Convergence Results for Neural Networks via Electrodynamics*

Rina Panigrahy[1], Ali Rahimi[2], Sushant Sachdeva[†3], and Qiuyi Zhang[4]

1 **Google Inc., Mountain View, CA, USA**
rinap@google.com
2 **Google Inc., Mountain View, CA, USA**
arahimi@google.com
3 **University of Toronto, Toronto, Canada**
sachdeva@cs.toronto.edu
4 **University of California Berkeley, Berkeley, CA, USA**
10zhangqiuyi@berkeley.edu

Abstract

We study whether a depth two neural network can learn another depth two network using gradient descent. Assuming a linear output node, we show that the question of whether gradient descent converges to the target function is equivalent to the following question in electrodynamics: Given k fixed protons in $\mathbb{R}^d$, and k electrons, each moving due to the attractive force from the protons and repulsive force from the remaining electrons, whether at equilibrium all the electrons will be matched up with the protons, up to a permutation. Under the standard electrical force, this follows from the classic Earnshaw's theorem. In our setting, the force is determined by the activation function and the input distribution. Building on this equivalence, we prove the existence of an activation function such that gradient descent learns at least one of the hidden nodes in the target network. Iterating, we show that gradient descent can be used to learn the entire network one node at a time.

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases Deep Learning, Learning Theory, Non-convex Optimization

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.22

1 Introduction

Deep learning has resulted in major strides in machine learning applications including speech recognition, image classification, and ad-matching. The simple idea of using multiple layers of nodes with a non-linear activation function at each node allows one to express any function. To learn a certain target function we just use (stochastic) gradient descent to minimize the loss; this approach has resulted in significantly lower error rates for several real world functions, such as those in the above applications. Naturally the question remains: how close are we to the optimal values of the network weight parameters? Are we stuck in some bad local minima? While there are several recent works [8, 11, 17] that have tried to study the presence of local minima, the picture is far from clear.

* The full version of this paper is available at [25], https://arxiv.org/abs/1702.00458
† Part of this work was done when this author was a research scientist at Google Inc., Mountain View, CA, USA

9th Innovations in Theoretical Computer Science Conference (ITCS 2018).
Editor: Anna R. Karlin; Article No. 22; pp. 22:1–22:19

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

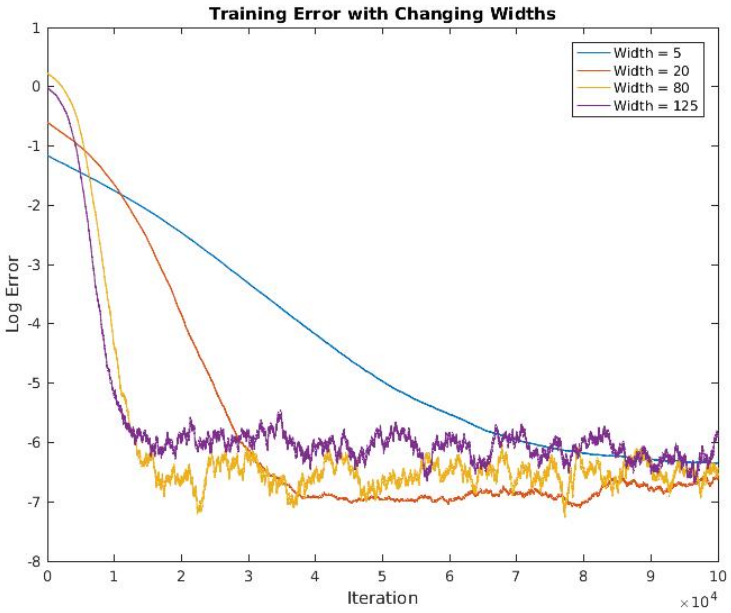

Figure 1 Test Error of Depth 2 Networks of Varying Width.

There has been some work on studying how well can neural networks learn some synthetic function classes (e.g. polynomials [1], decision trees). In this work we study how well can neural networks learn neural networks with gradient descent? Our focus here, via the framework of proper learning, is to understand if a neural network can learn a function from the same class (and hence achieve vanishing error).

Specifically, if the target function is a neural network with randomly initialized weights, and we attempt to learn it using a network with the same architecture, then, will gradient descent converge to the target function?

Experimental simulations (see Figure 1 and Section 5 for further details) show that for depth 2 networks of different widths, with random network weights, stochastic gradient descent of a hypothesis network with the same architecture converges to a squared ℓ_2 error that is a small percentage of a random network, indicating that SGD can learn these shallow networks with random weights. Because our activations are sigmoidal from -1 to 1, the training error starts from a value of about 1 (random guessing) and diminishes quickly to under 0.002. This seems to hold even when the width, the number of hidden nodes, is substantially increased (even up to 125 nodes), but depth is held constant at 2.

In this paper, we attempt to understand this phenomenon theoretically. We prove that, under some assumptions, depth-2 neural networks can learn functions from the same class with vanishingly small error using gradient descent.

1.1 Results and Contributions

We theoretically investigate the question of convergence for networks of depth two. Our main conceptual contribution is that for depth 2 networks where the top node is a sum node, the question of whether gradient descent converges to the desired target function is equivalent to the following question in electrodynamics: Given k fixed protons in $\mathbb{R}^d$, and k moving electrons, with all the electrons moving under the influence of the electrical force of

Table 1 Activation, Potentials, and Convergence Results Summary

Name of Activation	Potential ($\Phi(\theta, w)$)	Convergence?
Almost λ-harmonic	Complicated (see Lem 13)	Yes, Thm 17
Sign	$1 - \frac{2}{\pi}\cos^{-1}(\theta^T w)$	Yes for d = 2, Lem 29
Polynomial	$(\theta^T w)^m$	Yes, for orthonormal w_i. Lem 30

attraction from the protons and repulsion from the remaining electrons, at equilibrium, are all the electrons matched up with all the fixed protons, up to a permutation?

In the above, k is the number of hidden units, d is the number of inputs, the positions of each fixed charge is the input weight vector of a hidden unit in the target network, and the initial positions of the moving charges are the initial values of the weight vectors for the hidden units in the learning network. The motion of the charges essentially tracks the change in the network during gradient descent. The force between a pair of charges is not given by the standard electrical force of $1/r^2$ (where r is the distance between the charges), but by a function determined by the activation and the input distribution. Thus the question of convergence in these simplified depth two networks can be resolved by studying the equivalent electrodynamics question with the corresponding force function.

▶ **Theorem 1** (informal statement of Theorem 5). *Applying gradient descent for learning the output of a depth two network with k hidden units with activation σ, and a linear output node, under squared loss, using a network of the same architecture, is equivalent to the motion of k charges in the presence of k fixed charges where the force between each pair of charges is given by a potential function that depends on σ and the input distribution.*

Based on this correspondence we prove the existence of an activation function such that the corresponding gradient descent dynamics under standard Gaussian inputs result in learning at least one of the hidden nodes in the target network. We then show that this allows us to learn the complete target network one node at a time. For more realistic activation functions, we only obtain partial results. We assume the sample complexity is close to its infinite limit.

▶ **Theorem 2** (informal statement of Theorem 12). *There is an activation function such that running gradient descent for minimizing the squared loss along with ℓ_2 regularization for standard Gaussian inputs, at convergence, we learn at least one of the hidden weights of the target neural network.*

We prove that the above result can be iterated to learn the entire network node-by-node using gradient descent (Theorem 17). Our algorithm learns a network with the same architecture and number of hidden nodes as the target network, in contrast with several existing improper learning results.

In the appendix, we show some weak results for more practical activations. For the sign activation, we show that for the loss with respect to a single node, the only local minima are at the hidden target nodes with high probability if the target network has a randomly picked top layer. For the polynomial activation, we derive a similar result under the assumption that the hidden nodes are orthonormal.

1.2 Intuition and Techniques

Note that for the standard electric potential function given by $\Phi = 1/r$ where r is the distance between the charges, it is known from Earnshaw's theorem that an electrodynamic system with some fixed protons and some moving electrons is at equilibrium only when the moving electrons coincide with the fixed protons. Given our translation above between electrodynamic systems and depth 2 networks (Section 2), this would imply learnability of depth 2 networks under gradient descent under ℓ_2 loss, if the activation function corresponds to the electrostatic potential. However, there exists no activation function σ corresponding to this Φ.

The proof of Earnshaw's theorem is based on the fact that the electrostatic potential is harmonic, *i.e*, its Laplacian (trace of its Hessian) is identically zero. This ensures that at every critical point, there is direction of potential reduction (unless the hessian is identically zero). We generalize these ideas to potential functions that are eigenfunctions of the Laplacians, λ-harmonic potentials (Section 3). However, these potentials are unbounded. Subsequently, we construct a non-explicit activation function such that the corresponding potential is bounded and is almost λ-harmonic, *i.e.*, it is λ-harmonic outside a small sphere (Section 4). For this activation function, we show at a stable critical point, we must learn at least one of the hidden nodes. Gradient descent (possibly with some noise, as in the work of Ge *et al.* [12]) is believed to converge to stable critical points. However, for simplicity, we descend along directions of negative curvature to escape saddle points. Our activation lacks some regularity conditions required in [12]. We believe the results in [16] can be adapted to our setting to prove that perturbed gradient descent converges to stable critical points.

