Globally Optimal Training of Generalized Polynomial Neural Networks with Nonlinear Spectral Methods

A. Gautier, Q. Nguyen and M. Hein
Department of Mathematics and Computer Science
Saarland Informatics Campus, Saarland University, Germany

Abstract

The optimization problem behind neural networks is highly non-convex. Training with stochastic gradient descent and variants requires careful parameter tuning and provides no guarantee to achieve the global optimum. In contrast we show under quite weak assumptions on the data that a particular class of feedforward neural networks can be trained globally optimal with a linear convergence rate with our nonlinear spectral method. Up to our knowledge this is the first practically feasible method which achieves such a guarantee. While the method can in principle be applied to deep networks, we restrict ourselves for simplicity in this paper to one and two hidden layer networks. Our experiments confirm that these models are rich enough to achieve good performance on a series of real-world datasets.

1   Introduction

Deep learning [13, 16] is currently the state of the art machine learning technique in many application areas such as computer vision or natural language processing. While the theoretical foundations of neural networks have been explored in depth see e.g. [1], the understanding of the success of training deep neural networks is a currently very active research area [5, 6, 9]. On the other hand the parameter search for stochastic gradient descent and variants such as Adagrad and Adam can be quite tedious and there is no guarantee that one converges to the global optimum. In particular, the problem is even for a single hidden layer in general NP hard, see [17] and references therein. This implies that to achieve global optimality efficiently one has to impose certain conditions on the problem.

A recent line of research has directly tackled the optimization problem of neural networks and provided either certain guarantees [2, 15] in terms of the global optimum or proved directly convergence to the global optimum [8, 11]. The latter two papers are up to our knowledge the first results which provide a globally optimal algorithm for training neural networks. While providing a lot of interesting insights on the relationship of structured matrix factorization and training of neural networks, Haeffele and Vidal admit themselves in their paper [8] that their results are "challenging to apply in practice". In the work of Janzamin et al. [11] they use a tensor approach and propose a globally optimal algorithm for a feedforward neural network with one hidden layer and squared loss. However, their approach requires the computation of the score function tensor which uses the density of the data-generating measure. However, the data generating measure is unknown and also difficult to estimate for high-dimensional feature spaces. Moreover, one has to check certain non-degeneracy conditions of the tensor decomposition to get the global optimality guarantee.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

In contrast our nonlinear spectral method just requires that the data is nonnegative which is true for all sorts of count data such as images, word frequencies etc. The condition which guarantees global optimality just depends on the parameters of the architecture of the network and boils down to the computation of the spectral radius of a small nonnegative matrix. The condition can be checked without running the algorithm. Moreover, the nonlinear spectral method has a linear convergence rate and thus the globally optimal training of the network is very fast. The two main changes compared to the standard setting are that we require nonnegativity on the weights of the network and we have to minimize a modified objective function which is the sum of loss and the negative total sum of the outputs. While this model is non-standard, we show in some first experimental results that the resulting classifier is still expressive enough to create complex decision boundaries. As well, we achieve competitive performance on some UCI datasets. As the nonlinear spectral method requires some non-standard techniques, we use the main part of the paper to develop the key steps necessary for the proof. However, some proofs of the intermediate results are moved to the supplementary material.

2 Main result
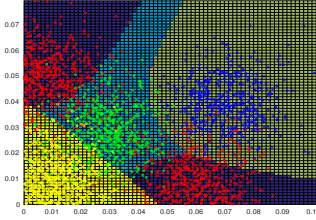
Figure 1: Classification decision boundaries in $\mathbb{R}^2$. (Best viewed in colors.)

In this section we present the algorithm together with the main theorem providing the convergence guarantee. We limit the presentation to one hidden layer networks to improve the readability of the paper. Our approach can be generalized to feedforward networks of *arbitrary* depth. In particular, we present in Section 4.1 results for two hidden layers.

We consider in this paper multi-class classification where d is the dimension of the feature space and K is the number of classes. We use the negative cross-entropy loss defined for label $y \in [K] := \{1, \ldots, K\}$ and classifier $f : \mathbb{R}^d \to \mathbb{R}^K$ as

$$L\big(y, f(x)\big) = -\log\left(\frac{e^{f_y(x)}}{\sum_{j=1}^K e^{f_j(x)}}\right) = -f_y(x) + \log\Big(\sum_{j=1}^K e^{f_j(x)}\Big).$$

The function class we are using is a feedforward neural network with one hidden layer with n_1 hidden units. As activation functions we use real powers of the form of a generalized polyomial, that is for $\alpha \in \mathbb{R}^{n_1}$ with $\alpha_i \geq 1$, $i \in [K]$, we define:

$$f_r(x) = f_r(w, u)(x) = \sum_{l=1}^{n_1} w_{rl}\Big(\sum_{m=1}^d u_{lm}x_m\Big)^{\alpha_l}, \tag{1}$$

where $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$ and $w \in \mathbb{R}_+^{K \times n_1}$, $u \in \mathbb{R}_+^{n_1 \times d}$ are the parameters of the network which we optimize. The function class in (1) can be seen as a generalized polynomial in the sense that the powers do not have to be integers. Polynomial neural networks have been recently analyzed in [15]. Please note that a ReLU activation function makes no sense in our setting as we require the data as well as the weights to be nonnegative. Even though nonnegativity of the weights is a strong constraint, one can model quite complex decision boundaries (see Figure 1, where we show the outcome of our method for a toy dataset in $\mathbb{R}^2$).

In order to simplify the notation we use $w = (w_1, \ldots, w_K)$ for the K output units $w_i \in \mathbb{R}_+^{n_1}$, $i = 1, \ldots, K$. All output units and the hidden layer are normalized. We optimize over the set

$$S_+ = \big\{(w, u) \in \mathbb{R}_+^{K \times n_1} \times \mathbb{R}_+^{n_1 \times d} \;\big|\; \|u\|_{p_u} = \rho_u, \; \|w_i\|_{p_w} = \rho_w, \; \forall i = 1, \ldots, K\big\}.$$

We also introduce S_{++} where one replaces $\mathbb{R}_+$ with $\mathbb{R}_{++} = \{t \in \mathbb{R} \mid t > 0\}$. The final optimization problem we are going to solve is given as

$$\max_{(w,u) \in S_+} \Phi(w, u) \quad \text{with} \tag{2}$$

$$\Phi(w, u) = \frac{1}{n}\sum_{i=1}^n \Big[-L\Big(y_i, f(w, u)(x^i)\Big) + \sum_{r=1}^K f_r(w, u)(x^i)\Big] + \epsilon\Big(\sum_{r=1}^K \sum_{l=1}^{n_1} w_{r,l} + \sum_{l=1}^{n_1} \sum_{m=1}^d u_{lm}\Big),$$

2

where $(x^i, y_i) \in \mathbb{R}^d_+ \times [K]$, $i = 1, \ldots, n$ is the training data. Note that this is a maximization problem and thus we use minus the loss in the objective so that we are effectively minimizing the loss. The reason to write this as a maximization problem is that our nonlinear spectral method is inspired by the theory of (sub)-homogeneous nonlinear eigenproblems on convex cones [14] which has its origin in the Perron-Frobenius theory for nonnegative matrices. In fact our work is motivated by the closely related Perron-Frobenius theory for multi-homogeneous problems developed in [7]. This is also the reason why we have nonnegative weights, as we work on the positive orthant which is a convex cone. Note that $\epsilon > 0$ in the objective can be chosen arbitrarily small and is added out of technical reasons.

In order to state our main theorem we need some additional notation. For $p \in (1, \infty)$, we let $p' = p/(p-1)$ be the Hölder conjugate of p, and $\psi_p(x) = \text{sign}(x)|x|^{p-1}$. We apply ψ_p to scalars and vectors in which case the function is applied componentwise. For a square matrix A we denote its spectral radius by $\rho(A)$. Finally, we write $\nabla_{w_i}\Phi(w, u)$ (resp. $\nabla_u \Phi(w, u)$) to denote the gradient of Φ with respect to w_i (resp. u) at (w, u). The mapping

$$G^\Phi(w, u) = \left(\frac{\rho_w \psi_{p'_w}(\nabla_{w_1}\Phi(w, u))}{\|\psi_{p'_w}(\nabla_{w_1}\Phi(w, u))\|_{p_w}}, \ldots, \frac{\rho_w \psi_{p'_w}(\nabla_{w_K}\Phi(w, u))}{\|\psi_{p'_w}(\nabla_{w_K}\Phi(w, u))\|_{p_w}}, \frac{\rho_u \psi_{p'_u}(\nabla_u \Phi(w, u))}{\|\psi_{p'_u}(\nabla_u \Phi(w, u))\|_{p_u}} \right),$$
(3)

defines a sequence converging to the global optimum of (2). Indeed, we prove:

Theorem 1. *Let $\{x^i, y_i\}_{i=1}^n \subset \mathbb{R}^d_+ \times [K]$, $p_w, p_u \in (1, \infty)$, $\rho_w, \rho_u > 0$, $n_1 \in \mathbb{N}$ and $\alpha \in \mathbb{R}^{n_1}$ with $\alpha_i \geq 1$ for every $i \in [n_1]$. Define $\rho_x, \xi_1, \xi_2 > 0$ as $\rho_x = \max_{i \in [n]} \|x^i\|_1$, $\xi_1 = \rho_w \sum_{l=1}^{n_1}(\rho_u \rho_x)^{\alpha_l}$, $\xi_2 = \rho_w \sum_{l=1}^{n_1} \alpha_l (\rho_u \rho_x)^{\alpha_l}$ and let $A \in \mathbb{R}^{(K+1) \times (K+1)}_{++}$ be defined as*

$$\begin{aligned} A_{l,m} &= 4(p'_w - 1)\xi_1, & A_{l,K+1} &= 2(p'_w - 1)(2\xi_2 + \|\alpha\|_\infty), & \forall m, l \in [K]. \\ A_{K+1,m} &= 2(p'_u - 1)(2\xi_1 + 1), & A_{K+1,K+1} &= 2(p'_u - 1)(2\xi_2 + \|\alpha\|_\infty - 1), \end{aligned}$$

If the spectral radius $\rho(A)$ of A satisfies $\rho(A) < 1$, then (2) has a unique global maximizer $(w^*, u^*) \in S_{++}$. Moreover, for every $(w^0, u^0) \in S_{++}$, there exists $R > 0$ such that

$$\lim_{k \to \infty}(w^k, u^k) = (w^*, u^*) \qquad and \qquad \|(w^k, u^k) - (w^*, u^*)\|_\infty \leq R\,\rho(A)^k \qquad \forall k \in \mathbb{N},$$

where $(w^{k+1}, u^{k+1}) = G^\Phi(w^k, u^k)$ for every $k \in \mathbb{N}$.

Note that one can check for a given model (number of hidden units n_1, choice of α, p_w, p_u, ρ_u, ρ_w) easily if the convergence guarantee to the global optimum holds by computing the spectral radius of a square matrix of size $K + 1$. As our bounds for the matrix A are very conservative, the "effective" spectral radius is typically much smaller, so that we have very fast convergence in only a few iterations, see Section 5 for a discussion. Up to our knowledge this is the first practically feasible algorithm to achieve global optimality for a non-trivial neural network model. Additionally, compared to stochastic gradient descent, there is no free parameter in the algorithm. Thus no careful tuning of the learning rate is required. The reader might wonder why we add the second term in the objective, where we sum over all outputs. The reason is that we need that the gradient of G^Φ is strictly positive in S_+, this is why we also have to add the third term for arbitrarily small $\epsilon > 0$. In Section 5 we show that this model achieves competitive results on a few UCI datasets.

Choice of α: It turns out that in order to get a non-trivial classifier one has to choose $\alpha_1, \ldots, \alpha_{n_1} \geq 1$ so that $\alpha_i \neq \alpha_j$ for every $i, j \in [n_1]$ with $i \neq j$. The reason for this lies in certain invariance properties of the network. Suppose that we use a permutation invariant componentwise activation function σ, that is $\sigma(Px) = P\sigma(x)$ for any permutation matrix P and suppose that A, B are globally optimal weight matrices for a one hidden layer architecture, then for any permutation matrix P,

$$A\sigma(Bx) = AP^T P\sigma(Bx) = AP^T \sigma(PBx),$$

which implies that $A' = AP^T$ and $B' = PB$ yield the same function and thus are also globally optimal. In our setting we know that the global optimum is *unique* and thus it has to hold that, $A = AP^T$ and $B = PB$ for all permutation matrices P. This implies that both A and B have rank one and thus lead to trivial classifiers. This is the reason why one has to use different α for every unit.

