{v}}\right)\cos\left(\theta_{\mathbf{w},\mathbf{v}}\right)\right), \end{aligned} \tag{7}$$

and

$$\theta_{\mathbf{w},\mathbf{v}} := \cos^{-1}\left(\frac{\mathbf{w}^\top\mathbf{v}}{\left|\left|\mathbf{w}\right|\right|\cdot\left|\left|\mathbf{v}\right|\right|}\right)$$

is the angle between two vectors $\mathbf{w}, \mathbf{v}$. The latter equality in Eq. (7) was shown in Cho & Saul (2009, section 2).

Using the above representation, Brutzkus & Globerson (2017) compute the gradient of $f\left(\mathbf{w}, \mathbf{v}\right)$ with respect to $\mathbf{w}$, given by

$$\begin{aligned} g\left(\mathbf{w}, \mathbf{v}\right) &:= \frac{\partial}{\partial\mathbf{w}} f\left(\mathbf{w}, \mathbf{v}\right) \\ &= \frac{1}{2\pi}\left(\left|\left|\mathbf{v}\right|\right|\sin\left(\theta_{\mathbf{w},\mathbf{v}}\right)\bar{\mathbf{w}} + \left(\pi - \theta_{\mathbf{w},\mathbf{v}}\right)\mathbf{v}\right). \end{aligned} \tag{8}$$

Which implies that $\nabla F\left(\mathbf{w}_1^n\right)$, the gradient of the objective with respect to $\mathbf{w}_1^n$, equals

$$\begin{aligned} &\nabla F\left(\mathbf{w}_1^n\right) = \\ &\quad \frac{1}{2}\mathbf{w}_1^n + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\tilde{g}\left(\mathbf{w}_i, \mathbf{w}_j\right) - \sum_{\substack{i\in[n]\\ j\in[k]}}\tilde{g}\left(\mathbf{w}_i, \mathbf{v}_j\right), \end{aligned}$$

where $\tilde{g}\left(\mathbf{w}_i, \mathbf{u}\right) \in \mathbb{R}^{kn}$ equals $g\left(\mathbf{w}_i, \mathbf{u}\right) \in \mathbb{R}^k$ on entries $k(i-1)+1$ through ki, and zero elsewhere. We now provide the Hessian of Eq. (7) based on the computation of the gradient in Eq. (8) (see Subsection A.4.1 in the appendix for the full derivation)

$$\begin{aligned} h_1\left(\mathbf{w}, \mathbf{v}\right) &:= \frac{\partial^2}{\partial\mathbf{w}^2} f\left(\mathbf{w}, \mathbf{v}\right) \\ &= \frac{\sin\left(\theta_{\mathbf{w},\mathbf{v}}\right)\left|\left|\mathbf{v}\right|\right|}{2\pi\left|\left|\mathbf{w}\right|\right|}\left(\mathbf{I} - \bar{\mathbf{w}}\bar{\mathbf{w}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}^\top\right), \end{aligned}$$

$$\begin{aligned} h_2\left(\mathbf{w}, \mathbf{v}\right) &:= \frac{\partial^2}{\partial\mathbf{w}\partial\mathbf{v}} f\left(\mathbf{w}, \mathbf{v}\right) \\ &= \frac{1}{2\pi}\left(\left(\pi - \theta_{\mathbf{w},\mathbf{v}}\right)\mathbf{I} + \bar{\mathbf{n}}_{\mathbf{w},\mathbf{v}}\bar{\mathbf{v}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{w}}^\top\right), \end{aligned}$$

where

$$\mathbf{n}_{\mathbf{v},\mathbf{w}} = \bar{\mathbf{v}} - \cos\left(\theta_{\mathbf{v},\mathbf{w}}\right)\bar{\mathbf{w}} \tag{9}$$

and $\bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}} = \frac{\mathbf{n}_{\mathbf{v},\mathbf{w}}}{\left|\left|\mathbf{n}_{\mathbf{v},\mathbf{w}}\right|\right|}$. To formally define the Hessian of F (a $kn\times kn$ matrix), we partition it into $n\times n$ blocks, each of size $k\times k$. Define $\tilde{h}_1\left(\mathbf{w}_i, \mathbf{u}\right) \in \mathbb{R}^{kn\times kn}$ to equal $h_1\left(\mathbf{w}_i, \mathbf{u}\right)$ on the i-th $d\times d$ diagonal block and zero elsewhere. For $\mathbf{w}_i, \mathbf{w}_j$ define $\tilde{h}_2\left(\mathbf{w}_i, \mathbf{w}_j\right) \in \mathbb{R}^{kn\times kn}$ to equal $h_2\left(\mathbf{w}_i, \mathbf{w}_j\right)$ on the i, j-th $k\times k$ block and zero elsewhere. We now have that the Hessian is given by