There is still a large gap between theory and practice. However, we believe our work can offer some theoretical explanations and guidelines for the design of better activation functions for gradient-based training algorithms. For example, better accuracy and training speed were reported when using the newly discovered exponential linear unit (ELU) activation function in [9, 21]. We hope for more theory-backed answers to these and many other questions in deep learning.

1.3 Related Work

If the activation functions are linear or if some independence assumptions are made, Kawaguchi shows that the only local minima are the global minima [17]. Under the spin-glass and other physical models, some have shown that the loss landscape admits well-behaving local minima that occur usually when the overall error is small [8, 11]. When only training error is considered, some have shown that a global minima can be achieved if the neural network contains sufficiently many hidden nodes [23]. Recently, Daniely has shown that SGD learns the conjugate kernel class [10]. Under simplifying assumptions, some results for learning ReLU's with gradient descent are given in [24, 7]. Our research is inspired by [1], where the authors show that for polynomial target functions, gradient descent on neural networks with one hidden layer converges to low error, given a large number of hidden nodes, and under complex perturbations, there are no robust local minima. Even more recently, similar results about the convergence of SGD for two-layer neural networks have been established for a polynomial activation function under a more complex loss function [13]. And in [19], they study the same problem as ours with the RELU activation and where lower layer of the network is close to identity and the upper layer has weights all one. This corresponds to the case where each electron is close to a distinct proton – under these assumptions they show that SGD learns the true network.

Under worst case assumptions, there has been hardness results for even simple networks. A neural network with one hidden unit and sigmoidal activation can admit exponentially many local minima [4]. Backprogration has been proven to fail in a simple network due to the abundance of bad local minima [6]. Training a 3-node neural network with one hidden layer is NP-complete [5]. But, these and many similar worst-case hardness results are based on worst case training data assumptions. However, by using a result in [18] that learning a neural network with threshold activation functions is equivalent to learning intersection of halfspaces, several authors showed that under certain cryptographic assumptions, depth-two neural networks are not efficiently learnable with smooth activation functions [20, 27, 26].

Due to the difficulty of analysis of the non convex gradient descent in deep learning, many have turned to improper learning and the study of non-gradient methods to train neural networks. Janzamin et. al use tensor decomposition methods to learn the shallow neural network weights, provided access to the score function of the training data distribution [15]. Eigenvector and tensor methods are also used to train shallow neural networks with quadratic activation functions in [20]. Combinatorial methods that exploit layerwise correlations in sparse networks have also been analyzed provably in [3]. Kernel methods, ridge regression, and even boosting were explored for regularized neural networks with smooth activation functions in [22, 27, 26]. Non-smooth activation functions, such as the ReLU, can be approximated by polynomials and are also amenable to kernel methods[14]. These methods however are very different from the simple popular SGD.

2 Deep Learning, Potentials, and Electron-Proton Dynamics

2.1 Preliminaries

We will work in the space $\mathcal{M} = \mathbb{R}^d$. We denote the gradient and Hessian as $\nabla_{\mathbb{R}^d} f$ and $\nabla^2_{\mathbb{R}^d} f$ respectively. The Laplacian is defined as $\Delta_{\mathbb{R}^d} f = \mathrm{Tr}(\nabla^2_{\mathbb{R}^d} f)$. If f is multivariate with variable x_i, then let f_{x_i} be a restriction of f onto the variable x_i with all other variables fixed. Let $\nabla_{x_i} f, \Delta_{x_i} f$ to be the gradient and Laplacian, respectively, of f_{x_i} with respect to x_i. Lastly, we say x is a critical point of f if ∇f does not exist or $\nabla f = 0$.

We focus on learning depth two networks with a linear activation on the output node. If the network takes inputs $x \in \mathbb{R}^d$ (say from some distribution $\mathcal{D}$), then the network output, denoted $f(x)$ is a sum over $k = \mathrm{poly}(d)$ hidden units with weight vectors $w_i \in \mathbb{R}^d$, activation $\sigma(x, w) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and output weights $b_i \in \mathbb{R}$. Thus, we can write $f(x) = \sum_{i=1}^k b_i \sigma(x, w_i)$. We denote this concept class $\mathcal{C}_{\sigma,k}$. Our hypothesis concept class is also $\mathcal{C}_{\sigma,k}$.

Let $\boldsymbol{a} = (a_1, ..., a_k)$ and $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$; similarly for $\boldsymbol{b}, \boldsymbol{w}$ and our guess is $\hat{f}(x) = \sum_{i=1}^k a_i \sigma(x, \theta_i)$. We define Φ, the **potential function** *corresponding to the activation* σ, as

$$\Phi(\theta, w) = \mathop{\mathbb{E}}_{X \sim D}[\sigma(X, \theta)\sigma(X, w)].$$

We work directly with the true squared loss error $L(a, \theta) = \mathbb{E}_{x \sim \mathcal{D}}[(f - \hat{f})^2]$. To simplify L, we re-parametrize a by $-a$ and expand.

$$\begin{aligned} L(\boldsymbol{a}, \boldsymbol{\theta}) &= \mathop{\mathbb{E}}_{X \sim D}\left[\left(\sum_{i=1}^k a_i \sigma(X, \theta_i) + \sum_{i=1}^k b_i \sigma(X, w_i)\right)^2\right] \\ &= \sum_{i=1}^k \sum_{j=1}^k a_i a_j \Phi(\theta_i, \theta_j) + 2 a_i b_j \Phi(\theta_i, w_j) + b_i b_j \Phi(w_i, w_j), \end{aligned} \tag{1}$$

Given $\mathcal{D}$, the activation function σ, and the loss L, we attempt to show that we can use some variant of gradient descent to learn, with high probability, an ϵ-approximation of w_j for some (or all) j. Note that our loss is jointly convex, though it is quadratic in $\boldsymbol{a}$.

In this paper, we restrict our attention to translationally invariant activations and potentials. Specifically, we may write $\Phi = h(\theta - w)$ for some function $h(x)$. Furthermore, a translationally invariant function $\Phi(r)$ is *radial* if it is a function of $r = \|x - y\|$.

▶ Remark. Translationally symmetric potentials satisfy $\Phi(\theta, \theta)$ *is a positive constant.* We normalize $\Phi(\theta, \theta) = 1$ for the rest of the paper.

We assume that our input distribution $\mathcal{D} = \mathcal{N}(0, \mathbf{I_{d\times d}})$ is fixed as the standard Gaussian in $\mathbb{R}^d$. This assumption is not critical and a simpler distribution might lead to better bounds. However, for arbitrary distributions, there are hardness results for PAC-learning halfspaces [18].

We call a potential function **realizable** if it corresponds to some activation σ. The following theorem characterizes realizable translationally invariant potentials under standard Gaussian inputs. Proofs and a similar characterization for rotationally invariant potentials can be found in Appendix B.

▶ **Theorem 3.** *Let $\mathcal{M} = \mathbb{R}^d$ and Φ is square-integrable and $\mathfrak{F}(\Phi)$ is integrable. Then, Φ is realizable under standard Gaussian inputs if $\mathfrak{F}(\Phi)(\omega) \geq 0$ and the corresponding activation is $\sigma(x) = (2\pi)^{d/4} e^{x^T x/4} \mathfrak{F}^{-1}(\sqrt{\mathfrak{F}(\Phi)})(x)$, where $\mathfrak{F}$ is the generalized Fourier transform in $\mathbb{R}^d$.*

2.2 Electron-Proton Dynamics

By interpreting the pairwise potentials as electrostatic attraction potentials, we notice that our dynamics is similar to electron-proton type dynamics under potential Φ, where w_i are fixed point charges in $\mathbb{R}^d$ and θ_i are moving point charges in $\mathbb{R}^d$ that are trying to find w_i. The total force on each charge is the sum of the pairwise forces, determined by the gradient of Φ. We note that standard dynamics interprets the force between particles as an acceleration vector. In gradient descent, it is interpreted as a velocity vector.

▶ **Definition 4.** Given a potential Φ and particle locations $\theta_1, ..., \theta_k \in \mathbb{R}^d$ along with their respective charges $a_1, ..., a_k \in \mathbb{R}$. We define **Electron-Proton Dynamics** under Φ with some subset $S \subseteq [k]$ of fixed particles to be the solution $(\theta_1(t), ..., \theta_k(t))$ to the following system of differential equations: For each pair (θ_i, θ_j), there is a force from θ_j exerted on θ_i that is given by $\mathbf{F}_i(\theta_j) = a_i a_j \nabla_{\theta_i} \Phi(\theta_i, \theta_j)$ and

$$-\frac{d\theta_i}{dt} = \sum_{j \neq i} \mathbf{F}_i(\theta_j)$$

for all $i \notin S$, with $\theta_i(0) = \theta_i$. For $i \in S$, $\theta_i(t) = \theta_i$.