3

Dependence of $\rho(A)$ on the model parameters: Let $Q, \tilde{Q} \in \mathbb{R}_+^{m \times m}$ and assume $0 \le Q_{i,j} \le \tilde{Q}_{i,j}$ for every $i, j \in [m]$, then $\rho(Q) \le \rho(\tilde{Q})$, see Corollary 3.30 [3]. It follows that $\rho(A)$ in Theorem 1 is increasing w.r.t. ρ_u, ρ_w, ρ_x and the number of hidden units n_1. Moreover, $\rho(A)$ is decreasing w.r.t. p_u, p_w and in particular, we note that for any fixed architecture $(n_1, \alpha, \rho_u, \rho_w)$ it is always possible to find p_u, p_w large enough so that $\rho(A) < 1$. Indeed, we know from the Collatz-Wielandt formula (Theorem 8.1.26 in [10]) that $\rho(A) = \rho(A^T) \le \max_{i \in [K+1]} (A^T v)_i / v_i$ for any $v \in \mathbb{R}_{++}^{K+1}$. We use this to derive lower bounds on p_u, p_w that ensure $\rho(A) < 1$. Let $v = (p_w - 1, \dots, p_w - 1, p_u - 1)$, then $(A^T v)_i < v_i$ for every $i \in [K+1]$ guarantees $\rho(A) < 1$ and is equivalent to

$$p_w > 4(K+1)\xi_1 + 3 \qquad \text{and} \qquad p_u > 2(K+1)(\|\alpha\|_\infty + 2\xi_2) - 1, \qquad (4)$$

where ξ_1, ξ_2 are defined as in Theorem 1. However, we think that our current bounds are sub-optimal so that this choice is quite conservative. Finally, we note that the constant R in Theorem 1 can be explicitly computed when running the algorithm (see Theorem 3).

Proof Strategy: The following main part of the paper is devoted to the proof of the algorithm. For that we need some further notation. We introduce the sets

$$V_+ = \mathbb{R}_+^{K \times n_1} \times \mathbb{R}_+^{n_1 \times d}, \quad V_{++} = \mathbb{R}_{++}^{K \times n_1} \times \mathbb{R}_{++}^{n_1 \times d}$$
$$B_+ = \left\{ (w, u) \in V_+ \mid \|u\|_{p_u} \le \rho_u, \ \|w_i\|_{p_w} \le \rho_w, \ \forall i = 1, \dots, K \right\},$$

and similarly we define B_{++} replacing V_+ by V_{++} in the definition. The high-level idea of the proof is that we first show that the global maximum of our optimization problem in (2) is attained in the "interior" of S_+, that is S_{++}. Moreover, we prove that any critical point of (2) in S_{++} is a fixed point of the mapping G^Φ. Then we proceed to show that there exists a unique fixed point of G^Φ in S_{++} and thus there is a unique critical point of (2) in S_{++}. As the global maximizer of (2) exists and is attained in the interior, this fixed point has to be the global maximizer.

Finally, the proof of the fact that G^Φ has a unique fixed point follows by noting that G^Φ maps B_{++} into B_{++} and the fact that B_{++} is a complete metric space with respect to the Thompson metric. We provide a characterization of the Lipschitz constant of G^Φ and in turn derive conditions under which G^Φ is a contraction. Finally, the application of the Banach fixed point theorem yields the uniqueness of the fixed point of G^Φ and the linear convergence rate to the global optimum of (2). In Section 4 we show the application of the established framework for our neural networks.

3 From the optimization problem to fixed point theory

Lemma 1. *Let $\Phi : V \to \mathbb{R}$ be differentiable. If $\nabla \Phi(w, u) \in V_{++}$ for every $(w, u) \in S_+$, then the global maximum of Φ on S_+ is attained in S_{++}.*

Proof. First note that as Φ is a continuous function on the compact set S_+ the global minimum and maximum are attained. A boundary point (w^b, u^b) of S_+ is characterized by the fact that at least one of the variables (w^b, u^b) has a zero component. Suppose w.l.o.g. that the subset $J \subset [n_1]$ of components of $w_1^b \in \mathbb{R}_+^{n_1}$ are zero, that is $w_{1,J}^b = 0$. The normal vector of the p_w-sphere at w_1^b is given by $\nu = \psi_{p_w}(w_1^b)$. The set of tangent directions is thus given by

$$T = \{v \in \mathbb{R}^{n_1} \mid \langle \nu, v \rangle = 0\}.$$

Note that if (w^b, u^b) is a local maximum, then

$$\left\langle \nabla_{w_1^b} \Phi(w^b, u^b), t \right\rangle \le 0, \quad \forall t \in T_+ = \{v \in \mathbb{R}_+^{n_1} \mid \langle \nu, v \rangle = 0\}, \qquad (5)$$

where T_+ is the set of "positive" tangent directions, that are pointing inside the set $\{w_1 \in \mathbb{R}_+^{n_1} \mid \|w_1\|_{p_w} = \rho_w\}$. Otherwise there would exist a direction of ascent which leads to a feasible point. Now note that ν has non-negative components as $w_1^b \in \mathbb{R}_+^{n_1}$. Thus

$$T_+ = \{v \in \mathbb{R}_+^{n_1} \mid v_i = 0 \text{ if } i \notin J\}.$$

4

However, by assumption $\nabla_{w_1^b}\Phi(w^b, u^b)$ is a vector with strictly positive components and thus (5) can never be fulfilled as T_+ contains only vectors with non-negative components and at least one of the components is strictly positive as $J \neq [n_1]$. Finally, as the global maximum is attained in S_+ and no local maximum exists at the boundary, the global maximum has to be attained in S_{++}. □

We now identify critical points of the objective Φ in S_{++} with fixed points of G^Φ in S_{++}.

Lemma 2. *Let $\Phi : V \to \mathbb{R}$ be differentiable. If $\nabla\Phi(w, u) \in V_{++}$ for all $(w, u) \in S_{++}$, then (w^*, u^*) is a critical point of Φ in S_{++} if and only if it is a fixed point of G^Φ.*

Proof. The Lagrangian of Φ constrained to the unit sphere S is given by

$$\mathcal{L}(w, u, \lambda) = \Phi(w, u) - \lambda_{K+1}(\|u\|_{p_u} - \rho_u) - \sum_{j=1}^{K} \lambda_i(\|w_j\|_{p_w} - \rho_w).$$

A necessary and sufficient condition [4] for $(w, u) \in S_{++}$ being a critical point of Φ is the existence of λ_i with

$$\nabla_{w_j}\Phi(w, u) = \lambda_j \psi_{p_w}(w_j) \quad \forall j \in [K] \qquad \text{and} \qquad \nabla_u\Phi(w, u) = \lambda_{K+1}\psi_{p_u}(u). \qquad (6)$$

Note that as $w_j \in \mathbb{R}_{++}^{n_1}$ and the gradients are strictly positive in S_{++} the λ_i, $i = 1, \dots, K+1$ have to be strictly positive. Noting that $\psi_{p'}(\psi_p(x)) = x$, we get

$$\psi_{p_w'}\big(\nabla_{w_j}\Phi(w, u)\big) = \lambda_j^{p_w'-1} w_j \quad \forall j \in [K] \qquad \text{and} \qquad \psi_{p_u'}\big(\nabla_u\Phi(w, u)\big) = \lambda_{K+1}^{p_u'-1} u. \qquad (7)$$

In particular, (w^*, u^*) is a critical point of Φ in S_{++} if and only if it satisfies (7). Finally, note that

$$\left(\frac{\rho_w \psi_{p_w'}\big(\nabla_{w_1}\Phi(w, u)\big)}{\|\psi_{p_w'}\big(\nabla_{w_1}\Phi(w, u)\big)\|_{p_w}}, \dots, \frac{\rho_w \psi_{p_w'}\big(\nabla_{w_K}\Phi(w, u)\big)}{\|\psi_{p_w'}\big(\nabla_{w_K}\Phi(w, u)\big)\|_{p_w}}, \frac{\rho_u \psi_{p_u'}\big(\nabla_u\Phi(w, u)\big)}{\|\psi_{p_u'}\big(\nabla_u\Phi(w, u)\big)\|_{p_u}} \right) \in S_{++},$$

as the gradient is strictly positive on S_{++} and thus $G^\Phi : S_{++} \to S_{++}$ defined in (3) is well-defined and if (w^*, u^*) is a critical point, then by (7) it holds $G^\Phi(w^*, u^*) = (w^*, u^*)$. On the other hand if $G^\Phi(w^*, u^*) = (w^*, u^*)$, then

$$\rho_w \frac{\psi_{p_w'}\big(\nabla_{w_j}\Phi(w^*, u^*)\big)}{\|\psi_{p_w'}\big(\nabla_{w_j}\Phi(w^*, u^*)\big)\|_{p_w}} = w_j^*, \quad j = 1, \dots, K, \qquad \rho_u \frac{\psi_{p_u'}\big(\nabla_u\Phi(w^*, u^*)\big)}{\|\psi_{p_u'}\big(\nabla_u\Phi(w^*, u^*)\big)\|_{p_u}} = u^*$$

and thus there exists $\lambda_j = \frac{\|\psi_{p_w'}(\nabla_{w_j}\Phi(w^*, u^*))\|_{p_w}}{\rho_w}$, $j \in [K]$ and $\lambda_{K+1} = \frac{\|\psi_{p_u'}(\nabla_u\Phi(w^*, u^*))\|_{p_u}}{\rho_u}$ such that (7) holds and thus (w^*, u^*) is a critical point of Φ in S_{++}. □

Our goal is to apply the Banach fixed point theorem to $G^\Phi : B_{++} \to S_{++} \subset B_{++}$. We recall this theorem for the convenience of the reader.

Theorem 2 (Banach fixed point theorem e.g. [12])**.** *Let (X, d) be a complete metric space with a mapping $T : X \to X$ such that $d(T(x), T(y)) \leq q\, d(x, y)$ for $q \in [0, 1)$ and all $x, y \in X$. Then T has a unique fixed-point x^* in X, that is $T(x^*) = x^*$ and the sequence defined as $x^{n+1} = T(x^n)$ with $x^0 \in X$ converges $\lim_{n \to \infty} x^n = x^*$ with linear convergence rate*

$$d(x^n, x^*) \leq \frac{q^n}{1-q}\, d(x^1, x^0).$$

So, we need to endow B_{++} with a metric μ so that (B_{++}, μ) is a complete metric space. A popular metric for the study of nonlinear eigenvalue problems on the positive orthant is the so-called Thompson metric $d \colon \mathbb{R}_{++}^m \times \mathbb{R}_{++}^m \to \mathbb{R}_+$ [18] defined as

$$d(z, \tilde{z}) = \|\ln(z) - \ln(\tilde{z})\|_\infty \qquad \text{where} \qquad \ln(z) = \big(\ln(z_1), \dots, \ln(z_m)\big).$$

Using the known facts that $(\mathbb{R}_{++}^n, d)$ is a complete metric space and its topology coincides with the norm topology (see e.g. Corollary 2.5.6 and Proposition 2.5.2 [14]), we prove:

5

Lemma 3. *For $p \in (1, \infty)$ and $\rho > 0$, $(\{z \in \mathbb{R}^n_{++} \mid \|z\|_p \leq \rho\}, d)$ is a complete metric space.*

Proof. Let $(z^k)_k \subset \{z \in \mathbb{R}^n_{++} \mid \|z\|_p \leq \rho\}$ be a Cauchy sequence w.r.t. to the metric d. We know from Proposition 2.5.2 in [14] that $(\mathbb{R}^n_{++}, d)$ is a complete metric space and thus there exists $z^* \in \mathbb{R}^n_{++}$ such that z^k converge to z^* w.r.t. d. Corollary 2.5.6 in [14] implies that the topology of $(\mathbb{R}^n_{++}, d)$ coincide with the norm topology implying that $\lim_{k\to\infty} z^k = z^*$ w.r.t. the norm topology. Finally, since $z \mapsto \|z\|_p$ is a continuous function, we get $\|z^*\|_p = \lim_{k\to\infty}\|z^k\|_p \leq \rho$, i.e. $z^* \in \{z \in \mathbb{R}^n_{++} \mid \|z\|_p \leq 1\}$ which proves our claim. $\qquad\square$

Now, the idea is to see B_{++} as a product of such metric spaces. For $i = 1, \ldots, K$, let $B^i_{++} = \{w_i \in \mathbb{R}^{n_1}_{++} \mid \|w_i\|_{p_w} \leq \rho_w\}$ and $d_i(w_i, \tilde{w}_i) = \gamma_i\|\ln(w_i) - \ln(\tilde{w}_i)\|_\infty$ for some constant $\gamma_i > 0$. Furthermore, let $B^{K+1}_{++} = \{u \in \mathbb{R}^{n_1 \times d}_{++} \mid \|u\|_{p_u} \leq \rho_u\}$ and $d_{K+1}(u, \tilde{u}) = \gamma_{K+1}\|\ln(u) - \ln(\tilde{u})\|_\infty$. Then (B^i_{++}, d_i) is a complete metric space for every $i \in [K+1]$ and $B_{++} = B^1_{++} \times \ldots \times B^K_{++} \times B^{K+1}_{++}$. It follows that (B_{++}, μ) is a complete metric space with $\mu\colon B_{++} \times B_{++} \to \mathbb{R}_+$ defined as

$$\mu\big((w, u), (\tilde{w}, \tilde{u})\big) = \sum_{i=1}^{K} \gamma_i\|\ln(w_i) - \ln(\tilde{w}_i)\|_\infty + \gamma_{K+1}\|\ln(u) - \ln(\tilde{u})\|_\infty.$$

The motivation for introducing the weights $\gamma_1, \ldots, \gamma_{K+1} > 0$ is given by the next theorem. We provide a characterization of the Lipschitz constant of a mapping $F\colon B_{++} \to B_{++}$ with respect to μ. Moreover, this Lipschitz constant can be minimized by a smart choice of γ. For $i \in [K], a, j \in [n_1], b \in [d]$, we write $F_{w_{i,j}}$ and $F_{u_{ab}}$ to denote the components of F such that $F = (F_{w_{1,1}}, \ldots, F_{w_{1,n_1}}, F_{w_{2,1}}, \ldots, F_{w_{K,n_1}}, F_{u_{11}}, \ldots, F_{u_{n_1 d}})$.