$$\begin{aligned} \nabla^2 F\left(\mathbf{w}_1^n\right) &= \frac{1}{2}\mathbf{I} + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\tilde{h}_1\left(\mathbf{w}_i, \mathbf{w}_j\right) \\ &\quad - \sum_{\substack{i\in[n]\\ j\in[k]}}\tilde{h}_1\left(\mathbf{w}_i, \mathbf{v}_j\right) + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\tilde{h}_2\left(\mathbf{w}_i, \mathbf{w}_j\right). \end{aligned} \tag{10}$$

4.3. Proof of Corollary 1

To show the first part of Corollary 1, we will use the following lemma:

Lemma 3. *Let $\mathbf{w}_1^n = \left(\mathbf{w}_1, \dots, \mathbf{w}_n\right)$, $V = \left(\mathbf{v}_1, \dots, \mathbf{v}_k\right)$ where $\mathbf{w}_i, \mathbf{v}_j \in \mathbb{R}^k$ for all $i \in [n], j \in [k]$. Denote for any natural $m \geq 0$, $\tilde{\mathbf{w}}_{1,m}^n = \left(\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_n\right)$, $\tilde{\mathbf{w}}_i = \left(\mathbf{w}_i, \mathbf{0}\right) \in \mathbb{R}^{k+m}$, $\tilde{V}_m = \left(\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_k\right)$, $\tilde{\mathbf{v}}_i = \left(\mathbf{v}_i, \mathbf{0}\right) \in \mathbb{R}^{k+m}$ and let $M \in \mathbb{R}^{n\times n}$ be the matrix with entries $M_{ij} =$*

$$\begin{cases} \frac{1}{2} + \sum\limits_{\substack{l=1\\ l\neq i}}^{n}\frac{\sin\left(\theta_{\mathbf{w}_i,\mathbf{w}_l}\right)\left|\left|\mathbf{w}_l\right|\right|}{2\pi\left|\left|\mathbf{w}_i\right|\right|} - \sum\limits_{l=1}^{k}\frac{\sin\left(\theta_{\mathbf{w}_i,\mathbf{v}_l}\right)\left|\left|\mathbf{v}_l\right|\right|}{2\pi\left|\left|\mathbf{w}_i\right|\right|}, & i = j \\ \frac{1}{2\pi}\left(\pi - \theta_{\mathbf{w}_i,\mathbf{w}_j}\right), & i \neq j \end{cases}.$$

Then the spectrum of $\nabla^2 F\left(\tilde{\mathbf{w}}_{1,m}^n\right)$ is comprised of the spectrum of $\nabla^2 F\left(\mathbf{w}_1^n\right)$ and the spectrum of M with multiplicity m. In particular, if $\nabla^2 F\left(\tilde{\mathbf{w}}_{1,1}^n\right) \succeq \lambda_{\min}\cdot\mathbf{I}$ then $\nabla^2 F\left(\tilde{\mathbf{w}}_{1,m}^n\right) \succeq \lambda_{\min}\cdot\mathbf{I}$, for any $m > 1$.

Proof. A straightforward substitution of $\tilde{\mathbf{w}}_{1,m}^n$ and $\tilde{V}_m$ in Eq. (10), and a permutation of the rows and columns of the resulting matrix reveals that

$$\nabla^2 F\left(\tilde{\mathbf{w}}_{1,m}^n\right) = \begin{bmatrix} \nabla^2 F\left(\mathbf{w}_1^n\right) & 0 & 0 & \cdots & 0 \\ 0 & M & 0 & \cdots & 0 \\ 0 & 0 & M & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & M \end{bmatrix}.$$

Now, diagonalizing the block diagonal $\nabla^2 F\left(\tilde{\mathbf{w}}_{1,m}^n\right)$ completes the proof of the lemma. □

Back to the first part of Corollary 1, we have from Lemma 3 that the lower bound on the smallest eigenvalue of $\nabla^2 F\left(\tilde{\mathbf{w}}_{1,1}^n\right)$ holds for $\nabla^2 F\left(\tilde{\mathbf{w}}_{1,m}^n\right)$ for any $m \geq 1$. Furthermore, since $\|\mathbf{w}_1^n\|_2 = \|\tilde{\mathbf{w}}_{1,m}^n\|_2$ for any $m \geq 0$ we have that the upper bound on the third order derivatives from Subsection A.2 in the appendix and the Lipschitz bound on the objective from Subsection A.3 in the appendix still hold, as well as the bound on the norm of the gradient. Therefore by running the simulations in Sec. 3 on $\tilde{\mathbf{w}}_{1,1}^n$ instead of $\mathbf{w}_1^n$, the results apply in any optimization space $\mathbb{R}^{n(k+m)}$, for natural $m \geq 0$, since the conditions for invoking Lemma 1 and Lemma 2 are met with the same exact constants[4], completing the first part of the corollary.