For the following theorem, we assume that $\boldsymbol{\theta}$ is fixed.

▶ **Theorem 5.** *Let Φ be a symmetric potential and L be as in (1). Running continuous gradient descent on $\frac{1}{2}L$ with respect to θ, initialized at $(\theta_1, ..., \theta_k)$ produces the same dynamics as Electron-Proton Dynamics under 2Φ with fixed particles at $w_1, ..., w_k$ with respective charges $b_1, .., b_k$ and moving particles at $\theta_1, ..., \theta_k$ with respective charges $a_1, ..., a_k$.*

3 Earnshaw's Theorem and Harmonic Potentials

When running gradient descent on a non-convex loss, we often can and do get stuck at a local minima. In this section, we use second-order information to deduce that for certain

classes of potentials, there are no spurious local minima. The potentials In this section are often *unbounded and un-realizable*. However, in the next section, we apply insights developed here to derive similar convergence results for approximations of these potentials.

Earnshaw's theorem in electrodynamics shows that there is no stable local minima for electron-proton dynamics. This hinges on the property that the electric potential $\Phi(\theta, w) = \|\theta - w\|^{2-d}, d \neq 2$ is harmonic, with $d = 3$ in natural setting. If $d = 2$, we instead have $\Phi(\theta, w) = -\ln(\|\theta - w\|)$. First, we notice that this is a symmetric loss, and our usual loss in (1) has constant terms that can be dropped to further simplify.

$$\overline{L}(a, \theta) = 2\sum_{i=1}^{k}\sum_{i<j} a_i a_j \Phi(\theta_i, \theta_j) + 2\sum_{i=1}^{k}\sum_{j=1}^{k} a_i b_j \Phi(\theta_i, w_j) \tag{2}$$

▶ **Definition 6.** $\Phi(\theta, w)$ is a **harmonic** potential on Ω if $\Delta_\theta \Phi(\theta, w) = 0$ for all $\theta \in \Omega$, except possibly at $\theta = w$.

▶ **Definition 7.** Let $\Omega \subseteq \mathbb{R}^d$ and consider a function $f : \Omega \to \mathbb{R}$. A critical point $x^* \in \Omega$ is a **local minimum** if there exists $\epsilon > 0$ such that $f(x^* + v) \geq f(x^*)$ for all $\|v\| \leq \epsilon$. It is a **strict local minimum** if the inequality is strict for all $\|v\| \leq \epsilon$.

▶ **Fact 8.** *Let x^* be a critical point of a function $f : \Omega \to \mathbb{R}$ such that f is twice differentiable at x^*. Then, if x^* is a local minimum then $\lambda_{min}(\nabla^2 f(x^*)) \geq 0$. Moreover, if $\lambda_{min}(\nabla^2 f(x^*)) > 0$, then x^* is a strict local minimum.*

Note that if $\lambda_{min}(\nabla^2 f(x^*)) < 0$ then moving along the direction of the corresponding eigenvector decreases f locally. If Φ is harmonic then it can be shown the trace of its Hessian is 0 so if there is any non zero eigenvalue then at least one eigenvalue is negative. This idea results in the following known theorem (see full proof in supplementary material) that is applicable to the electric potential function $1/r$ in 3-dimensions since is harmonic. It implies that a configuration of n electrons and n protons cannot be in a strict local minimum even if one of the mobile charges is isolated (however note that this potential function goes to ∞ at $r = 0$ and may not be realizable).

▶ **Theorem 9.** (Earnshaw's Theorem. See [2]) *Let $\mathcal{M} = \mathbb{R}^d$ and let Φ be harmonic and L be as in* (2)*. Then, L admits no differentiable strict local minima.*

Note that the Hessian of a harmonic potential can be identically zero. To avoid this possibility we generalize harmonic potentials.

3.1 λ-Harmonic Potentials

In order to relate our loss function with its Laplacian, we consider potentials that are non-negative eigenfunctions of the Laplacian operator. Since the zero eigenvalue case simply gives rise to harmonic potentials, we restrict our attention to positive eigenfunctions.

▶ **Definition 10.** A potential Φ is **λ-harmonic** on Ω if there exists $\lambda > 0$ such that for every $\theta \in \Omega$, $\Delta_\theta \Phi(\theta, w) = \lambda \Phi(\theta, w)$, except possibly at $\theta = w$.

Note that there are realizable versions of these potentials; for example $\Phi(a, b) = e^{-\|a-b\|_1}$ in $\mathbb{R}^1$. In the next section, we construct realizable potentials that are λ-harmonic almost everywhere except when θ and w are very close.

▶ **Theorem 11.** *Let Φ be λ-harmonic and L be as in* (1)*. Then, L admits no local minima $(\boldsymbol{a}, \boldsymbol{\theta})$, except when $L(\boldsymbol{a}, \boldsymbol{\theta}) = L(0, \boldsymbol{\theta})$ or $\theta_i = w_j$ for some i, j.*

Proof. Let $(\boldsymbol{a}, \boldsymbol{\theta})$ be a critical point of L. On the contrary, we assume that $\theta_i \neq w_j$ for all i, j. WLOG, we can partition $[k]$ into $S_1, ..., S_r$ such that for all $u \in S_i, v \in S_j$, we have $\theta_u = \theta_v$ iff $i = j$. Let $S_1 = \{\theta_1, \dots, \theta_l\}$. We consider changing all $\theta_1, \dots, \theta_l$ by the same v and define $H(\mathbf{a}, v) = L(\mathbf{a}, \theta_1 + v, ..., \theta_l + v, \theta_{l+1} \dots, \theta_k)$.

The optimality conditions on $\mathbf{a}$ are $0 = \frac{\partial L}{\partial a_i} = 2\sum_j a_j \Phi(\theta_i, \theta_j) + 2\sum_{j=1}^k b_j \Phi(\theta_i, w_j)$. Thus, by the definition of λ-harmonic potentials, we may differentiate as $\theta_i \neq w_j$ and compute the Laplacian as

$$\begin{aligned}\Delta_v H &= \lambda \sum_{i=1}^{l} a_i \left(2\sum_{j=1}^{k} b_j \Phi(\theta_i, w_j) + 2\sum_{j=l+1}^{k} a_j \Phi(\theta_i, \theta_j) \right) \\ &= \lambda \sum_{i=1}^{l} a_i \left(-2\sum_{j=1}^{l} a_j \Phi(\theta_i, \theta_j) \right) = -2\lambda \sum_{i=1}^{l} a_i \left(\sum_{j=1}^{l} a_j \right) = -2\lambda \left(\sum_{i=1}^{l} a_i \right)^2\end{aligned}$$

If $\sum_{i=1}^l a_i \neq 0$, then we conclude that the Laplacian is strictly negative, so we are not at a local minimum. Similarly, we can conclude that for each S_i, $\sum_{u \in S_i} a_u = 0$. In this case, since $\sum_{i=1}^k a_i \sigma(\theta_i, x) = 0$, $L(\boldsymbol{a}, \boldsymbol{\theta}) = L(0, \boldsymbol{\theta})$. ◀

4 Realizable Potentials with Convergence Guarantees

In this section, we derive convergence guarantees for realizable potentials that are almost λ-harmonic, specifically, they are λ-harmonic outside of a small neighborhood around the origin. First, we prove the existence of activation functions such that the corresponding potentials are almost λ-harmonic. Then, we reason about the Laplacian of our loss, as in the previous section, to derive our guarantees. We show that at a stable minima, each of the θ_i is close to some w_j in the target network. We may end up with a many to one mapping of the learned hidden weights to the true hidden weights, instead of a bijection. To make sure that $\|a\|$ remains controlled throughout the optimization process, we add a quadratic regularization term to L and instead optimize $G = L + \|a\|^2$.

Our optimization procedure is a slightly altered version of gradient descent, where we incorporate a second-order method (which we call Hessian descent as in Algorithm 1) that is used when the gradient is small and progress is slow. The descent algorithm (Algorithm 2) allows us to converge to points with small gradient and small negative curvature. Namely, for smooth functions, in poly$(1/\epsilon)$ iterations, we reach a point in $\mathcal{M}_{G,\epsilon}$, where

$$\mathcal{M}_{G,\epsilon} = \left\{ x \in \mathcal{M} \middle| \|\nabla G(x)\| \leq \epsilon \text{ and } \lambda_{min}(\nabla^2 G(x)) \geq -\epsilon \right\}$$

We show that if $(\boldsymbol{a}, \boldsymbol{\theta})$ is in $\mathcal{M}_{G,\epsilon}$ for ϵ small, then θ_i is close to w_j for some j. Finally, we show how to initialize $(\boldsymbol{a^{(0)}}, \boldsymbol{\theta^{(0)}})$ and run second-order GD to converge to $\mathcal{M}_{G,\epsilon}$, proving our main theorem.