Lemma 4. *Suppose that $F \in C^1(B_{++}, V_{++})$ and $A \in \mathbb{R}^{(K+1)\times(K+1)}_+$ satisfies*

$$\big\langle |\nabla_{w_k} F_{w_{i,j}}(w, u)|, w_k \big\rangle \leq A_{i,k} F_{w_{i,j}}(w, u), \quad \big\langle |\nabla_u F_{w_{i,j}}(w, u)|, u \big\rangle \leq A_{i,K+1} F_{w_{i,j}}(w, u)$$

and

$$\big\langle |\nabla_{w_k} F_{u_{ab}}(w, u)|, w_k \big\rangle \leq A_{K+1,k} F_{u_{ab}}(w, u), \quad \big\langle |\nabla_u F_{u_{ab}}(w, u)|, u \big\rangle \leq A_{K+1,K+1} F_{u_{ab}}(w, u)$$

for all $i, k \in [K]$, $a, j \in [n_1]$, $b \in [d]$ and $(w, u) \in B_{++}$. Then, for every $(w, u), (\tilde{w}, \tilde{u}) \in B_{++}$ it holds

$$\mu\big(F(w, u), F(\tilde{w}, \tilde{u})\big) \leq U \, \mu\big((w, u), (\tilde{w}, \tilde{u})\big) \qquad \text{with} \qquad U = \max_{k \in [K+1]} \frac{(A^T\gamma)_k}{\gamma_k}.$$

Proof. Let $i \in [K]$ and $j \in [n_1]$. First, we show that for $(w, u), (\tilde{w}, \tilde{u}) \in B_{++}$, one has

$$\left| \ln\big(F_{w_{i,j}}(w, u)\big) - \ln\big(F_{w_{i,j}}(\tilde{w}, \tilde{u})\big) \right| \leq \sum_{s=1}^{K} A_{i,s}\|\ln(w_s) - \ln(\tilde{w}_s)\|_\infty + A_{i,K+1}\|\ln(u) - \ln(\tilde{u})\|_\infty.$$

Let $(w, u), (\tilde{w}, \tilde{u}) \in B_{++}$. If $(w, u) = (\tilde{w}, \tilde{u})$, there is nothing to prove. So, suppose that $(w, u) \neq (\tilde{w}, \tilde{u})$. Let $\hat{B} = \{\ln(v) \mid v \in B_{++}\} \subset V$ and $g\colon \hat{B} \to \mathbb{R}$ with

$$g(\overline{w}, \overline{u}) = \ln\left(F_{w_{i,j}}\big(\exp(\overline{w}, \overline{u})\big)\right) \quad \text{where} \quad \exp(\overline{w}, \overline{u}) = \big(e^{\overline{w}_{1,1}}, \ldots, e^{\overline{w}_{K,n_1}}, e^{\overline{u}_{11}}, \ldots, e^{\overline{u}_{n_1 d}}\big).$$

Note that g is continuously differentiable because it is the composition of continuously differentiable mappings. Set $(v, x) = \big(\ln(w), \ln(u)\big)$ and $(y, z) = \big(\ln(\tilde{w}), \ln(\tilde{u})\big)$. By the mean value theorem, there exists $t \in (0, 1)$, such that $(\hat{w}, \hat{u}) = t(v, x) + (1 - t)(y, z)$ satisfies

$$g(v, x) - g(y, z) = \langle \nabla g(\hat{w}, \hat{u}), (v, x) - (y, z) \rangle.$$

Let $(\overline{w}, \overline{u}) = \exp(\hat{w}, \hat{u})$. It follows that

$$\ln\big(F_{w_{i,j}}(w, u)\big) - \ln\big(F_{w_{i,j}}(\tilde{w}, \tilde{u})\big) = \langle \nabla g(\hat{w}, \hat{u}), (v, x) - (y, z) \rangle$$

$$= \frac{\big\langle \big(\nabla F_{w_{i,j}}(\overline{w}, \overline{u})\big) \circ (\overline{w}, \overline{u}), (v, x) - (y, z)\big\rangle}{F_{w_{i,j}}(\overline{w}, \overline{u})}$$

6

where ∘ denotes the Hadamard product. Note that by convexity of the exponential and $\|\cdot\|_p$ we have $(\overline{w}, \overline{u}) \in B_{++}$ as

$$\|\overline{w}\|_{p_w} = \|e^{tv+(1-t)y}\|_{p_w} \le \|te^v + (1-t)e^y\|_{p_w} \le t\|e^v\|_{p_w} + (1-t)\|e^y\|_{p_w}$$
$$= t\|w\|_{p_w} + (1-t)\|\tilde{w}\|_{p_w} \le \rho_w,$$

where we have used in the second step convexity of the exponential and that the p-norms are order preserving, $\|x\|_p \le \|y\|_p$ if $x_i \le z_i$ for all i. A similar argument shows $\|\overline{u}\|_{p_u} \le \rho_u$. Hence

$$\left| \ln\left(F_{w_{i,j}}(w,u)\right) - \ln\left(F_{w_{i,j}}(\tilde{w}, \tilde{u})\right) \right|$$

$$\le \sum_{k=1}^{K} \frac{\left|\left\langle \left(\nabla_{w_k} F_{w_{i,j}}(\overline{w}, \overline{u})\right) \circ \overline{w}_k, v_k - y_k \right\rangle\right|}{F_{w_{i,j}}(\overline{w}, \overline{u})} + \frac{\left|\left\langle \left(\nabla_u F_{w_{i,j}}(\overline{w}, \overline{u})\right) \circ \overline{u}, x - z \right\rangle\right|}{F_{w_{i,j}}(\overline{w}, \overline{u})}$$

$$\le \sum_{k=1}^{K} \frac{\left\|\left(\nabla_{w_k} F_{w_{i,j}}(\overline{w}, \overline{u})\right) \circ \overline{w}_k\right\|_1}{F_{w_{i,j}}(\overline{w}, \overline{u})} \|v_k - y_k\|_\infty + \frac{\left\|\left(\nabla_u F_{w_{i,j}}(\overline{w}, \overline{u})\right) \circ \overline{u}\right\|_1}{F_{w_{i,j}}(\overline{w}, \overline{u})} \|x - z\|_\infty$$

$$= \sum_{k=1}^{K} \frac{\left\langle \left|\nabla_{w_k} F_{w_{i,j}}(\overline{w}, \overline{u})\right|, \overline{w}_k \right\rangle}{F_{w_{i,j}}(\overline{w}, \overline{u})} \|\ln(w_k) - \ln(\tilde{w}_k)\|_\infty + \frac{\left\langle \left|\nabla_u F_{w_{i,j}}(\overline{w}, \overline{u})\right|, \overline{u} \right\rangle}{F_{w_{i,j}}(\overline{w}, \overline{u})} \|\ln(u) - \ln(\tilde{u})\|_\infty$$

$$\le \sum_{k=1}^{K} A_{i,k} \|\ln(w_k) - \ln(\tilde{w}_k)\|_\infty + A_{i,K+1} \|\ln(u) - \ln(\tilde{u})\|_\infty.$$

In particular, taking the maximum over $j \in [n_1]$ shows that for every $i \in [K]$ we have

$$\|\ln\left(F_{w_i}(w,u)\right) - \ln\left(F_{w_i}(\tilde{w}, \tilde{u})\right)\|_\infty \le \sum_{s=1}^{K} A_{i,s} \|\ln(w_s) - \ln(\tilde{w}_s)\|_\infty + A_{i,K+1} \|\ln(u) - \ln(\tilde{u})\|_\infty.$$

A similar argument shows that $\|\ln\left(F_u(w,u)\right) - \ln\left(F_u(\tilde{w}, \tilde{u})\right)\|_\infty$ is upper bounded by

$$\sum_{s=1}^{K} A_{K+1,s} \|\ln(w_s) - \ln(\tilde{w}_s)\|_\infty + A_{K+1,K+1} \|\ln(u) - \ln(\tilde{u})\|_\infty.$$

So, we finally get

$$\mu\left(F(w,u), F(\tilde{w}, \tilde{u})\right) = \sum_{i=1}^{K} \gamma_i \|\ln\left(F_{w_i}(w,u)\right) - \ln\left(F_{w_i}(\tilde{w}, \tilde{u})\right)\|_\infty$$

$$+ \gamma_{K+1} \|\ln\left(F_u(w,u)\right) - \ln\left(F_u(\tilde{w}, \tilde{u})\right)\|_\infty$$

$$\le \sum_{s=1}^{K} (A^T \gamma)_s \|\ln(w_s) - \ln(\tilde{w}_s)\|_\infty + (A^T \gamma)_{K+1} \|\ln(u) - \ln(\tilde{u})\|_\infty$$

$$\le U \mu\left((w,u), (\tilde{w}, \tilde{u})\right). \qquad \square$$

Note that, from the Collatz-Wielandt ratio for nonnegative matrices, we know that the constant U in Lemma 4 is lower bounded by the spectral radius $\rho(A)$ of A. Indeed, by Theorem 8.1.31 in [10], we know that if A^T has a positive eigenvector $\gamma \in \mathbb{R}_{++}^{K+1}$, then

$$\max_{i \in [K+1]} \frac{(A^T \gamma)_i}{\gamma_i} = \rho(A) = \min_{\tilde{\gamma} \in \mathbb{R}_{++}^{K+1}} \max_{i \in [K+1]} \frac{(A^T \tilde{\gamma})_i}{\tilde{\gamma}_i}. \tag{8}$$

Therefore, in order to obtain the minimal Lipschitz constant U in Lemma 4, we choose the weights of the metric μ to be the components of γ. A combination of Theorem 2, Lemma 4 and this observation implies the following result.

Theorem 3. *Let* $\Phi \in C^1(V, \mathbb{R}) \cap C^2(B_{++}, \mathbb{R})$ *with* $\nabla\Phi(S_+) \subset V_{++}$. *Let* $G^\Phi \colon B_{++} \to B_{++}$ *be defined as in* (3). *Suppose that there exists a matrix* $A \in \mathbb{R}_+^{(K+1)\times(K+1)}$ *such that* G^Φ *and* A *satisfies the assumptions of Lemma 4 and* A^T *has a positive eigenvector* $\gamma \in \mathbb{R}_{++}^{K+1}$. *If*

7

$\rho(A) < 1$, then Φ has a unique critical point (w^*, u^*) in S_{++} which is the global maximum of the optimization problem (2). Moreover, the sequence $((w^k, u^k))_k$ defined for any $(w^0, u^0) \in S_{++}$ as $(w^{k+1}, u^{k+1}) = G^\Phi(w^k, u^k)$, $k \in \mathbb{N}$, satisfies $\lim_{k\to\infty}(w^k, u^k) = (w^*, u^*)$ and

$$\|(w^k, u^k) - (w^*, u^*)\|_\infty \le \rho(A)^k \left(\frac{\mu((w^1, u^1), (w^0, u^0))}{(1 - \rho(A)) \min\left\{\frac{\gamma_{K+1}}{\rho_u}, \min_{t\in[K]}\frac{\gamma_t}{\rho_w}\right\}} \right) \qquad \forall k \in \mathbb{N},$$

where the weights in the definition of μ are the entries of γ.