For the second part of the corollary, we note that if $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are chosen i.i.d. from $\mathcal{N}(\mathbf{0}, c\mathbf{I})$, then by standard concentration arguments, for any $\epsilon > 0$ and high enough dimension d (depending on k, ϵ), it holds with probability at least $1-\exp(-\Omega(d))$ that $|\frac{1}{\sqrt{cd}}||\mathbf{v}_i||-1| \leq \epsilon$ and $|\frac{1}{cd}\mathbf{v}_i^\top \mathbf{v}_{i'}| \leq \epsilon$ for all $i, i' \in \{1, \ldots, k\}$ (see Ledoux (2005)). Therefore, regardless of which distribution we are considering, with probability at least $1 - \exp(-\Omega(d))$, we can find a scalar $a > 0$ and an orthogonal matrix M, such that $||aM\mathbf{v}_i - \mathbf{e}_i|| \leq \epsilon$ for all i, where $\mathbf{e}_i$ is the i-th standard basis vector. Note that this strongly uses the orthonormal reparameterization invariance of gradient descent.

Letting F be our objective function (w.r.t. the randomly chosen $\mathbf{v}_1, \ldots, \mathbf{v}_k$), and using the rotational symmetry of the Gaussian distribution and the positive-homogeneity of the ReLU function, we have that $F(\mathbf{w}_1^n)$ equals

$$\frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[\left(\sum_{i=1}^n [\mathbf{w}_i^\top \mathbf{x}]_+ - \sum_{i=1}^k [\mathbf{v}_i^\top \mathbf{x}]_+\right)^2\right]$$

$$= \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[\left(\sum_{i=1}^n [\mathbf{w}_i^\top (M^\top \mathbf{x})]_+ - \sum_{i=1}^k [\mathbf{v}_i^\top (M^\top \mathbf{x})]_+\right)^2\right]$$

$$= \frac{1}{2a^2}\mathbb{E}_{\mathbf{x}}\left[\left(\sum_{i=1}^n [(aM\mathbf{w}_i)^\top \mathbf{x}]_+ - \sum_{i=1}^k [(aM\mathbf{v}_i)^\top \mathbf{x}]_+\right)^2\right],$$

where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. It follows that F has the same local minima as

$$\tilde{F}(\mathbf{w}_1^n) := \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\sum_{i=1}^n [\mathbf{w}_i^\top \mathbf{x}]_+ - \sum_{i=1}^k [(aM\mathbf{v}_i)^\top \mathbf{x}]_+\right)^2\right],$$

since they are equivalent after scaling and rotation, as $\tilde{F}(\mathbf{w}_1^n) = a^2 F(\frac{1}{a}M^\top \mathbf{w}_1^n)$. Thus, it is enough to prove existence of local minima for $\tilde{F}$.

By the argument above, we can rewrite $\tilde{F}(\mathbf{w}_1^n)$ as

$$\tilde{F}(\mathbf{w}_1^n) := \tilde{F}_{\tilde{\mathbf{e}}_1,\ldots,\tilde{\mathbf{e}}_k}(\mathbf{w}_1^n) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\sum_{i=1}^n [\mathbf{w}_i^\top \mathbf{x}]_+ - \sum_{i=1}^k [\tilde{\mathbf{e}}_i^\top \mathbf{x}]_+\right)^2\right],$$

where (with high probability) each $\tilde{\mathbf{e}}_i$ is ϵ-close to the standard basis vector $\mathbf{e}_i$. If $\mathbf{e}_i = \tilde{\mathbf{e}}_i$, we have already shown that there is some local minimum $\mathbf{w}_1^{*n}$, which is in the interior of a sphere S such that $F(\mathbf{w}_1^{*n}) < \min_{\mathbf{w}_1^n \in S} F(\mathbf{w}_1^n)$, and moreover, the ball B enclosed by S does not contain global minima (see Eq. (4)) since Thm. 4 in the appendix and the condition in Eq. (5) imply that the minimal value in the ball enclosing $\mathbf{w}_1^n$ is strictly positive. In particular, let $\epsilon_0 > 0$ be such that

$$F(\mathbf{w}_1^{*n}) < \min_{\mathbf{w}_1^n \in S} F(\mathbf{w}_1^n) - \epsilon_0$$

and

$$\min_{\mathbf{w}_1^n \in B} F(\mathbf{w}_1^n) > \inf_{\mathbf{w}_1^n} F(\mathbf{w}_1^n) + \epsilon_0.$$