▶ **Theorem 12.** *Let $\mathcal{M} = \mathbb{R}^d$ for $d \equiv 3 \mod 4$ and $k = \text{poly}(d)$. For all $\epsilon \in (0, 1)$, we can construct an activation σ_ϵ such that if $w_1, ..., w_k \in \mathbb{R}^d$ with w_i randomly chosen from $w_i \sim \mathcal{N}(\mathbf{0}, O(d \log d)\mathbf{I_{d \times d}})$ and $b_1, ..., b_k$ be randomly chosen at uniform from $[-1, 1]$, then with high probability, we can choose an initial point $(\boldsymbol{a^{(0)}}, \boldsymbol{\theta^{(0)}})$ such that after running SecondGD (Algorithm 2) on the regularized objective $G(\boldsymbol{a}, \boldsymbol{\theta})$ for at most $(d/\epsilon)^{O(d)}$ iterations, there exists an i, j such that $\|\theta_i - w_j\| < \epsilon$.*

We start by stating a lemma concerning the construction of an almost λ-harmonic function on $\mathbb{R}^d$. The construction is given in Appendix B and uses a linear combination of realizable

Algorithm 1 $x = HD(L, x_0, T, \alpha)$

Input: $L : \mathcal{M} \to \mathbb{R}$; $x_0 \in \mathcal{M}$; $T \in \mathbb{N}$; $\alpha \in \mathbb{R}$
Initialize $x \leftarrow x_0$
for $i = 1$ **to** T **do**
  Find unit eigenvector v_{min} corresponding to $\lambda_{min}(\nabla^2 f(x))$
  $\beta \leftarrow -\alpha\lambda_{min}(\nabla^2 f(x))\text{sign}(\nabla f(x)^T v_{min})$
  $x \leftarrow x + \beta v_{min}$

Algorithm 2 $x = SecondGD(L, x_0, T, \alpha, \eta, \gamma)$

Input: $L : \mathcal{M} \to \mathbb{R}$; $x_0 \in \mathcal{M}$; $T \in \mathbb{N}$; $\alpha, \eta, \gamma \in \mathbb{R}$
for $i = 1$ **to** T **do**
  if $\|\nabla L(x_{i-1})\| \geq \eta$ **then** $x_i \leftarrow x_{i-1} - \alpha\nabla L(x_{i-1})$
  else $x_i \leftarrow HD(L, x_{i-1}, 1, \alpha)$
  if $L(x_i) \geq L(x_{i-1}) - min(\alpha\eta^2/2, \alpha^2\gamma^3/2)$ **then return** x_{i-1}

potentials that correspond to an activation function of the indicator function of a n-sphere. By using Fourier analysis and Theorem 3, we can finish the construction of our almost λ-harmonic potential.

▶ **Lemma 13.** *Let $\mathcal{M} = \mathbb{R}^d$ for $d \equiv 3 \mod 4$. Then, for any $\epsilon \in (0, 1)$, we can construct a radial activation $\sigma_\epsilon(r)$ such that the corresponding radial potential $\Phi_\epsilon(r)$ is λ-harmonic for $r \geq \epsilon$.*

Furthermore, we have ${\Phi_\epsilon}^{(d-1)}(r) \geq 0$ for all $r > 0$, ${\Phi_\epsilon}^{(k)}(r) \geq 0$, and ${\Phi_\epsilon}^{(k+1)}(r) \leq 0$ for all $r > 0$ and $d - 3 \geq k \geq 0$ even.

When $\lambda = 1$, $|\Phi_\epsilon^{(k)}(r)| \leq O((d/\epsilon)^{2d})$ for all $0 \leq k \leq d - 1$. And when $r \geq \epsilon$, $\Omega(e^{-r}r^{2-d}(d/\epsilon)^{-2d}) \leq \Phi_\epsilon(r) \leq O((1 + r)^d e^{1-r}(r)^{2-d})$ and $\Omega(e^{-r}r^{1-d}(d/\epsilon)^{-2d}) \leq |\Phi'_\epsilon(r)| \leq O((d + r)(1 + r)^d e^{1-r}r^{1-d})$

Our next lemma use the almost λ-harmonic properties to show that at an almost stationary point of G, we must have converged close to some w_j as long as our charges a_i are not too small. The proof is similar to Theorem 11. Then, the following lemma relates the magnitude of the charges a_i to the progress made in the objective function.

▶ **Lemma 14.** *Let $\mathcal{M} = \mathbb{R}^d$ for $d \equiv 3 \mod 4$ and let G be the regularized loss corresponding to the activation σ_ϵ given by Lemma 13 with $\lambda = 1$. For any $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$, if $(\boldsymbol{a}, \boldsymbol{\theta}) \in \mathcal{M}_{G,\delta}$, then for all i, either 1) there exists j such that $\|\theta_i - w_j\| < k\epsilon$ or 2) $a_i^2 < 2kd\delta$.*

▶ **Lemma 15.** *Assume the conditions of Lemma 14. If $\sqrt{G(\mathbf{a}, \boldsymbol{\theta})} \leq \sqrt{G(\mathbf{0}, \mathbf{0})} - \delta$ and $(\boldsymbol{a}, \boldsymbol{\theta}) \in \mathcal{M}_{G,\delta^2/(2k^3d)}$, then there exists some i, j such that $\|\theta_i - w_j\| < k\epsilon$.*

Finally, we guarantee that our initialization substantially decreases our objective function. Together with our previous lemmas, it will imply that we must be close to some w_j upon convergence. This is the overview of the proof of Theorem 12, presented below.

▶ **Lemma 16.** *Assume the conditions of Theorem 12 and Lemma 14. With high probability, we can initialize $(\boldsymbol{a}^{(\mathbf{0})}, \boldsymbol{\theta}^{(\mathbf{0})})$ such that $\sqrt{G(\mathbf{a}^{(\mathbf{0})}, \boldsymbol{\theta}^{(\mathbf{0})})} \leq \sqrt{G(\mathbf{0}, \mathbf{0})} - \delta$ with $\delta = (d/\epsilon)^{-O(d)}$.*

Proof of Theorem 12. Let our potential $\Phi_{\epsilon/k}$ be the one as constructed in Lemma 13 that is 1-harmonic for all $r \geq \epsilon/k$ and as always, $k = \text{poly}(d)$. First, by Lemma 16, we can

Algorithm 3 Node-wise Descent Algorithm

Input: $(\boldsymbol{a}, \boldsymbol{\theta}) = (a_1, ..., a_k, \theta_1, ..., \theta_k), a_i \in \mathbb{R}, \theta_i \in \mathcal{M}$; $T \in \mathbb{N}$; L; $\alpha, \eta, \gamma \in \mathbb{R}$;
for $i = 1$ **to** k **do**
  Initialize (a_i, θ_i)
  $(a_i, \theta_i) = SecondGD\left(L_{a_i,\theta_i}, (a_i, \theta_i), T, \alpha, \eta, \gamma\right)$
return $a = (a_1, ..., a_k), \theta = (\theta_1, ..., \theta_k)$

initialize $(\boldsymbol{a^{(0)}}, \boldsymbol{\theta^{(0)}})$ such that $\sqrt{G(\boldsymbol{a^{(0)}}, \boldsymbol{\theta^{(0)}})} \leq \sqrt{G(\mathbf{0}, \mathbf{0})} - \delta$ for $\delta = (d/\epsilon)^{-O(d)}$. If we set $\alpha = (d/\epsilon)^{-O(d)}$ and $\eta = \gamma = \delta^2/(2k^3 d)$, then running Algorithm 2 will terminate and return some $(\boldsymbol{a}, \boldsymbol{\theta})$ in at most $(d/\epsilon)^{O(d)}$ iterations. This is because our algorithm ensures that our objective function decreases by at least $\min(\alpha\eta^2/2, \alpha^2\gamma^3/2)$ at each iteration, $G(\mathbf{0}, \mathbf{0})$ is bounded by $O(k)$, and $G \geq 0$ is non-negative.

Let $\boldsymbol{\theta} = (\theta_1, ...\theta_k)$. If there exists θ_i, w_j such that $\|\theta_i - w_j\| < \epsilon$, then we are done. Otherwise, we claim that $(\boldsymbol{a}, \boldsymbol{\theta}) \in \mathcal{M}_{G,\delta^2/(2k^3 d)}$. For the sake of contradiction, assume otherwise. By our algorithm termination conditions, then it must be that after one step of gradient or Hessian descent from $(\boldsymbol{a}, \boldsymbol{\theta})$, we reach some $(\boldsymbol{a'}, \boldsymbol{\theta'})$ and $G(\boldsymbol{a'}, \boldsymbol{\theta'}) > G(\boldsymbol{a}, \boldsymbol{\theta}) - \min(\alpha\eta^2/2, \alpha^2\gamma^3/2)$.