Proof. As γ is a positive eigenvector of A^T, by (8), we know that $A^T\gamma = \rho(A)\gamma$. It follows from Lemma 4 that $\mu(G^\Phi(w, u), G^\Phi(\tilde{w}, \tilde{u})) < \rho(A)\,\mu((w, u), (\tilde{w}, \tilde{u}))$ for every $(w, u), (\tilde{w}, \tilde{u}) \in B_{++}$, i.e. G^Φ is a strict contraction on the complete metric space (B_{++}, μ) and by the Banach fixed point theorem 2 we know that G^Φ has a unique fixed point (w^*, u^*) in S_{++}. From Lemma 1 we know that Φ attains its global maximum in S_{++} and, by Lemma 2, this maximum is a fixed point of G^Φ. Hence, (w^*, u^*) is the unique global maximum of Φ in S_{++}. Finally, Theorem 2 implies that

$$\mu((w^k, u^k), (w^*, u^*)) \le \frac{\rho(A)^k}{1 - \rho(A)}\mu((w^1, u^1), (w^0, u^0)) \qquad \forall k \in \mathbb{N}.$$

The mean value theorem implies that for every $r \in \mathbb{R}$, we have

$$|e^s - e^t| \le |s - t| \max_{\xi \in (-\infty, r]} e^\xi = e^r|s - t| \qquad \forall s, t \in (-\infty, r].$$

In particular, we have

$$\ln(w_{a,b}^k), \ln(w_{a,b}^*) \in (-\infty, \ln(\rho_w)] \qquad \text{and} \qquad \ln(u_{st}^k), \ln(u_{st}^*) \in (-\infty, \ln(\rho_u)].$$

It follows that

$$\mu((w^k, u^k), (w^*, u^*)) = \sum_{t=1}^{K} \gamma_t \| \ln(w_t^k) - \ln(w_t^*)\|_\infty + \gamma_{K+1}\|\ln(u^k) - \ln(u^*)\|_\infty$$

$$\ge \sum_{t=1}^{K} \frac{\gamma_t}{\rho_w}\|w_t^k - w_t^*\|_\infty + \frac{\gamma_{K+1}}{\rho_u}\|u^k - u^*\|_\infty$$

$$\ge \max\left\{ \max_{t\in[K]} \frac{\gamma_t}{\rho_w}\|w_t^k - w_t^*\|_\infty, \frac{\gamma_{K+1}}{\rho_u}\|u^k - u^*\|_\infty \right\}$$

$$\ge \min\left\{ \frac{\gamma_{K+1}}{\rho_u}, \min_{t\in[K]} \frac{\gamma_t}{\rho_w} \right\}\|(w^k, u^k) - (w^*, u^*)\|_\infty$$

and thus

$$\|(w^k, u^k) - (w^*, u^*)\|_\infty \le \frac{\mu((w^k, u^k), (w^*, u^*))}{\min\left\{\frac{\gamma_{K+1}}{\rho_u}, \min_{t\in[K]}\frac{\gamma_t}{\rho_w}\right\}}$$

$$\le \rho(A)^k \left(\frac{\mu((w^1, u^1), (w^0, u^0))}{(1 - \rho(A)) \min\left\{\frac{\gamma_{K+1}}{\rho_u}, \min_{t\in[K]}\frac{\gamma_t}{\rho_w}\right\}} \right). \qquad \square$$

4   Application to Neural Networks

In the previous sections we have outlined the proof of our main result for a general objective function satisfying certain properties. The purpose of this section is to prove that the properties hold for our optimization problem for neural networks.

We recall our objective function from (2)

$$\Phi(w, u) = \frac{1}{n} \sum_{i=1}^{n} \left[-L(y_i, f(w, u)(x^i)) + \sum_{r=1}^{K} f_r(w, u)(x^i) \right] + \epsilon \left(\sum_{r=1}^{K} \sum_{l=1}^{n_1} w_{r,l} + \sum_{l=1}^{n_1} \sum_{m=1}^{d} u_{lm} \right)$$

8

and the function class we are considering from (1)

$$f_r(x) = f_r(w, u)(x) = \sum_{l=1}^{n_1} w_{r,l} \Big(\sum_{m=1}^{d} u_{lm} x_m \Big)^{\alpha_l},$$

The arbitrarily small ϵ in the objective is needed to make the gradient strictly positive on the boundary of V_+. We note that the assumption $\alpha_i \geq 1$ for every $i \in [n_1]$ is crucial in the following lemma in order to guarantee that $\nabla \Phi$ is well defined on S_+.

Lemma 5. *Let Φ be defined as in (2), then $\nabla \Phi(w, u)$ is strictly positive for any $(w, u) \in S_+$.*

Proof. One can compute

$$\frac{\partial \Phi(w, u)}{\partial w_{a,b}} = \frac{1}{n} \sum_{i=1}^{n} \Big(\delta_{y_i a} - \frac{e^{f_a(x^i)}}{\sum_{j=1}^{K} e^{f_j(x^i)}} + 1 \Big) \Big(\sum_{m=1}^{d} u_{bm} x_m^i \Big)^{\alpha_b} + \epsilon,$$

where $\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{else} . \end{cases}$ denotes the Kronecker delta, and

$$\frac{\partial \Phi(w, u)}{\partial u_{ab}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{K} \Big(\delta_{y_i r} - \frac{e^{f_r(x^i)}}{\sum_{j=1}^{K} e^{f_j(x^i)}} + 1 \Big) \Big(w_{r,a} \alpha_a \big(\sum_{m=1}^{d} u_{am} x_m^i \big)^{\alpha_a - 1} x_b^i \Big) + \epsilon.$$

Note that

$$\delta_{y_i a} - \frac{e^{f_a(x_i)}}{\sum_{j=1}^{K} e^{f_j(x)}} + 1 > 0,$$

and thus using $(w, u) \in S_+$, we deduce that even without the additional ϵ, the gradient would be strictly positive on S_{++} assuming that there exists $k \in [n]$ such that $x^k \neq 0$. However, at the boundary it can happen that the gradient is zero and thus an arbitrarily small ϵ is sufficent to guarantee that the gradient is strictly positive on S_{++}. $\qquad \square$

Next, we derive the matrix $A \in \mathbb{R}^{(K+1) \times (K+1)}$ in order to apply Theorem 3 to G^Φ with Φ defined in (2). As discussed in its proof, the matrix A given in the following theorem has a smaller spectral radius than that of Theorem 1. To express this matrix, we consider $\Psi_{p,q}^\alpha : \mathbb{R}_{++}^{n_1} \times \mathbb{R}_{++} \to \mathbb{R}_{++}$ defined for $p, q \in (1, \infty)$ and $\alpha \in \mathbb{R}_{++}^{n_1}$ as

$$\Psi_{p,q}^\alpha(\delta, t) = \Bigg(\Big[\sum_{l \in J} (\delta_l \, t^{\alpha_l})^{\frac{p \, q}{q - \bar{\alpha} p}} \Big]^{1 - \frac{\bar{\alpha} p}{q}} + \max_{j \in J^c} (\delta_j \, t^{\alpha_j})^p \Bigg)^{1/p}, \tag{9}$$

where $J = \{ l \in [n_1] \mid \alpha_l p \leq q \}$, $J^c = \{ l \in [n_1] \mid \alpha_l p > q \}$ and $\bar{\alpha} = \min_{l \in J} \alpha_l$.

Theorem 4. *Let Φ be defined as above and G^Φ be as in (3). Set $C_w = \rho_w \, \Psi_{p'_w, p_u}^\alpha(\mathbf{1}, \rho_u \rho_x)$, $C_u = \rho_w \, \Psi_{p'_w, p_u}^\alpha(\alpha, \rho_u \rho_x)$ and $\rho_x = \max_{i \in [n]} \|x^i\|_{p'_u}$. Then A and G^Φ satisfy all assumptions of Lemma 4 with*

$$A = 2 \, \mathrm{diag} \, \big(p'_w - 1, \dots, p'_w - 1, p'_u - 1 \big) \begin{pmatrix} Q_{w,w} & Q_{w,u} \\ Q_{u,w} & Q_{u,u} \end{pmatrix}$$

where $Q_{w,w} \in \mathbb{R}_{++}^{K \times K}, Q_{w,u} \in \mathbb{R}_{++}^{K \times 1}, Q_{u,w} \in \mathbb{R}_{++}^{1 \times K}$ and $Q_{u,u} \in \mathbb{R}_{++}$ are defined as

$$\begin{aligned} Q_{w,w} &= 2 C_w \mathbf{1}\mathbf{1}^T, & Q_{w,u} &= (2 C_u + \|\alpha\|_\infty) \mathbf{1}, \\ Q_{u,w} &= (2 C_w + 1) \mathbf{1}^T, & Q_{u,u} &= (2 C_u + \|\alpha\|_\infty - 1). \end{aligned}$$

In the supplementary material, we prove that $\Psi_{p,q}^\alpha(\delta, t) \leq \sum_{l=1}^{n_1} \delta_l t^{\alpha_l}$ which yields the weaker bounds ξ_1, ξ_2 given in Theorem 1. In particular, this observation combined with Theorems 3 and 4 implies Theorem 1.

Proof. We omit in the following the summation intervals as this is clear from the definition of the variables e.g. if $w \in \mathbb{R}^{K \times n_1}$ then $\sum_{a,b} w_{a,b} = \sum_{a=1}^{K} \sum_{b=1}^{n_1} w_{a,b}$. We split the proof in three steps so that we can reuse some of them in the proof of the similar theorem for two hidden layers.

9

Proposition 1. *Suppose there exist $M_{i,j}, C_i > 0$, $i,j \in \{w, u\}$ such that for every $x \in \{x^1, \ldots, x^n\}$ and $r, s \in [K]$, we have*

$$\sum_t w_{s,t} \frac{\partial f_r(x)}{\partial w_{s,t}} \le C_w, \qquad \sum_{a,b} u_{ab} \frac{\partial f_r(x)}{\partial u_{ab}} \le C_u,$$

$$\sum_t w_{s,t} \frac{\partial^2 f_r(x)}{\partial w_{s,t} \partial w_{a,b}} \le M_{w,w} \frac{\partial f_r(x)}{\partial w_{a,b}}, \qquad \sum_{a,b} u_{ab} \frac{\partial^2 f_r(x)}{\partial u_{ab} \partial w_{s,t}} \le M_{w,u} \frac{\partial f(x)}{\partial w_{s,t}},$$

$$\sum_t w_{s,t} \frac{\partial^2 f_r(x)}{\partial w_{s,t} \partial u_{ab}} \le M_{u,w} \frac{\partial f_r(x)}{\partial u_{ab}}, \qquad \sum_{a,b} u_{ab} \frac{\partial^2 f_r(x)}{\partial u_{ab} \partial u_{st}} \le M_{u,u} \frac{\partial f(x)}{\partial u_{st}}.$$

Then A and $F = G^\Phi$ satisfy all assumptions of Lemma 4 for

$$A = 2 \operatorname{diag} \left(p'_w - 1, \ldots, p'_w - 1, p'_u - 1\right) \begin{pmatrix} Q_{w,w} & Q_{w,u} \\ Q_{u,w} & Q_{u,u} \end{pmatrix}$$

where $Q_{w,w} \in \mathbb{R}_{++}^{K \times K}, Q_{w,u} \in \mathbb{R}_{++}^{K \times 1}, Q_{u,w} \in \mathbb{R}_{++}^{1 \times K}, Q_{u,u} \in \mathbb{R}_{++}$ are constant matrices given by

$$\begin{aligned} Q_{w,w} &= (2C_w + M_{w,w})\mathbf{1}\mathbf{1}^T, & Q_{w,u} &= (2C_u + M_{w,u})\mathbf{1}, \\ Q_{u,w} &= (2C_w + M_{u,w})\mathbf{1}^T, & Q_{u,u} &= (2C_u + M_{u,u}). \end{aligned}$$