It is easily verified that by setting ϵ small enough (depending only on $\mathbf{w}_1^{*n}, B, \epsilon_0$ which are all fixed), we can ensure that

$$\max_{\mathbf{w}_1^n \in B} |\tilde{F}_{\tilde{\mathbf{e}}_1,\ldots,\tilde{\mathbf{e}}_k}(\mathbf{w}_1^n) - \tilde{F}_{\mathbf{e}_1,\ldots,\mathbf{e}_k}(\mathbf{w}_1^n)| \leq \frac{\epsilon_0}{3},$$

therefore $\tilde{F}_{\tilde{\mathbf{e}}_1,\ldots,\tilde{\mathbf{e}}_k}(\mathbf{w}_1^{*n}) < \min_{\mathbf{w}_1^n \in S} \tilde{F}_{\tilde{\mathbf{e}}_1,\ldots,\tilde{\mathbf{e}}_k}(\mathbf{w}_1^n)$, as well as $\min_{\mathbf{w}_1^n \in B} \tilde{F}_{\tilde{\mathbf{e}}_1,\ldots,\tilde{\mathbf{e}}_k}(\mathbf{w}_1^n) > \inf_{\mathbf{w}_1^n} \tilde{F}_{\tilde{\mathbf{e}}_1,\ldots,\tilde{\mathbf{e}}_k}(\mathbf{w}_1^n)$, which implies that any minimizer of $\tilde{F}_{\tilde{\mathbf{e}}_1,\ldots,\tilde{\mathbf{e}}_k}$ over B must be a local (non-global) minimum.

[4] Note that for $m = 0$ the eigenvalue lower bound constant may change, since the spectrum of M has no impact on the spectrum of $\nabla^2 F\left(\tilde{\mathbf{w}}_{1,0}^n\right)$. This, however, can only result in a stronger lower bound and does not affect the validity on the results obtained when running the experiments in Sec. 3 with $m = 1$.

References

Auer, P., Herbster, M., and Warmuth, M. K. Exponentially many local minima for single neurons. In *NIPS*, 1996.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.

Boob, D. and Lan, G. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

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

Cho, Y. and Saul, L. K. Kernel methods for deep learning. In *Advances in neural information processing systems*, pp. 342–350, 2009.

Du, S. S., Lee, J. D., Tian, Y., Poczos, B., and Singh, A. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint arXiv:1712.00779*, 2017.

Feizi, S., Javadi, H., Zhang, J., and Tse, D. Porcupine neural networks:(almost) all local optima are global. *arXiv preprint arXiv:1710.02196*, 2017.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

Ge, R., Lee, J. D., and Ma, T. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Haeffele, B. D. and Vidal, R. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

Janzamin, M., Sedghi, H., and Anandkumar, A. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *CoRR abs/1506.08473*, 2015.

Ledoux, M. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2005.

Li, Y. and Yuan, Y. Convergence analysis of two-layer neural networks with relu activation. *arXiv preprint arXiv:1705.09886*, 2017.

Livni, R., Shalev-Shwartz, S., and Shamir, O. On the computational efficiency of training neural networks. In *NIPS*, pp. 855–863, 2014.

Nguyen, Q. and Hein, M. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

Poston, T., Lee, C.-N., Choie, Y., and Kwon, Y. Local minima and back propagation. In *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on*, volume 2, pp. 173–176. IEEE, 1991.

Safran, I. and Shamir, O. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pp. 774–782, 2016.

Shamir, O. Distribution-specific hardness of learning neural networks. *arXiv preprint arXiv:1609.01037*, 2016.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of overparameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

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

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Swirszcz, G., Czarnecki, W. M., and Pascanu, R. Local minima in training of deep networks. *arXiv preprint arXiv:1611.06310*, 2016.

Tian, Y. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, Q., Panigrahy, R., Sachdeva, S., and Rahimi, A. Electron-proton dynamics in deep learning. *arXiv preprint arXiv:1702.00458*, 2017.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.