Now, Lemma 13 ensures all first three derivatives of $\Phi_{\epsilon/k}$ are bounded by $O((dk/\epsilon)^{2d})$, except at $w_1, ..., w_k$. Furthermore, since there do not exist θ_i, w_j such that $\|\theta_i - w_j\| < \epsilon$, G is three-times continuously differentiable within a $\alpha(dk/\epsilon)^{2d} = (d/\epsilon)^{-O(d)}$ neighborhood of $\boldsymbol{\theta}$. Therefore, by Lemma 18 and 19 in the appendix, we must have $G(\boldsymbol{a'}, \boldsymbol{\theta'}) \leq G(\boldsymbol{a}, \boldsymbol{\theta}) - \min(\alpha\eta^2/2, \alpha^2\gamma^3/2)$, a contradiction. Lastly, since our algorithm maintains that our objective function is decreasing, so $\sqrt{G(\boldsymbol{a}, \boldsymbol{\theta})} \leq \sqrt{G(\mathbf{0}, \mathbf{0})} - \delta$. Finally, we conclude by Lemma 15. ◀

4.1 Node-by-Node Analysis

We cannot easily analyze the convergence of gradient descent to the global minima when all θ_i are simultaneously moving since the pairwise interaction terms between the θ_i present complications, even with added regularization. Instead, we run a greedy node-wise descent (Algorithm 3) to learn the hidden weights, i.e. we run a descent algorithm with respect to (a_i, θ_i) sequentially. The main idea is that after running SGD with respect to θ_1, θ_1 should be close to some w_j for some j. Then, we can carefully induct and show that θ_2 must be some other w_k for $k \neq j$ and so on.

Let $L_1(a_1, \theta_1)$ be the objective L restricted to a_1, θ_1 being variable, and $a_2, ..., a_k = 0$ are fixed. The tighter control on the movements of θ_1 allows us to remove our regularization. While our previous guarantees before allow us to reach a ϵ-neighborhood of w_j when running SGD on L_1, we will strengthen our guarantees to reach a $(d/\epsilon)^{-O(d)}$-neighborhood of w_j, by reasoning about the first derivatives of our potential in an ϵ-neighborhood of w_j. By similar argumentation as before, we will be able to derive the following convergence guarantees for node-wise training.

▶ **Theorem 17.** *Let $\mathcal{M} = \mathbb{R}^d$ and $d \equiv 3 \mod 4$ and let L be as in 1 and $k = \text{poly}(d)$. For all $\epsilon \in (0, 1)$, we can construct an activation σ_ϵ such that if $w_1, ..., w_k \in \mathbb{R}^d$ with w_i randomly chosen from $w_i \sim \mathcal{N}(\mathbf{0}, O(d \log d)\mathbf{I_{d\times d}})$ and $b_1, ..., b_k$ be randomly chosen at uniform from $[-1, 1]$, then with high probability, after running nodewise descent (Algorithm 3) on the objective L for at most $(d/\epsilon)^{O(d)}$ iterations, $(\boldsymbol{a}, \boldsymbol{\theta})$ is in a $(d/\epsilon)^{-O(d)}$ neighborhood of the global minima.*

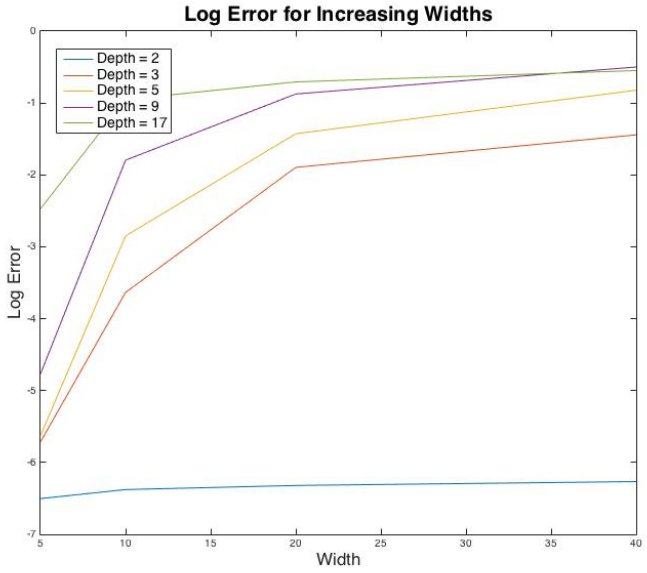

Figure 2 Test Error of Varying-Depth Networks vs. Width

Table 2 Test Error of Learning Neural Networks of Various Depth and Width

	Width 5	Width 10	Width 20	Width 40
Depth 2	0.0015	0.0017	0.0018	0.0019
Depth 3	0.0033	0.0264	0.1503	0.2362
Depth 5	0.0036	0.0579	0.2400	0.4397
Depth 9	0.0085	0.1662	0.4171	0.6071
Depth 17	0.0845	0.3862	0.4934	0.5777

5 Experiments

For our experiments, our training data is given by $(x_i, f(x_i))$, where x_i are randomly chosen from a standard Gaussian in $\mathbb{R}^d$ and f is a randomly generated neural network with weights chosen from a standard Gaussian. We run gradient descent (Algorithm 4) on the empirical loss, with stepsize around $\alpha = 10^{-5}$, for $T = 10^6$ iterations. The nonlinearity used at each node is sigmoid from -1 to 1, including the output node, unlike the assumptions in the theoretical analysis. A random guess for the network will result in a mean squared error of around 1. Our experiments (see Fig 1) show that for depth-2 neural networks, even with non-linear outputs, the training error diminishes quickly to under 0.002. This seems to hold even when the width, the number of hidden nodes, is substantially increased (even up to 125 nodes), but depth is held constant; although as the number of nodes increases, the rate of decrease is slower. This substantiates our claim that depth-2 neural networks are learnable.

However, it seems that for depth greater than 2, the test error becomes significant when width is high (see Fig 2). Even for depth 3 networks, the increase in depth impedes the learnability of the neural network and the training error does not get close enough to 0. It seems that for neural networks with greater depth, positive convergence results in practice are elusive. We note that we are using training error as a measure of success, so it's possible that the true underlying parameters are not learned.

References

1 Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

2 Vladimir I Arnold, Valery V Kozlov, and Anatoly I Neishtadt. Mathematical aspects of classical and celestial mechanics. *Encyclopaedia Math. Sci*, 3:1–291, 1985.

3 Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *ICML*, pages 584–592, 2014.

4 Peter Auer, Mark Herbster, and Manfred K. Warmuth. Exponentially many local minima for single neurons. In David S. Touretzky, Michael Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, November 27-30, 1995*, pages 316–322. MIT Press, 1995. URL: `http://papers.nips.cc/paper/1028-exponentially-many-local-minima-for-single-neurons`.

5 Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is np-complete. In David Haussler and Leonard Pitt, editors, *Proceedings of the First Annual Workshop on Computational Learning Theory, COLT '88, Cambridge, MA, USA, August 3-5, 1988.*, pages 9–18. ACM/MIT, 1988. URL: `http://dl.acm.org/citation.cfm?id=93033`.

6 Martin L Brady, Raghu Raghavan, and Joseph Slawny. Back propagation fails to separate where perceptrons succeed. *IEEE Transactions on Circuits and Systems*, 36(5):665–674, 1989.

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

8 Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

9 Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289, 2015. `arXiv:1511.07289`.

10 Amit Daniely. Sgd learns the conjugate kernel class of the network. *arXiv preprint arXiv:1702.08503*, 2017.

11 Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

12 Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

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

14 Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. *arXiv preprint arXiv:1611.10258*, 2016.

15 Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Generalization bounds for neural networks through tensor factorization. *CoRR*, abs/1506.08473, 2015. `arXiv:1506.08473`.

16 Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

17 Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

18 Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 553–562. IEEE, 2006.

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

20 Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

21 Anish Shah, Eashan Kadam, Hena Shah, Sameer Shinde, and Sandip Shingade. Deep residual networks with exponential linear unit. In *Proceedings of the Third International Symposium on Computer Vision and the Internet*, pages 59–65. ACM, 2016.

22 Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2011.

23 Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *CoRR*, abs/1605.08361, 2016. `arXiv:1605.08361`.

24 Yuandong Tian. 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.

25 Qiuyi Zhang, Rina Panigrahy, and Sushant Sachdeva. Electron-proton dynamics in deep learning. *CoRR*, abs/1702.00458, 2017. `arXiv:1702.00458`.

26 Yuchen Zhang, Jason D Lee, and Michael I Jordan. l1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pages 993–1001, 2016.

27 Yuchen Zhang, Jason D. Lee, Martin J. Wainwright, and Michael I. Jordan. Learning halfspaces and neural networks with random initialization. *CoRR*, abs/1511.07948, 2015. `arXiv:1511.07948`.