Proof. In order to save space we make the proof w.r.t. to abstract variables g, h, that is $g, h \in \{w_1, \ldots, w_K, u\}$. First of all, note that we have

$$
\begin{aligned}
\frac{\partial G_{h_a}^\Phi}{\partial g_s} &= \frac{\partial}{\partial g_s} \frac{\rho_h \left(\frac{\partial \Phi}{\partial h_a}\right)^{p'_h - 1}}{\|\psi_{p'_h}(\nabla_h \Phi)\|_{p_h}} = \frac{\partial}{\partial g_s} \frac{\rho_h \left(\frac{\partial \Phi}{\partial h_a}\right)^{p'_h - 1}}{\|\nabla_h \Phi\|_{p'_h}^{p'_h - 1}} \\
&= (p'_h - 1) \rho_h \left[\frac{\left(\frac{\partial \Phi}{\partial h_a}\right)^{p'_h - 2} \frac{\partial^2 \Phi}{\partial g_s \partial h_a}}{\|\nabla_h \Phi\|_{p'_h}^{p'_h - 1}} - \frac{\left(\frac{\partial \Phi}{\partial h_a}\right)^{p'_h - 1} \|\nabla_h \Phi\|_{p'_h}^{-1} \sum_c \left(\frac{\partial \Phi}{\partial h_c}\right)^{p'_h - 1} \frac{\partial^2 \Phi}{\partial g_s \partial h_c}}{\|\nabla_h \Phi\|_{p'_h}^{2p'_h - 2}} \right] \\
&= \left[\frac{\frac{\partial^2 \Phi}{\partial g_s \partial h_a}}{\frac{\partial \Phi}{\partial h_a}} - \frac{\sum_c \left(\frac{\partial \Phi}{\partial h_c}\right)^{p'_h - 1} \frac{\partial^2 \Phi}{\partial g_s \partial h_c}}{\|\nabla_h \Phi\|_{p'_h}^{p'_h}} \right] (p'_h - 1) G_{h_a}^\Phi,
\end{aligned}
$$

where $\rho_h = \rho_w$ if $h \in \{w_1, \ldots, w_K\}$ and $\rho = \rho_u$ if $h = u$. Thus,

$$\sum_s g_s \left| \frac{\partial G_{h_a}^\Phi}{\partial g_s} \right| = \left[\sum_s g_s \left| \frac{\frac{\partial^2 \Phi}{\partial g_s \partial h_a}}{\frac{\partial \Phi}{\partial h_a}} - \frac{\sum_c \left(\frac{\partial \Phi}{\partial h_c}\right)^{p'_h - 1} \frac{\partial^2 \Phi}{\partial g_s \partial h_c}}{\|\nabla_h \Phi\|_{p'_h}^{p'_h}} \right| \right] (p'_h - 1) G_{h_a}^\Phi.$$

Now, suppose that there exists $R_{h,g} > 0$ such that

$$\sum_s g_s \left| \frac{\partial^2 \Phi}{\partial g_s \partial h_a} \right| \le R_{h,g} \frac{\partial \Phi}{\partial h_a}, \tag{10}$$

10

then we get

$$A_{h,g} = (p'_h - 1) \max_a \sum_s g_s \left| \frac{\frac{\partial^2 \Phi}{\partial g_s \partial h_a}}{\frac{\partial \Phi}{\partial h_a}} - \frac{\sum_c \left(\frac{\partial \Phi}{\partial h_c} \right)^{p'_h - 1} \frac{\partial^2 \Phi}{\partial g_s \partial h_c}}{\|\nabla_h \Phi\|_{p'_h}^{p'_h}} \right|$$

$$\leq (p' - 1) \max_a \sum_s g_s \left| \frac{\frac{\partial^2 \Phi}{\partial g_s \partial h_a}}{\frac{\partial \Phi}{\partial h_a}} \right| + (p'_h - 1) \sum_s g_s \left| \frac{\sum_c \left(\frac{\partial \Phi}{\partial h_c} \right)^{p'_h - 1} \frac{\partial^2 \Phi}{\partial g_s \partial h_c}}{\|\nabla_h \Phi\|_{p'_h}^{p'_h}} \right|$$

$$\leq (p'_h - 1) \max_a \sum_s g_s \left| \frac{\frac{\partial^2 \Phi}{\partial g_s \partial h_a}}{\frac{\partial \Phi}{\partial h_a}} \right| + (p'_h - 1) \frac{\sum_c \left(\frac{\partial \Phi}{\partial h_c} \right)^{p'_h - 1} \sum_s g_s \left| \frac{\partial^2 \Phi}{\partial g_s \partial h_c} \right|}{\|\nabla_h \Phi\|_{p'_h}^{p'_h}}$$

$$\leq (p'_h - 1) R_{h,g} + (p'_h - 1) R_{h,g} \frac{\sum_c \left(\frac{\partial \Phi}{\partial h_c} \right)^{p'_h}}{\|\nabla_h \Phi\|_{p'_h}^{p'_h}}$$

$$= 2(p'_h - 1) R_{h,g}.$$

It follows that, if we define $A_{h,g} = 2(p'_h - 1) R_{h,g}$ for all $g, h \in \{w_1, \ldots, w_K, u\}$, then A and G^Φ satisfy all assumptions of Lemma 4. In order to conclude the proof, we show that $R_{h,g} = 2C_g + M_{h,g}$. We compute the derivatives of the cross-entropy loss

$$-\frac{\partial L}{\partial f_r(x)} = \delta_{yr} - \frac{e^{f_r(x)}}{\sum_{j=1}^K e^{f_j(x)}},$$

$$-\frac{\partial^2 L}{\partial f_q(x) \partial f_r(x)} = -\delta_{qr} \frac{e^{f_r(x)}}{\sum_{j=1}^K e^{f_j(x)}} + \frac{e^{f_r(x)} e^{f_q(x)}}{\left(\sum_{j=1}^K e^{f_j(x)} \right)^2},$$

We get for the derivatives of the objective with respect to the abstract variables g, h

$$\frac{\partial \Phi}{\partial g_s} = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^K \left(-\frac{\partial L}{\partial f_r} \Big|_{f(x^i)} + 1 \right) \frac{\partial f_r}{\partial g_s} \Big|_{x^i} + \epsilon \tag{11}$$

and

$$\frac{\partial \Phi}{\partial g_s \partial h_a} = \frac{1}{n} \sum_{i=1}^n \sum_{q,r=1}^K \left(-\frac{\partial^2 L}{\partial f_q \partial f_r} \Big|_{f(x^i)} \right) \frac{\partial f_r}{\partial h_a} \Big|_{x^i} \frac{\partial f_q}{\partial g_s} \Big|_{x^i} + \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^K \left(-\frac{\partial L}{\partial f_r} \Big|_{f(x^i)} + 1 \right) \frac{\partial^2 f_r}{\partial g_s \partial h_a} \Big|_{x^i}$$

We then can upper bound

$$\sum_s g_s \left| \sum_{q,r=1}^K \left(-\frac{\partial^2 L}{\partial f_q \partial f_r} \Big|_{f(x^i)} \right) \frac{\partial f_r}{\partial h_a} \Big|_{x^i} \frac{\partial f_q}{\partial g_s} \Big|_{x^i} \right|$$

$$\leq \sum_{q,r=1}^K \left| \left(\frac{\partial^2 L}{\partial f_q \partial f_r} \Big|_{f(x^i)} \right) \right| \frac{\partial f_r}{\partial h_a} \Big|_{x^i} \sum_s g_s \frac{\partial f_q}{\partial g_s} \Big|_{x^i}$$

$$\leq \max_{l=1,\ldots,K} \sum_s g_s \frac{\partial f_l}{\partial g_s} \Big|_{x^i} \sum_{r=1}^K \frac{e^{f_r(x^i)}}{\sum_{j=1}^K e^{f_j(x^i)}} \left[\sum_{q \neq r} \frac{e^{f_q(x^i)}}{\sum_{j=1}^K e^{f_j(x)}} + \left(1 - \frac{e^{f_r(x^i)}}{\sum_{j=1}^K e^{f_j(x^i)}} \right) \right] \frac{\partial f_r}{\partial h_a} \Big|_{x^i}$$

$$= 2 \max_{l=1,\ldots,K} \sum_s g_s \frac{\partial f_l}{\partial g_s} \Big|_{x^i} \sum_{r=1}^K \frac{e^{f_r(x^i)}}{\sum_{j=1}^K e^{f_j(x^i)}} \left(1 - \frac{e^{f_r(x^i)}}{\sum_{j=1}^K e^{f_j(x^i)}} \right) \frac{\partial f_r}{\partial h_a} \Big|_{x^i}$$

$$\leq 2 \max_{l=1,\ldots,K} \sum_s g_s \frac{\partial f_l}{\partial g_s} \Big|_{x^i} \sum_{r=1}^K \left(1 - \frac{e^{f_r(x^i)}}{\sum_{j=1}^K e^{f_j(x^i)}} + \delta_{y_i r} \right) \frac{\partial f_r}{\partial h_a} \Big|_{x^i}$$

11

where we have used the non-negativity of the derivatives of the function class in the first step and

$$\sum_{q \neq r} \frac{e^{f_q(x)}}{\sum_{j=1}^{K} e^{f_j(x)}} = 1 - \frac{e^{f_r(x)}}{\sum_{j=1}^{K} e^{f_j(x)}}.$$

Now, note that

$$\sum_s g_s \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{K} \left(-\frac{\partial L}{\partial f_r}\Big|_{f(x^i)} + 1 \right) \frac{\partial^2 f_r}{\partial h_a \partial g_s}\Big|_{x^i} \leq M_{h,g} \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{K} \left(-\frac{\partial L}{\partial f_r}\Big|_{f(x^i)} + 1 \right) \frac{\partial f_r}{\partial h_a}\Big|_{x^i}$$

$$\leq M_{h,g} \frac{\partial \Phi}{\partial h_a}.$$

This shows that

$$\sum_s g_s \left| \frac{\partial^2 \Phi}{\partial g_s \partial h_a} \right|$$

$$\leq 2 \max_{k,l} \sum_s g_s \frac{\partial f_l}{\partial g_s}\Big|_{x^k} \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{K} \left(1 - \frac{e^{f_r(x^i)}}{\sum_{j=1}^{K} e^{f_j(x^i)}} + \delta_{y_i r} \right) \frac{\partial f_r}{\partial h_a}\Big|_{x^i} + M_{h,g} \frac{\partial \Phi}{\partial h_a}$$

$$\leq \left[2 \max_{k,l} \sum_s g_s \frac{\partial f_l}{\partial g_s}\Big|_{x^k} + M_{h,g} \right] \frac{\partial \Phi}{\partial h_a} \leq (2C_g + M_{h,g}) \frac{\partial \Phi}{\partial h_a}.$$

Thus, we get $R_{h,g} \leq 2C_g + M_{h,g}$. $\qquad\qquad\square$

The following lemma is useful for the estimation of the bounds C_w, C_u in Proposition 1.

Lemma 6. *Let $\alpha \in \mathbb{R}_{++}^r$, $p, p_u \in [1, \infty)$ and $\Psi_{p,p_u}^\alpha : \mathbb{R}_{++}^r \times \mathbb{R}_{++} \to \mathbb{R}_{++}$ defined as in (9). Then, for $x \in \mathbb{R}_+^s$ and $u \in \mathbb{R}_{++}^{r \times s}$ satisfying $\|u\|_{p_u} \leq \rho_u$, we have*

$$\left[\sum_l \delta_l^p \left(\sum_m u_{lm} x_m \right)^{p\alpha_l} \right]^{1/p} \leq \Psi_{p,p_u}^\alpha(\delta, \rho_u \|x\|_{p_u'}).$$

Moreover, if $(\delta, t) \leq (\tilde{\delta}, \tilde{t})$ then $\Psi_{p,p_u}^\alpha(\delta, t) \leq \Psi_{p,p_u}^\alpha(\tilde{\delta}, \tilde{t}) \leq \sum_{l=1}^{r} \tilde{\delta}_l \tilde{t}^{\alpha_l}$.