A Electron-Proton Dynamics

▶ **Theorem 5.** *Let Φ be a symmetric potential and L be as in* (1)*. Running continuous gradient descent on $\frac{1}{2}L$ with respect to θ, initialized at $(\theta_1, ..., \theta_k)$ produces the same dynamics as Electron-Proton Dynamics under 2Φ with fixed particles at $w_1, ..., w_k$ with respective charges $b_1, .., b_k$ and moving particles at $\theta_1, ..., \theta_k$ with respective charges $a_1, ..., a_k$.*

Proof. The initial values are the same. Notice that continuous gradient descent on $L(\boldsymbol{a}, \boldsymbol{\theta})$ with respect to θ produces dynamics given by $\frac{d\theta_i(t)}{dt} = -\nabla_{\theta_i} L(\boldsymbol{a}, \boldsymbol{\theta})$. Therefore,

$$\frac{d\theta_i(t)}{dt} = -2\sum_{j\neq i} a_i a_j \nabla_{\theta_i}\Phi(\theta_i, \theta_j) - 2\sum_{j=1}^{k} a_i b_j \nabla_{\theta_i}\Phi(\theta_i, w_j)$$

And gradient descent does not move w_i. By definition, the dynamics corresponds to Electron-Proton Dynamics as claimed. ◀

B Realizable Potentials

This section can be found in the full version of this paper on ArXiv [25].

C Earnshaw's Theorem

▶ **Theorem 9.** (Earnshaw's Theorem. See [2]) *Let $\mathcal{M} = \mathbb{R}^d$ and let Φ be harmonic and L be as in* (2)*. Then, L admits no differentiable strict local minima.*

Proof. If $(\boldsymbol{a}, \boldsymbol{\theta})$ is a differentiable strict local minima, then for any i, we must have

$$\nabla_{\theta_i} L = 0, \text{ and } \mathrm{Tr}(\nabla^2_{\theta_i} L) > 0.$$

Algorithm 4 $x = GD(L, x_0, T, \alpha)$

Input: $L : \mathcal{M} \to \mathbb{R}$; $x_0 \in \mathcal{M}$; $T \in \mathbb{N}$; $\alpha \in \mathbb{R}$
Initialize $x = x_0$
for $i = 1$ **to** T **do**
  $x = x - \alpha \nabla L(x)$
  $x = \Pi_{\mathcal{M}} x$

Since Φ is harmonic, we also have

$$\mathrm{Tr}(\nabla^2_{\theta_i} L(\theta_1, ..., \theta_n)) = \Delta_{\theta_i} L = 2\sum_{j \neq i} a_i a_j \Delta_{\theta_i} \Phi(\theta_i, \theta_j) + 2\sum_{j=1}^{k} a_i b_j \Delta_{\theta_i} \Phi(\theta_i, w_j) = 0,$$

which is a contradiction. In the first line, there is a factor of 2 by symmetry. ◀

D Descent Lemmas and Iteration Bounds

▶ **Lemma 18.** *Let $f : \Omega \to \mathbb{R}$ be a thrice differentiable function such that $|f(y)| \leq B_0, \|\nabla f(y)\| \leq B_1, \|\nabla^2 f(y)\| \leq B_2, \|\nabla^2 f(z) - \nabla^2 L(y)\| \leq B_3 \|z - y\|$ for all y, z in a (αB_1)-neighborhood of x. If $\|\nabla f(x)\| \geq \eta$ and x' is reached after one iteration of gradient descent (Algorithm 4) with stepsize $\alpha \leq \frac{1}{B_2}$, then $\|x' - x\| \leq \alpha B_1$ and $f(x') \leq f(x) - \alpha\eta^2/2$.*

Proof. The gradient descent step is given by $x' = x - \alpha \nabla f(x)$. The bound on $\|x' - x\|$ is clear since $\|\nabla f(x)\| \leq B_1$.

$$\begin{aligned} f(x') &\leq f(x) - \alpha \nabla f(x)^T \nabla f(x)^T + \alpha^2 \frac{B_2}{2} \|\nabla f(x)\|^2 \\ &\leq f(x) - (\alpha - \alpha^2 \frac{B_2}{2})\eta^2 \end{aligned}$$

For $0 \leq \alpha \leq \frac{1}{B_2}$, we have $\alpha - \alpha^2 B_2/2 \geq \alpha/2$, and our lemma follows. ◀

▶ **Lemma 19.** *Let $f : \Omega \to \mathbb{R}$ be a thrice differentiable function such that $|f(y)| \leq B_0, \|\nabla f(y)\| \leq B_1, \|\nabla^2 f(y)\| \leq B_2, \|\nabla^2 f(z) - \nabla^2 L(y)\| \leq B_3 \|z - y\|$ for all y, z in a (αB_2)-neighborhood of x. If $\lambda_{min}(\nabla^2 f(x)) \leq -\gamma$ and x' is reached after one iteration of Hessian descent (Algorithm 1) with stepsize $\alpha \leq \frac{1}{B_3}$, then $\|x' - x\| \leq \alpha B_2$ and $f(x') \leq f(x) - \alpha^2\gamma^3/2$.*

Proof. The gradient descent step is given by $x' = x + \beta v_{min}$, where v_{min} is the unit eigenvector corresponding to $\lambda_{min}(\nabla^2 f(x))$ and $\beta = -\alpha \lambda_{min}(\nabla^2 f(x)) sgn(\nabla f(x)^T v_{min})$. Our bound on $\|x' - x\|$ is clear since $|\lambda_{min}(\nabla^2 f(x))| \leq B_2$.

$$\begin{aligned} f(x') &\leq f(x) + \beta \nabla f(x)^T v_{min} + \beta^2 v_{min}^T \nabla^2 f(x) v_{min} + \frac{B_3}{6} |\beta|^3 \|v_{min}\|^3 \\ &\leq f(x) - |\beta|^2 \gamma + \frac{B_3}{6} |\beta|^3 \end{aligned}$$

The last inequality holds since the sign of β is chosen so that $\beta \nabla f(x)^T v_{min} \leq 0$. Now, since $|\beta| = \alpha\gamma \leq \frac{\gamma}{B_3}$, $-|\beta|^2\gamma + \frac{B_3}{6}|\beta|^3 \leq -\alpha^2\gamma^3/2$. ◀

E Convergence of Almost λ-Harmonic Potentials

▶ **Lemma 20.** *Let $\mathcal{M} = \mathbb{R}^d$ for $d \equiv 3 \mod 4$ and let G be the regularized loss corresponding to the activation σ_ϵ given by Lemma 13 with $\lambda = 1$. For any $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$,*

if $(\boldsymbol{a}, \boldsymbol{\theta}) \in \mathcal{M}_{G,\delta}$*, then for all* i*, either 1) there exists* j *such that* $\|\theta_i - w_j\| < k\epsilon$ *or 2)* $a_i^2 < 2kd\delta$*.*

Proof. The proof is similar to Theorem 11. Let Φ_ϵ be the realizable potential in 13 such that $\Phi_\epsilon(r)$ is λ-harmonic when $r \geq \epsilon$ with $\lambda = 1$. Note that $\Phi_\epsilon(0) = 1$ is normalized. And let $(\boldsymbol{a}, \boldsymbol{\theta}) \in \mathcal{M}_{G,\delta}$.

WLOG, consider θ_1 and a initial set $S_0 = \{\theta_1\}$ containing it. For a finite set of points S and a point x, define $d(x, S) = \min_{y \in S} \|x - y\|$. Then, we consider the following set growing process. If there exists $\theta_i, w_i \notin S_j$ such that $d(\theta_i, S_j) < \epsilon$ or $d(w_i, S_j) < \epsilon$, add θ_i, w_i to S_j to form S_{j+1}. Otherwise, we stop the process. We grow S_0 to until the process terminates and we have the grown set S.

If there is some $w_j \in S$, then it must be the case that there exists $j_1, \cdots j_q$ such that $\|\theta_1 - \theta_{j_1}\| < \epsilon$ and $\|\theta_{j_i} - \theta_{j_{i+1}}\| < \epsilon$, and $\|\theta_{j_q} - w_j\| < \epsilon$ for some w_j. So, there exists j, such that $\|\theta_1 - w_j\| < k\epsilon$.

Otherwise, notice that for each $\theta_i \in S$, $\|w_j - \theta_i\| \geq \epsilon$ for all j, and $\|\theta_i - \theta_j\| \geq \epsilon$ for all $\theta_j \notin S$. WLOG, let $S = \{\theta_1, \ldots, \theta_l\}$.