Proof. Let $J = \{l \in [r] \mid \alpha_l p < p_u\}$ and $J^c = \{l \in [r] \mid \alpha_l p \geq p_u\}$, we start with some observations. On the one hand, as $\|u\|_{p_u} \leq \rho_u$, we have

$$\sum_{l \in J^c} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{p\alpha_l / p_u} \leq \sum_{l \in J^c} \frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \leq \frac{\|u\|_{p_u}^{p_u}}{\rho_u^{p_u}} \leq 1.$$

On the other hand, if $\overline{\alpha} = \min_{l \in J} \alpha_l$, then $\overline{p} = \frac{p_u}{\overline{\alpha} p} > 1$, $\overline{p}' = \frac{p_u}{p_u - \overline{\alpha} p}$, and

$$\sum_{l \in J} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{\overline{p}(p\alpha_l / p_u)} = \sum_{l \in J} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{\frac{\alpha_l}{\overline{\alpha}}} \leq \sum_{l \in J} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right) \leq 1.$$

It follows that

$$\sum_l \delta_l^p \left(\sum_m u_{lm} x_m \right)^{p\alpha_l} \leq \sum_l \left(\sum_m u_{lm}^{p_u} \right)^{\frac{p\alpha_l}{p_u}} \delta_l^p \|x\|_{p_u'}^{p\alpha_l}$$

$$= \sum_{l \in J} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{p\alpha_l / p_u} \delta_l^p (\rho_u \|x\|_{p_u'})^{p\alpha_l} + \sum_{l \in J^c} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{p\alpha_l / p_u} \delta_l^p (\rho_u \|x\|_{p_u'})^{p\alpha_l}$$

$$\leq \sum_{l \in J} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{p\alpha_l / p_u} \delta_l^p (\rho_u \|x\|_{p_u'})^{p\alpha_l} + \max_{j \in J^c} \delta_j^p (\rho_u \|x\|_{p_u'})^{p\alpha_j} \sum_{l \in J^c} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{p\alpha_l / p_u}$$

$$\leq \left[\sum_{l \in J} \left(\frac{\sum_m u_{lm}^{p_u}}{\rho_u^{p_u}} \right)^{\overline{p}(p\alpha_l / p_u)} \right]^{1/\overline{p}} \left[\sum_{l \in J} (\delta_l^p (\rho_u \|x\|_{p_u'})^{p\alpha_l})^{\overline{p}'} \right]^{1/\overline{p}'} + \max_{j \in J^c} \delta_j^p (\rho_u \|x\|_{p_u'})^{p\alpha_j}$$

$$\leq \left[\sum_{l \in J} (\delta_l^p (\rho_u \|x\|_{p_u'})^{p\alpha_l})^{\overline{p}'} \right]^{1/\overline{p}'} + \max_{j \in J^c} \delta_j^p (\rho_u \|x\|_{p_u'})^{p\alpha_j} = \left[\Psi_{p,p_u}^\alpha(\delta, \rho_u \|x\|_{p_u'}) \right]^p$$

12

In particular, we see that

$$\Psi^{\alpha}_{p,p_u}(\delta,t) = \left(\Big[\sum_{l\in J}(\delta_l\, t^{\alpha_l})^{p\,\overline{p}'} \Big]^{1/\overline{p}'} + \max_{j\in J^c}(\delta_j\, t^{\alpha_j})^p \right)^{1/p}.$$

Finally, let $(\delta,t) \le (\tilde{\delta},\tilde{t})$. By monotonicity of $\|\cdot\|_{\overline{p}}$ and $\|\cdot\|_p$, we have $\Psi^{\alpha}_{p,p_u}(\delta,t) \le \Psi^{\alpha}_{p,p_u}(\tilde{\delta},\tilde{t})$. Now, using $\|\cdot\|_p \le \|\cdot\|_1$ and $\|\cdot\|_{p\,\overline{p}'} \le \|\cdot\|_1$, we get

$$\begin{aligned}
\Psi^{\alpha}_{p,p_u}(\delta,t) &= \left(\Big(\Big[\sum_{l\in J}(\delta_l\, t^{\alpha_l})^{p\,\overline{p}'} \Big]^{1/(p\,\overline{p}')} \Big)^p + \max_{j\in J^c}(\delta_j\, t^{\alpha_j})^p \right)^{1/p} \\
&\le \Big[\sum_{l\in J}(\delta_l\, t^{\alpha_l})^{p\,\overline{p}'} \Big]^{1/(p\,\overline{p}')} + \max_{j\in J^c}\delta_j\, t^{\alpha_j} \le \sum_{l\in J}\delta_l\, t^{\alpha_l} + \max_{j\in J^c}\delta_j\, t^{\alpha_j} \\
&\le \sum_{l\in[r]}\delta_l\, t^{\alpha_l}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square
\end{aligned}$$

Finally, using the Lemma above, we explicit the bounds of Proposition 1.

Lemma 7. *Let f be defined as in (1) and let $\rho_x = \max_{i=1\in[n]} \|x^i\|_{p'_u}$, then the bounds in Proposition 1 are given by $C_w = \rho_w \Psi^{\alpha}_{p'_w,p_u}(\mathbf{1},\rho_u\rho_x)$, $C_u = \rho_w \Psi^{\alpha}_{p'_w,p_u}(\alpha,\rho_u\rho_x)$, $M_{w,w} = 0$, $M_{w,u} = \|\alpha\|_\infty$, $M_{u,w} = 1$ and $M_{u,u} = \|\alpha\|_\infty - 1$.*

Proof. Let $x \in \{x^1,\ldots,x^n\}$ and, for $(w,u) \in B_{++}$, we have

$$\begin{aligned}
\frac{\partial f_r(x)}{\partial w_{a,b}} &= \delta_{ra}\Big(\sum_m u_{bm}x_m \Big)^{\alpha_b}, \\
\frac{\partial f_r(x)}{\partial u_{ab}} &= w_{r,a}\alpha_a\Big(\sum_m u_{am}x_m \Big)^{\alpha_a-1}x_b, \\
\frac{\partial^2 f_r(x)}{\partial w_{s,t}\partial w_{a,b}} &= 0, \\
\frac{\partial^2 f_r(x)}{\partial w_{s,t}\partial u_{ab}} &= \delta_{rs}\delta_{at}\alpha_a\Big(\sum_m u_{am}x_m \Big)^{\alpha_a-1}x_b, \\
\frac{\partial^2 f_r(x)}{\partial u_{st}\partial u_{ab}} &= \delta_{as}w_{r,a}\alpha_a(\alpha_a-1)\Big(\sum_m u_{am}x_m \Big)^{\alpha_a-2}x_bx_t.
\end{aligned}$$

With Lemma 6, we have

$$\begin{aligned}
\sum_b w_{a,b}\frac{\partial f_r(x)}{\partial w_{a,b}} &= \delta_{ra}\sum_b w_{a,b}\Big(\sum_m u_{bm}x_m \Big)^{\alpha_b} \le \delta_{ra}\|w_a\|_{p_w}\Big(\sum_b \big(\sum_m u_{bm}x_m \big)^{\alpha_b p'_w} \Big)^{1/p'_w} \\
&\le \delta_{ra}\|w_a\|_{p_w}\Psi^{\alpha}_{p'_w,p_u}(\mathbf{1},\rho_u\|x\|_{p'_u}) \le \rho_w\Psi^{\alpha}_{p'_w,p_u}(\mathbf{1},\rho_u\rho_x) = C_w,
\end{aligned}$$

and

$$\begin{aligned}
\sum_{a,b} u_{ab}\frac{\partial f_r(x)}{\partial u_{ab}} &\le \sum_a w_{r,a}\alpha_a\Big(\sum_m u_{am}x_m \Big)^{\alpha_a} \le \|w_r\|_{p_w}\Psi^{\alpha}_{p'_w,p_u}(\alpha,\rho_u\|x\|_{p'_u}) \\
&\le \rho_w\Psi^{\alpha}_{p'_w,p_u}(\alpha,\rho_u\rho_x) = C_u.
\end{aligned}$$

13

Furthermore,

$$\sum_t w_{s,t}\frac{\partial^2 f_r(x)}{\partial w_{s,t}\partial w_{a,b}} = 0 \qquad \implies \quad M_{w,w} = 0$$

$$\sum_{a,b} u_{ab}\frac{\partial^2 f_r(x)}{\partial u_{ab}\partial w_{s,t}} = \delta_{rs}\alpha_t\Big(\sum_m u_{tm}x_m\Big)^{\alpha_t} = \alpha_t\frac{\partial f(x)}{\partial w_{s,t}} \le \|\alpha\|_\infty\frac{\partial f(x)}{\partial w_{s,t}}$$

$$\implies \quad M_{w,u} = \|\alpha\|_\infty$$

$$\sum_t w_{s,t}\frac{\partial^2 f_r(x)}{\partial w_{s,t}\partial u_{ab}} = \delta_{rs}w_{s,a}\alpha_a\Big(\sum_m u_{am}x_m\Big)^{\alpha_a-1}x_b = \delta_{sr}\frac{\partial f_r(x)}{\partial u_{ab}} \le \frac{\partial f_r(x)}{\partial u_{ab}}$$

$$\implies \quad M_{u,w} = 1$$

$$\sum_{a,b} u_{ab}\frac{\partial^2 f_r(x)}{\partial u_{st}\partial u_{ab}} = w_{r,s}\alpha_s(\alpha_s-1)\Big(\sum_m u_{sm}x_m\Big)^{\alpha_s-1}x_t = (\alpha_s-1)\frac{\partial f_r(x)}{\partial u_{st}}$$

$$\le \|\alpha-\mathbf{1}\|_\infty\frac{\partial f_r(x)}{\partial u_{st}} \qquad \implies \quad M_{u,u} = \|\alpha-\mathbf{1}\|_\infty.$$

Finally, as $\alpha_i \ge 1$ for every $i \in [n_1]$, we have $\|\alpha-\mathbf{1}\|_\infty = \|\alpha\|_\infty - 1$. $\qquad\square$

Combining Lemma 7 and Proposition 1 concludes the proof. Finally, we note that the upper bound on $\Psi_{p,q}^\alpha$ proved in Lemma 6 implies that the matrix A in Theorem 4 is componentwise smaller or equal than that of Theorem 1 and thus has a smaller spectral radius by Corollary 3.30 [3]. $\qquad\square$

4.1 Neural networks with two hidden layers

We show how to extend our framework for neural networks with 2 hidden layers. In future work we will consider the general case. We briefly explain the major changes. Let $n_1, n_2 \in \mathbb{N}$ and $\alpha \in \mathbb{R}_{++}^{n_1}, \beta \in \mathbb{R}_{++}^{n_2}$ with $\alpha_i, \beta_j \ge 1$ for all $i \in [n_1], j \in [n_2]$, our function class is:

$$f_r(x) = f_r(w,v,u)(x) = \sum_{l=1}^{n_2} w_{r,l}\Big(\sum_{m=1}^{n_1} v_{lm}\Big(\sum_{s=1}^{d} u_{ms}x_s\Big)^{\alpha_m}\Big)^{\beta_l}$$

and the optimization problem becomes

$$\max_{(w,v,u)\in S_+} \Phi(w,v,u) \qquad \text{where} \qquad V_+ = \mathbb{R}_+^{K\times n_2} \times \mathbb{R}_+^{n_2\times n_1} \times \mathbb{R}_+^{n_1\times d}, \qquad (12)$$

$S_+ = \{(w_1,\dots,w_K,v,u)\in V_+ \mid \|w_i\|_{p_w} = \rho_w, \|v\|_{p_v} = \rho_v, \|u\|_{p_u} = \rho_u\}$ and

$$\Phi(w,v,u) = \frac{1}{n}\sum_{i=1}^n\Big[-L\big(y_i, f(x^i)\big)+\sum_{r=1}^K f_r(x^i)\Big]+\epsilon\Big(\sum_{r=1}^K\sum_{l=1}^{n_2} w_{r,l}+\sum_{l=1}^{n_2}\sum_{m=1}^{n_1} v_{lm}+\sum_{m=1}^{n_1}\sum_{s=1}^d u_{ms}\Big).$$

The map $G^\Phi\colon S_{++} \to S_{++} = \{z\in S_+ \mid z > 0\}$, $G^\Phi = (G_{w_1}^\Phi,\dots,G_{w_K}^\Phi, G_v^\Phi, G_u^\Phi)$, becomes

$$G_{w_i}^\Phi(w,v,u) = \rho_w\frac{\psi_{p'_w}(\nabla_{w_i}\Phi(w,u))}{\|\psi_{p'_w}(\nabla_{w_i}\Phi(w,v,u))\|_{p_w}} \qquad \forall i \in [K] \qquad (13)$$

and

$$G_v^\Phi(w,v,u) = \rho_v\frac{\psi_{p'_v}(\nabla_v\Phi(w,v,u))}{\|\psi_{p'_v}(\nabla_v\Phi(w,v,u))\|_{p_v}}, \qquad G_u^\Phi(w,v,u) = \rho_u\frac{\psi_{p'_u}(\nabla_u\Phi(w,v,u))}{\|\psi_{p'_u}(\nabla_u\Phi(w,v,u))\|_{p_u}}.$$

We have the following equivalent of Theorem 1 for 2 hidden layers.