We consider changing all $\theta_1, \ldots, \theta_l$ by the same v and define

$$H(\mathbf{a}, v) = G(\mathbf{a}, \theta_1 + v, ..., \theta_l + v, \theta_{l+1} \ldots, \theta_k).$$

The optimality conditions on $\mathbf{a}$ are

$$\left|\frac{\partial H}{\partial a_i}\right| = |4a_i + 2\sum_{j \neq i} a_j \Phi_\epsilon(\theta_i, \theta_j) + 2\sum_{j=1}^{k} b_j \Phi_\epsilon(\theta_i, w_j)| \leq \delta$$

Next, since $\Phi_\epsilon(r)$ is λ-harmonic for $r \geq \epsilon$, we may calculate the Laplacian of H as

$$\begin{aligned}
\Delta_v H &= \sum_{i=1}^{l} \lambda \left(2\sum_{j=1}^{k} a_i b_j \Phi_\epsilon(\theta_i, w_j) + 2\sum_{j=l+1}^{k} a_i a_j \Phi_\epsilon(\theta_i, \theta_j) \right) \\
&\leq \sum_{i=1}^{l} \lambda \left(-4a_i^2 - 2\sum_{j=1, j \neq i}^{l} a_i a_j \Phi_\epsilon(\theta_i, \theta_j) \right) + \delta \sum_{i=1}^{l} \lambda |a_i| \\
&= -2\lambda\, \mathbb{E}\left[\left(\sum_{i=1}^{l} a_i \sigma(\theta_i, X) \right)^2 \right] - 2\lambda \sum_{i=1}^{l} a_i^2 + \delta\lambda \sum_{i=1}^{l} |a_i|
\end{aligned}$$

The second line follows from our optimality conditions and the third line follows from completing the square. Since $(\boldsymbol{a}, \boldsymbol{\theta}) \in \mathcal{M}_{G,\delta}$, we have $\Delta_v H \geq -2kd\delta$. Let $S = \sum_{i=1}^{l} a_i^2$. Then, by Cauchy-Schwarz, we have $-2\lambda S + \delta\lambda\sqrt{k}\sqrt{S} \geq -2kd\delta$. When $S \geq \delta^2 k$, we see that $-\lambda S \geq -2\lambda S + \delta\lambda\sqrt{k}\sqrt{S} \geq -2kd\delta$. Therefore, $S \leq 2kd\delta/\lambda$.

We conclude that $S \leq \max(\delta^2 k, 2kd\delta/\lambda) \leq 2kd\delta/\lambda$ since $\delta \leq 1 \leq 2d/\lambda$ and $\lambda = 1$. Therefore, $a_i^2 \leq 2kd\delta$. ◀

▶ **Lemma 21.** *Assume the conditions of Lemma 14. If* $\sqrt{G(\mathbf{a}, \boldsymbol{\theta})} \leq \sqrt{G(\mathbf{0}, \mathbf{0})} - \delta$ *and* $(\boldsymbol{a}, \boldsymbol{\theta}) \in \mathcal{M}_{G,\delta^2/(2k^3 d)}$*, then there exists some* i, j *such that* $\|\theta_i - w_j\| < k\epsilon$*.*

Proof. If there does not exists i, j such that $\|\theta_i - w_j\| < k\epsilon$, then by Lemma 14, this implies $a_i^2 < \delta^2/k^2$ for all i. Now, for a integrable function $f(x)$, $\|f\|_X = \sqrt{\mathbb{E}_X[f(X)^2]}$ is a norm. Therefore, if $f(x) = \sum_i b_i \sigma(w_i, x)$ be our true target function, we conclude that by triangle

inequality

$$\sqrt{G(\boldsymbol{a},\boldsymbol{\theta})} \geq \left\| \sum_{i=1}^{k} a_i \sigma(\theta_i, x) - f(x) \right\|_X \geq \|f(x)\|_X - \sum_{i=1}^{k} \|a_i \sigma(\theta_i, x)\|_X \geq \sqrt{G(\mathbf{0},\mathbf{0})} - \delta$$

This gives a contradiction, so we conclude that there must exist i, j such that θ_i is in a $k\epsilon$ neighborhood of w_j. ◀

▶ **Lemma 22.** *Assume the conditions of Theorem 12 and Lemma 14. With high probability, we can initialize* $(\boldsymbol{a^{(0)}}, \boldsymbol{\theta^{(0)}})$ *such that* $\sqrt{G(\mathbf{a^{(0)}}, \boldsymbol{\theta^{(0)}})} \leq \sqrt{G(\mathbf{0},\mathbf{0})} - \delta$ *with* $\delta = (d/\epsilon)^{-O(d)}$.

Proof. Consider choosing $\theta_1 = \mathbf{0}$ and then optimizing a_1. Given θ_1, the loss decrease is:

$$G(a_1, \mathbf{0}) - G(\mathbf{0},\mathbf{0}) = \min_{a_1} 2a_1^2 + 2\sum_{j=1}^{k} a_1 b_j \Phi_\epsilon(\mathbf{0}, w_j) = -\frac{1}{2}\left(\sum_{j=1}^{k} b_j \Phi_\epsilon(\mathbf{0}, w_j)\right)^2$$

Because w_j are random Gaussians with variance $O(d \log d)$, we have $\|w_j\| \leq O(d \log d)$ with high probability for all j. By Lemma 13, our potential satisfies $\Phi_\epsilon(\mathbf{0}, w_j) \geq (d/\epsilon)^{-O(d)}$. And since b_j are uniformly chosen in $[-1, 1]$, we conclude that with high probability over the choices of b_j, $-\frac{1}{2}\left(\sum_{j=1}^{k} b_j \Phi(\theta_1, w_j)\right)^2 \geq (d/\epsilon)^{-O(d)}$ by appealing to Chebyshev's inequality on the squared term.

Therefore, we conclude that with high probability, $G(a_1, \mathbf{0}) \leq G(\mathbf{0},\mathbf{0}) - \frac{1}{2}(d/\epsilon)^{-O(d)}$. Let $\sqrt{G(a_1,\mathbf{0})} = \sqrt{G(\mathbf{0},\mathbf{0})} - \Delta \geq 0$. Squaring and rearranging gives $\Delta \geq \frac{1}{4\sqrt{G(\mathbf{0},\mathbf{0})}}(d/\epsilon)^{-O(d)}$. Since $G(\mathbf{0},\mathbf{0}) \leq O(k) = O(\text{poly}(d))$, we are done. ◀

E.1 Node by Node Analysis

The proofs in this section can be found in the full version of this paper on ArXiv [25].

▶ **Lemma 23.** *Let* $\mathcal{M} = \mathbb{R}^d$ *for* $d \equiv 3 \mod 4$ *and let* L_1 *be the loss restricted to* (a_1, θ_1) *corresponding to the activation function* σ_ϵ *given by Lemma 13 with* $\lambda = 1$. *For any* $\epsilon \in (0,1)$ *and* $\delta \in (0,1)$, *we can construct* σ_ϵ *such that if* $(a_1, \theta_1) \in \mathcal{M}_{L_1,\delta}$, *then for all* i, *either 1) there exists* j *such that* $\|\theta_1 - w_j\| < \epsilon$ *or 2)* $a_1^2 < 2d\delta$.

▶ **Lemma 24.** *Assume the conditions of Lemma 23. If* $\sqrt{L_1(a_1,\theta_1)} \leq \sqrt{L_1(0,0)} - \delta$ *and* $(a_1, \theta_1) \in \mathcal{M}_{G,\delta^2/(2d)}$, *then there exists some* j *such that* $\|\theta_1 - w_j\| < \epsilon$.

▶ **Lemma 25.** *Assume the conditions of Theorem 27 and Lemma 23. If* $\|\theta_1 - w_j\| \leq d$ *and* $|b_j| \geq 1/\text{poly}(d)$ *and* $|a_1 - a_1^*(\theta_1)| \leq (d/\epsilon)^{-O(d)}$ *is almost optimal and for* i, $\|w_i - w_j\| \geq \Omega(d \log d)$, *then* $-\nabla_{\theta_1} L_1 = \zeta \frac{w_j - \theta_1}{\|\theta_1 - w_j\|} + \xi$ *with* $\zeta \geq \frac{1}{\text{poly}(d)}(d/\epsilon)^{-8d}$ *and* $\xi \leq (d/\epsilon)^{-O(d)}$.

▶ **Lemma 26** (Node-wise Initialization)**.** *Assume the conditions of Theorem 27 and Lemma 23. With high probability, we can initialize* $(a_1^{(0)}, \theta_1^{(0)})$ *such that* $\sqrt{L(a_1^{(0)}, \theta_1^{(0)})} \leq \sqrt{L(0,0)} - \delta$ *with* $\delta = \frac{1}{\text{poly}(d)}(d/\epsilon)^{-18d}$ *in time* $\log(d)^{O(d)}$.

▶ **Lemma 27.** *Assume the conditions of Lemma 23. Also, assume* $b_1, ..., b_k$ *are any numbers in* $[-1, 1]$ *and* $w_1, ..., w_k \in \mathbb{R}^d$ *satisfy* $\|w_i\| \leq O(d \log d)$ *for all* i *and there exists some* $|b_j| \geq 1/\text{poly}(d)$ *with* $\|w_i - w_j\| \geq \Omega(d \log d)$ *for all* i.

Then with high probability, we can choose an initial point $(a_1^{(0)}, \theta_1^{(0)})$ *such that after running SecondGD (Algorithm 2) on the restricted regularized objective* $L_1(a_1, \theta_1)$ *for at most* $(d/\epsilon)^{O(d)}$ *iterations, there exists some* w_j *such that* $\|\theta_1 - w_j\| < \epsilon$. *Furthermore, if* $|b_j| \geq 1/\text{poly}(d)$ *and* $\|w_i - w_j\| \geq \Omega(d \log d)$ *for all* i, *then* $\|\theta_1 - w_j\| < (d/\epsilon)^{-O(d)}$ *and* $|a + b_j| < (d/\epsilon)^{-O(d)}$.

▶ **Theorem 17.** *Let $\mathcal{M} = \mathbb{R}^d$ and $d \equiv 3 \mod 4$ and let L be as in 1 and $k = \text{poly}(d)$. For all $\epsilon \in (0,1)$, we can construct an activation σ_ϵ such that if $w_1, ..., w_k \in \mathbb{R}^d$ with w_i randomly chosen from $w_i \sim \mathcal{N}(\mathbf{0}, O(d \log d)\mathbf{I_{d \times d}})$ and $b_1, ..., b_k$ be randomly chosen at uniform from $[-1, 1]$, then with high probability, after running nodewise descent (Algorithm 3) on the objective L for at most $(d/\epsilon)^{O(d)}$ iterations, $(\boldsymbol{a}, \boldsymbol{\theta})$ is in a $(d/\epsilon)^{-O(d)}$ neighborhood of the global minima.*

F Common Activations

First, we consider the sign activation function. Under restrictions on the size of the input dimension or the number of hidden units, we can prove convergence results under the sign activation function, as it gives rise to a harmonic potential.

▶ **Assumption 1.** *All output weights $b_i = 1$ and therefore the output weights $a_i = -b_i = -1$ are fixed throughout the learning algorithm.*

▶ **Lemma 28.** *Let $\mathcal{M} = S^1$ and let Assumption 1 hold. Let L be as in (2) and σ is the sign activation function. Then L admits no strict local minima, except at the global minima.*

We cannot simply analyze the convergence of GD on all θ_i simultaneously since as before, the pairwise interaction terms between the θ_i present complications. Therefore, we now only consider the convergence guarantee of gradient descent on the first node, θ_1, to some w_j, while the other nodes are inactive (i.e. $a_2, ..., a_k = 0$). In essence, we are working with the following simplified loss function.

$$L(a_1, \theta_1) = a_1^2 \Phi(\theta_1, \theta_1) + 2 \sum_{j=1}^{k} a_1 b_j \Phi(\theta_1, w_j) \tag{3}$$

▶ **Lemma 29.** *Let $\mathcal{M} = S^1$ and L be as in (3) and σ is the sign activation function. Then, almost surely over random choices of $b_1, ..., b_k$, all local minima of L are at $\pm w_j$.*

For the polynomial activation and potential functions, we also can show convergence under orthogonality assumptions on w_j. Note that the realizability of polynomial potentials is guaranteed in Section B.

▶ **Theorem 30.** *Let $\mathcal{M} = S^{d-1}$. Let $w_1, ..., w_k$ be orthonormal vectors in $\mathbb{R}^d$ and Φ is of the form $\Phi(\theta, w) = (\theta^T w)^l$ for some fixed integer $l \geq 3$. Let L be as in (3). Then, all critical points of L are not local minima, except when $\theta_1 = w_j$ for some j.*

F.1 Convergence of Sign Activation

▶ **Lemma 31.** *Let $\mathcal{M} = S^1$ and let Assumption 1 hold. Let L be as in (2) and σ is the sign activation function. Then L admits no strict local minima, except at the global minima.*

Proof. We will first argue that unless all the electrons and protons have matched up as a permutation it cannot be a strict local minimum and then argue that the global minimum is a strict local minimum.

First note that if some electron and proton have merged, we can remove such pairs and argue about the remaining configuration of charges. So WLOG we assume there are no such overlapping electron and proton.

First consider the case when there is an isolated electron e and there is no charge diagonally opposite to it. In this case look at the two semicircles on the left and the right half of the circle around the isolated electron – let q_1 and q_2 be the net charges in the left and the right semi-circles. Note that $q_1 \neq q_2$ since they are integers and $q_1 + q_2 = +1$ which is odd. So by moving the electron slightly to the side with the larger charge you decrease the potential.

If there is a proton opposite the isolated electron the argument becomes simpler as the proton benefits the motion of the electron in either the left or right direction. So the only way the electron does not benefit by moving in either direction is that $q_1 = -1$ and $q_2 = -1$ which is impossible.

If there is an electron opposite the isolated electron then the combination of these two diagonally opposing electrons have a zero effect on every other charge. So it is possible rotate this pair jointly keeping them opposed in any way and not change the potential. So this is not a strict local minimum.

Next if there is a clump of isolated electrons with no charge on the diagonally opposite point then again as before if $q_1 \neq q_2$ we are done. If $q_1 = q_2$ then the the electrons in the clump locally are unaffected by the remaining charges. So now by splitting the clump into two groups and moving them apart infinitesimally we will decrease the potential.

Now if there is only protons in the diagonally opposite position an isolated electron again we are done as in the case when there is one electron diagonally opposite one proton.

Finally if there is only electrons diagonally opposite a clump of electrons again we are done as we have found at least one pair of opposing electrons that can be jointly rotated in any way.

Next we will argue that a permutation matching up is a strict local minumum. For this we will assume that no two protons are diagonally opposite each other (as they can be removed without affecting the function). Now given a perfect matching up of electrons and protons, if we perturb the electrons in any way infinitesimally, then any isolated clump of electrons can be moved slightly to the left or right to improve the potential. ◀

▶ **Lemma 32.** *Let $\mathcal{M} = S^1$ and L be as in* (3) *and σ is the sign activation function. Then, almost surely over random choices of $b_1, ..., b_k$, all local minima of L are at $\pm w_j$.*

Proof. In S^1, notice that the pairwise potential function is $\Phi(\theta, w) = 1 - 2\cos^{-1}(\theta^T w)/\pi = 1 - 2\alpha/\pi$, where α is the angle between θ, w. So, let us parameterize in polar coordinates, calling our true parameters as $\widetilde{w_1}, ..., \widetilde{w_k} \in [0, 2\pi]$ and rewriting our loss as a function of $\widetilde{\theta} \in [0, 2\pi]$.

Since Φ is a linear function of the angle between θ, w_j, each w_j exerts a constant gradient on $\widetilde{\theta}$ towards $\widetilde{w_j}$, with discontinuities at $\widetilde{w_j}, \pi + \widetilde{w_j}$. Almost surely over $b_1, .., b_k$, the gradient is non-zero almost everywhere, except at the discontinuities, which are at $\widetilde{w_j}, \pi + \widetilde{w_j}$ for some j. ◀

F.2 Convergence of Polynomial Potentials

▶ **Theorem 30.** *Let $\mathcal{M} = S^{d-1}$. Let $w_1, ..., w_k$ be orthonormal vectors in $\mathbb{R}^d$ and Φ is of the form $\Phi(\theta, w) = (\theta^T w)^l$ for some fixed integer $l \geq 3$. Let L be as in* (3)*. Then, all critical points of L are not local minima, except when $\theta_1 = w_j$ for some j.*

Proof. WLOG, we can consider $w_1, ..., w_d$ to be the basis vectors $e_1, ..., e_d$. Note that this is a manifold optimization problem, so our optimality conditions are given by introducing a

Lagrange multiplier λ, as in [12].

$$\frac{\partial L}{\partial a} = 2\sum_{i=1}^{d} ab_i(\theta_i)^l + 2a = 0$$

$$(\nabla_\theta L)_i = 2ab_i l(\theta_i)^{l-1} - 2\lambda\theta_i = 0$$

where λ is chosen that minimizes

$$\lambda = \arg\min_{\lambda} \sum_i (ab_i l(\theta_i)^{l-1} - \lambda\theta_i)^2 = \sum ab_i l(\theta_i)^l$$

Therefore, either $\theta_i = 0$ or $b_i(\theta_i)^{l-2} = \lambda/(al)$. From [12], we consider the constrained Hessian, which is a diagonal matrix with diagonal entry:

$$(\nabla^2 L)_{ii} = 2ab_i l(l-1)(\theta_i)^{l-2} - 2\lambda$$

Assume that there exists $\theta_i, \theta_j \neq 0$, then we claim that θ is not a local minima. First, our optimality conditions imply $b_i(\theta_i)^{l-2} = b_j(\theta_j)^{l-2} = \lambda/(al)$. So,

$$(\nabla^2 L)_{ii} = (\nabla^2 L)_{jj} = 2ab_i l(l-1)(\theta_i)^{l-2} - 2\lambda$$

$$= 2(l-2)\lambda = -2(l-2)la^2$$

Now, there must exist a vector $v \in S^{d-1}$ such that $v_k = 0$ for $k \neq i, j$ and $v^T\theta = 0$, so v is in the tangent space at θ. Finally, $v^T(\nabla^2 L)v = -2(l-2)la^2 < 0$, implying θ is not a local minima when $a \neq 0$. Note that $a = 0$ occurs with probability 0 since our objective function is non-increasing throughout the gradient descent algorithm and is almost surely initialized to be negative with a optimized upon initialization, as by observed before. ◀

Under a node-wise descent algorithm, we can show polynomial-time convergence to global minima under orthogonality assumptions on w_j for these polynomial activations/potentials. We will not include the proof but it follows from similar techniques presented for nodewise convergence in Section E.