Theorem 5. *Let $\{x^i, y_i\}_{i=1}^n \subset \mathbb{R}_+^d \times [K]$, $p_w, p_v, p_u \in (1,\infty)$, $\rho_w, \rho_v, \rho_u > 0$, $n_1, n_2 \in \mathbb{N}$ and $\alpha \in \mathbb{R}_{++}^{n_1}, \beta \in \mathbb{R}_{++}^{n_2}$ with $\alpha_i, \beta_j \ge 1$ for all $i \in [n_1], j \in [n_2]$. Let $\rho_x = \max_{i\in[n]}\|x^i\|_{p'_u}$,*

$$\theta = \rho_v\Psi_{p'_v,p_u}^\alpha(\mathbf{1}, \rho_u\rho_x), \quad C_w = \rho_w\Psi_{p'_w,p_v}^\beta(\mathbf{1},\theta), \quad C_v = \rho_w\Psi_{p'_w,p_v}^\beta(\beta,\theta), \quad C_u = \|\alpha\|_\infty C_v,$$

14

and define $A \in \mathbb{R}_{++}^{(K+2)\times(K+2)}$ *as*

$$
\begin{aligned}
A_{m,l} &= 4(p_w' - 1)C_w, & A_{m,K+1} &= 2(p_w' - 1)(2C_v + \|\beta\|_\infty) \\
A_{m,K+2} &= 2(p_w' - 1)(2C_u + \|\alpha\|_\infty\|\beta\|_\infty), & A_{K+1,l} &= 2(p_v' - 1)(2C_w + 1) \\
A_{K+1,K+1} &= 2(p_v' - 1)(2C_v + \|\beta\|_\infty - 1), & A_{K+1,K+2} &= 2(p_v' - 1)(2C_u + \|\alpha\|_\infty\|\beta\|_\infty) \\
A_{K+2,l} &= 2(p_u' - 1)(2C_w + 1), & A_{K+2,K+1} &= 2(p_u' - 1)(2C_v + \|\beta\|_\infty), \\
A_{K+2,K+2} &= 2(p_u' - 1)(2C_u + \|\alpha\|_\infty\|\beta\|_\infty - 1) & & \forall m, l \in [K].
\end{aligned}
$$

If $\rho(A) < 1$, then (12) has a unique global maximizer $(w^*, v^*, u^*) \in S_{++}$. Moreover, for every $(w^0, v^0, u^0) \in S_{++}$, there exists $R > 0$ such that

$$
\lim_{k\to\infty}(w^k, v^k, u^k) = (w^*, v^*, u^*) \qquad \text{and} \qquad \|(w^k, v^k, u^k) - (w^*, v^*, u^*)\|_\infty \le R\,\rho(A)^k \quad \forall k \in \mathbb{N}
$$

where $(w^{k+1}, v^{k+1}, u^{k+1}) = G^\Phi(w^k, v^k, u^k)$ for every $k \in \mathbb{N}$ and G^Φ is defined as in (13).

Proof. The proof of Theorem 3, can be extended by considering the weighted Thompson metric $\mu\colon V_{++} \times V_{++} \to \mathbb{R}_+$ defined as

$$
\mu\big((w, v, u), (\tilde{w}, \tilde{v}, \tilde{u})\big) = \sum_{i=1}^{K} \gamma_i \|\ln(w_i) - \ln(\tilde{w}_i)\|_\infty + \gamma_{K+1}\|\ln(v) - \ln(\tilde{v})\|_\infty + \gamma_{K+2}\|\ln(u) - \ln(\tilde{u})\|.
$$

In particular, when $\rho(A) < 1$, the convergence rate becomes

$$
\|(w^k, v^k, u^k) - (w^*, v^*, u^*)\|_\infty \le \rho(A)^k \left(\frac{\mu\big((w^1, v^1, u^1), (w^0, v^0, u^0)\big)}{(1 - \rho(A)) \min\big\{ \frac{\gamma_{K+2}}{\rho_u}, \frac{\gamma_{K+1}}{\rho_v}, \min_{t\in[K]} \frac{\gamma_t}{\rho_w} \big\}} \right) \quad \forall k \in \mathbb{N},
$$

where the weights in the definition of μ are the components of the positive eigenvector of A^T. As $A^T \in \mathbb{R}_{++}^{(K+2)\times(K+2)}$, this vector always exists by the Perron-Frobenius theorem (see for instance Theorem 8.4.4 in [10]). Proposition 1 generalizes straightforwardly. So, we need to compute the corresponding quantities $C_g, M_{g,h} > 0$ for $g, h \in \{w_1, \ldots, w_K, u, v\}$. To shorten notations, we write

$$
(ux)^\alpha = \big((ux)_1^{\alpha_1}, \ldots, (ux)_{n_1}^{\alpha_{n_1}}\big) \qquad \text{and} \qquad (v(ux)^\alpha)^\beta = \big((v(ux)^\alpha)_1^{\beta_1}, \ldots, (v(ux)^\alpha)_{n_2}^{\beta_{n_2}}\big),
$$

where

$$
(ux)_i = \sum_{t=1}^{d} u_{it} x_t \qquad \text{and} \qquad (v(ux)^\alpha)_j = \sum_{s=1}^{n_1} v_{js}(ux)_s^{\alpha_s} \qquad \forall i \in [n_1],\, j \in [n_2].
$$

Again, we omit the summation intervals in the following. First, we compute

$$
\frac{\partial f_r(x)}{\partial w_{a,b}} = \delta_{ra}(v(ux)^\alpha)_b^{\beta_b}
$$

$$
\frac{\partial f_r(x)}{\partial v_{ab}} = w_{r,a}\beta_a(v(ux)^\alpha)_a^{\beta_a - 1}(ux)_b^{\alpha_b}
$$

$$
\frac{\partial f_r(x)}{\partial u_{ab}} = \sum_j w_{r,j}\beta_j(v(ux)^\alpha)_j^{\beta_j - 1} v_{ja}\alpha_a(ux)_a^{\alpha_a - 1}x_b
$$

Hence

$$
\sum_b w_{a,b}\frac{\partial f_r(x)}{\partial w_{a,b}} = \delta_{ra}\sum_b w_{r,b}(v(ux)^\alpha)_b^{\beta_b} = \delta_{ra} f_r(x)
$$

$$
\sum_{a,b} v_{ab}\frac{\partial f_r(x)}{\partial v_{ab}} = \sum_a w_{r,a}\beta_a(v(ux)^\alpha)_a^{\beta_a - 1}\sum_b v_{ab}(ux)_b^{\alpha_b} = \sum_a \beta_a w_{r,a}(v(ux)^\alpha)_a^{\beta_a}
$$

$$
\sum_{a,b} u_{ab}\frac{\partial f_r(x)}{\partial u_{ab}} = \sum_j w_{r,j}\beta_j(v(ux)^\alpha)_j^{\beta_j - 1}\sum_a v_{ja}\alpha_a(ux)_a^{\alpha_a - 1}\sum_b u_{ab}x_b
$$

$$
= \sum_j \beta_j w_{r,j}(v(ux)^\alpha)_j^{\beta_j - 1}\sum_a \alpha_a v_{ja}(ux)_a^{\alpha_a}.
$$

15

It follows with Lemma 6 that, with $\theta = \rho_v\,\Psi^\alpha_{p'_v,p_u}(\mathbf{1},\rho_u\rho_x)$, we have

$$\sum_b w_{a,b}\frac{\partial f_r(x)}{\partial w_{a,b}} = \delta_{ra}\sum_b w_{a,b}(v(ux)^\alpha)^{\beta_b}_b \le \|w_a\|_{p_w}\Big(\sum_b (v(ux)^\alpha)^{\beta_b p'_w}_b\Big)^{1/p'_w}$$

$$\le \|w_a\|_{p_w}\Psi^\beta_{p'_w,p_v}(\mathbf{1},\rho_v\|(ux)^\alpha\|_{p'_v}) \le \|w_a\|_{p_w}\Psi^\beta_{p'_w,p_v}\big(\mathbf{1},\rho_v\,\Psi^\alpha_{p'_v,p_u}(\mathbf{1},\rho_u\|x\|_{p'_u})\big)$$

$$\le \rho_w\Psi^\beta_{p'_w,p_v}\big(\mathbf{1},\theta\big) = C_w$$

$$\sum_{a,b} v_{ab}\frac{\partial f_r(x)}{\partial v_{ab}} = \sum_a \beta_a w_{r,a}(v(ux)^\alpha)^{\beta_a}_a \le \|w_r\|_{p_w}\Big(\sum_a \big[\beta_a(v(ux)^\alpha)^{\beta_a}_a\big]^{p'_w}\Big)^{1/p'_w}$$

$$\le \|w_r\|_{p_w}\Psi^\beta_{p'_w,p_v}(\beta,\rho_v\|(ux)^\alpha\|_{p'_v}) \le \rho_w\Psi^\beta_{p'_w,p_v}(\beta,\theta) = C_v$$

$$\sum_{a,b} u_{ab}\frac{\partial f_r(x)}{\partial u_{ab}} = \sum_j \beta_j w_{r,j}(v(ux)^\alpha)^{\beta_j-1}_j \sum_a \alpha_a v_{ja}(ux)^{\alpha_a}_a \le \|\alpha\|_\infty \sum_j \beta_j w_{r,j}(v(ux)^\alpha)^{\beta_j}_j$$

$$\le \|\alpha\|_\infty C_v = C_u.$$

Now, for the bound involving the second derivatives, we have

$$\frac{\partial^2 f_r(x)}{\partial w_{s,t}\partial w_{a,b}} = 0 \qquad \Longrightarrow \qquad M_{w,w} = 0,$$

and

$$\frac{\partial^2 f_r(x)}{\partial v_{st}\partial v_{ab}} = \delta_{sa}w_{r,a}\beta_a(\beta_a-1)(v(ux)^\alpha)^{\beta_a-2}_a(ux)^{\alpha_b}_b(ux)^{\alpha_t}_t$$

so that

$$\sum_{s,t} v_{st}\frac{\partial^2 f_r(x)}{\partial v_{st}\partial v_{ab}} = w_{r,a}\beta_a(\beta_a-1)(v(ux)^\alpha)^{\beta_a-2}_a(ux)^{\alpha_b}_b\sum_t v_{at}(ux)^{\alpha_t}_t$$

$$= w_{r,a}(\beta_a-1)\beta_a(v(ux)^\alpha)^{\beta_a-1}_a(ux)^{\alpha_b}_b$$

$$= (\beta_a-1)\frac{\partial f_r(x)}{\partial v_{ab}} \le \|\beta-\mathbf{1}\|_\infty\frac{\partial f_r(x)}{\partial v_{ab}} \qquad \Longrightarrow \qquad M_{v,v} = \|\beta-\mathbf{1}\|_\infty.$$

Moreover, we have

$$\frac{\partial^2 f_r(x)}{\partial u_{st}\partial u_{ab}} = \sum_j w_{r,j}\beta_j v_{ja}\alpha_a(ux)^{\alpha_a-1}_a x_b(\beta_j-1)(v(ux)^\alpha)^{\beta_j-2}_j v_{js}\alpha_s(ux)^{\alpha_s-1}_s x_t$$

$$+ \delta_{as}\sum_j w_{r,j}\beta_j(v(ux)^\alpha)^{\beta_j-1}_j v_{ja}\alpha_a x_b(\alpha_a-1)(ux)^{\alpha_a-2}_a x_t,$$

thus

$$\sum_{s,t} u_{st}\frac{\partial^2 f_r(x)}{\partial u_{st}\partial u_{ab}} = \sum_j w_{r,j}\beta_j v_{ja}\alpha_a(ux)^{\alpha_a-1}_a x_b(\beta_j-1)(v(ux)^\alpha)^{\beta_j-2}_j\sum_s v_{js}\alpha_s(ux)^{\alpha_s-1}_s\sum_t u_{st}x_t$$

$$+ \sum_j w_{r,j}\beta_j(v(ux)^\alpha)^{\beta_j-1}_j v_{ja}\alpha_a x_b(\alpha_a-1)(ux)^{\alpha_a-2}_a\sum_t u_{at}x_t$$

$$= \sum_j w_{r,j}\beta_j v_{ja}\alpha_a(ux)^{\alpha_a-1}_a x_b(\beta_j-1)(v(ux)^\alpha)^{\beta_j-2}_j\sum_s v_{js}\alpha_s(ux)^{\alpha_s}_s$$

$$+ \sum_j w_{r,j}\beta_j(v(ux)^\alpha)^{\beta_j-1}_j v_{ja}\alpha_a x_b(\alpha_a-1)(ux)^{\alpha_a-1}_a$$

$$\le \|\alpha\|_\infty \sum_j w_{r,j}\beta_j v_{ja}\alpha_a(ux)^{\alpha_a-1}_a x_b(\beta_j-1)(v(ux)^\alpha)^{\beta_j-2}_j\sum_s v_{js}(ux)^{\alpha_s}_s$$

$$+ \|\alpha-\mathbf{1}\|_\infty \sum_j w_{r,j}\beta_j(v(ux)^\alpha)^{\beta_j-1}_j v_{ja}\alpha_a x_b(ux)^{\alpha_a-1}_a$$

$$\le (\|\alpha\|_\infty\|\beta-\mathbf{1}\|_\infty + \|\alpha-\mathbf{1}\|_\infty)\frac{\partial f_r(x)}{\partial u_{ab}}$$

$$\Longrightarrow \qquad M_{u,u} = \|\alpha\|_\infty\|\beta-\mathbf{1}\|_\infty + \|\alpha-\mathbf{1}\|_\infty.$$

16

As $\alpha_i, \beta_j \geq 1$ for every $i \in [n_1], j \in [n_2]$ by assumption, we have $\|\alpha - \mathbf{1}\|_\infty = \|\alpha\|_\infty - 1$ and $\|\beta - \mathbf{1}\|_\infty = \|\beta\|_\infty - 1$ so that $M_{u,u} = \|\alpha\|_\infty \|\beta\|_\infty - 1$ and $M_{v,v} = \|\beta\|_\infty - 1$. Now, we look at the mixed second derivatives. We have

$$\frac{\partial^2 f_r(x)}{\partial v_{st} \partial w_{a,b}} = \delta_{ra}\delta_{bs}\beta_s(v(ux)^\alpha)_s^{\beta_s - 1}(ux)_t^{\alpha_t}.$$

It follows that

$$\sum_b w_{a,b}\frac{\partial^2 f_r(x)}{\partial v_{st}\partial w_{a,b}} = \delta_{ra}w_{a,s}\beta_s(v(ux)^\alpha)_s^{\beta_s-1}(ux)_t^{\alpha_t} \leq w_{r,s}\beta_s(v(ux)^\alpha)_s^{\beta_s-1}(ux)_t^{\alpha_t} = \frac{\partial f_r(x)}{\partial v_{st}}$$

$$\implies \qquad M_{v,w} = 1,$$

and

$$\sum_{s,t} v_{st}\frac{\partial^2 f_r(x)}{\partial v_{st}\partial w_{a,b}} = \delta_{ra}\beta_b(v(ux)^\alpha)_b^{\beta_b-1}\sum_t v_{bt}(ux)_t^{\alpha_t} = \beta_b\frac{\partial f_r(x)}{\partial w_{a,b}} \leq \|\beta\|_\infty\frac{\partial f_r(x)}{\partial w_{a,b}}$$

$$\implies \qquad M_{w,v} = \|\beta\|_\infty.$$

Furthermore, it holds

$$\frac{\partial^2 f_r(x)}{\partial u_{st}\partial w_{ab}} = \delta_{ra}\beta_b(v(ux)^\alpha)_b^{\beta_b-1}v_{bs}\alpha_s(ux)_s^{\alpha_s-1}x_t,$$

so that

$$\sum_b w_{a,b}\frac{\partial^2 f_r(x)}{\partial w_{ab}\partial u_{st}} = \delta_{ra}\sum_b w_{a,b}\beta_b(v(ux)^\alpha)_b^{\beta_b-1}v_{bs}\alpha_s(ux)_s^{\alpha_s-1}x_t$$

$$\leq \sum_b w_{r,b}\beta_b(v(ux)^\alpha)_b^{\beta_b-1}v_{bs}\alpha_s(ux)_s^{\alpha_s-1}x_t = \frac{\partial f_r(x)}{\partial u_{st}}$$

$$\implies \qquad M_{u,w} = 1,$$

and

$$\sum_{s,t} u_{st}\frac{\partial^2 f_r(x)}{\partial w_{ab}\partial u_{st}} = \delta_{ra}\beta_b(v(ux)^\alpha)_b^{\beta_b-1}\sum_s v_{bs}\alpha_s(ux)_s^{\alpha_s-1}\sum_t u_{st}x_t$$

$$= \delta_{ra}\beta_b(v(ux)^\alpha)_b^{\beta_b-1}\sum_s v_{bs}\alpha_s(ux)_s^{\alpha_s}$$

$$\leq \|\alpha\|_\infty\delta_{ra}\beta_b(v(ux)^\alpha)_b^{\beta_b} \leq \|\alpha\|_\infty\|\beta\|_\infty\frac{\partial f_r(x)}{\partial w_{a,b}}$$

$$\implies \qquad M_{w,u} = \|\alpha\|_\infty\|\beta\|_\infty.$$

Finally, we have

$$\frac{\partial^2 f_r(x)}{\partial u_{st}\partial v_{ab}} = w_{r,a}\beta_a(ux)_b^{\alpha_b}(\beta_a - 1)(v(ux)^\alpha)_a^{\beta_a-2}v_{as}\alpha_s(ux)_s^{\alpha_s-1}x_t$$

$$+ \delta_{sb}w_{r,a}\beta_a(v(ux)^\alpha)_a^{\beta_a-1}\alpha_b(ux)_b^{\alpha_b-1}x_t.$$

Hence,

$$\sum_{a,b} v_{ab}\frac{\partial^2 f_r(x)}{\partial u_{st}\partial v_{ab}} = \sum_a w_{r,a}\beta_a(\beta_a - 1)(v(ux)^\alpha)_a^{\beta_a-2}v_{as}\alpha_s(ux)_s^{\alpha_s-1}x_t\sum_b v_{ab}(ux)_b^{\alpha_b}$$

$$+ \sum_a w_{r,a}\beta_a(v(ux)^\alpha)_a^{\beta_a-1}v_{as}\alpha_s(ux)_s^{\alpha_s-1}x_t$$

$$= \sum_a w_{r,a}\beta_a(\beta_a - 1)(v(ux)^\alpha)_a^{\beta_a-1}v_{as}\alpha_s(ux)_s^{\alpha_s-1}x_t + \frac{\partial f_r(x)}{\partial u_{st}}$$

$$\leq (\|\beta - \mathbf{1}\|_\infty + 1)\frac{\partial f_r(x)}{\partial u_{st}}$$

$$\implies M_{u,v} = \|\beta - \mathbf{1}\|_\infty + 1 = \|\beta\|_\infty,$$

17

and

$$\sum_{s,t} u_{st} \frac{\partial^2 f_r(x)}{\partial u_{st} \partial v_{ab}} = w_{r,a}\beta_a(ux)_b^{\alpha_b}(\beta_a-1)(v(ux)^\alpha)_a^{\beta_a-2} \sum_s v_{as}\alpha_s(ux)_s^{\alpha_s-1} \sum_t u_{st}x_t$$

$$+ w_{r,a}\beta_a(v(ux)^\alpha)_a^{\beta_a-1}\alpha_b(ux)_b^{\alpha_b-1} \sum_t u_{bt}x_t$$

$$= w_{r,a}\beta_a(ux)_b^{\alpha_b}(\beta_a-1)(v(ux)^\alpha)_a^{\beta_a-2} \sum_s v_{as}\alpha_s(ux)_s^{\alpha_s}$$

$$+ w_{r,a}\beta_a(v(ux)^\alpha)_a^{\beta_a-1}\alpha_b(ux)_b^{\alpha_b}$$

$$\leq \|\alpha\|_\infty w_{r,a}\beta_a(ux)_b^{\alpha_b}(\beta_a-1)(v(ux)^\alpha)_a^{\beta_a-2} \sum_s v_{as}(ux)_s^{\alpha_s}$$

$$+ \|\alpha\|_\infty w_{r,a}\beta_a(v(ux)^\alpha)_a^{\beta_a-1}(ux)_b^{\alpha_b}$$

$$= \|\alpha\|_\infty w_{r,a}\beta_a(ux)_b^{\alpha_b}(\beta_a-1)(v(ux)^\alpha)_a^{\beta_a-1} + \|\alpha\|_\infty \frac{\partial f_r(x)}{\partial v_{ab}}$$

$$\leq \|\alpha\|_\infty(\|\beta-\mathbf{1}\|_\infty+1)\frac{\partial f_r(x)}{\partial v_{ab}}$$

$$\implies M_{v,u} = \|\alpha\|_\infty(\|\beta-\mathbf{1}\|_\infty+1) = \|\alpha\|_\infty\|\beta\|_\infty.$$

$\square$

As for the case with one hidden layer, for any fixed architecture $\rho_w, \rho_v, \rho_u > 0$, $n_1, n_2 \in \mathbb{N}$ and $\alpha \in \mathbb{R}_{++}^{n_1}, \beta \in \mathbb{R}_{++}^{n_2}$ with $\alpha_i, \beta_j \geq 1$ for all $i \in [n_1], j \in [n_2]$, it is possible to derive lower bounds on p_w, p_v, p_u that guarantee $\rho(A) < 1$ in Theorem 5. Indeed, it holds

$$C_w \leq \zeta_1 = \rho_w \sum_{j=1}^{n_2}\left[\rho_v \sum_{l=1}^{n_1}(\rho_u\tilde{\rho}_x)^{\alpha_l}\right]^{\beta_j} \quad \text{and} \quad C_v \leq \zeta_2 = \rho_w \sum_{j=1}^{n_2}\beta_j\left[\rho_v \sum_{l=1}^{n_1}(\rho_u\tilde{\rho}_x)^{\alpha_l}\right]^{\beta_j},$$

with $\tilde{\rho}_x = \max_{i \in [n]}\|x^i\|_1$. Hence, the two hidden layers equivalent of (4) becomes

$$p_w > 4(K+2)\zeta_1 + 5, \ p_v > 2(K+2)\big[2\zeta_2 + \|\beta\|_\infty\big] - 1, \ p_u > 2(K+2)\|\alpha\|_\infty(2\zeta_2 + \|\beta\|_\infty) - 1. \tag{14}$$

5 Experiments

Table 1: Test accuracy on UCI datasets

Dataset	NLSM1	NLSM2	ReLU1	ReLU2	SVM
Cancer	96.4	96.4	95.7	93.6	95.7
Iris	90.0	96.7	100	93.3	100
Banknote	97.1	96.4	100	97.8	100
Blood	76.0	76.7	76.0	76.0	77.3
Haberman	75.4	75.4	70.5	72.1	72.1
Seeds	88.1	90.5	90.5	92.9	95.2
Pima	79.2	80.5	76.6	79.2	79.9
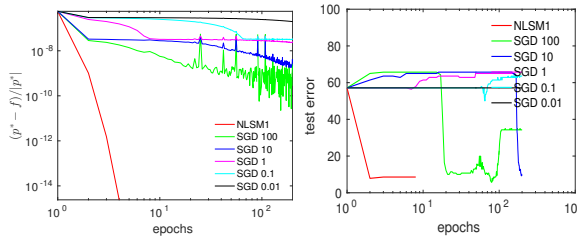
Figure 2: Training score (left) w.r.t. the optimal score p^* and test error (right) of NLSM1 and Batch-SGD with different step-sizes.

The shown experiments should be seen as a proof of concept. We do not have yet a good understanding of how one should pick the parameters of our model to achieve good performance. However, the other papers which have up to now discussed global optimality for neural networks [11, 8] have not included any results on real datasets. Thus, up to our knowledge, we show for the first time a globally optimal algorithm for neural networks that leads to non-trivial classification results.

We test our methods on several low dimensional UCI datasets and denote our algorithms as NLSM1 (one hidden layer) and NLSM2 (two hidden layers). We choose the parameters of our model out of 100 randomly generated combinations of $(n_1, \alpha, \rho_w, \rho_u) \in [2, 20] \times [1, 4] \times (0, 1]^2$

18

<center>Nonlinear Spectral Method for 1 hidden layer</center>

Input: Model $n_1 \in \mathbb{N}$, $p_w, p_u \in (1, \infty)$, $\rho_w, \rho_u > 0$, $\alpha_1, \ldots, \alpha_{n_1} \geq 1$, $\epsilon > 0$ so that the matrix A of Theorem 1 satisfies $\rho(A) < 1$. Accuracy $\tau > 0$ and $(w^0, u^0) \in S_{++}$.
1   Let $(w^1, u^1) = G^\Phi(w^0, u^0)$ and compute R as in Theorem 3
2   **Repeat**
3        $(w^{k+1}, u^{k+1}) = G^\Phi(w^k, u^k)$
4        $k \leftarrow k + 1$
5   **Until** $k \geq \ln\big(\tau/R\big)/\ln\big(\rho(A)\big)$
Output: (w^k, u^k) fulfills $\|(w^k, u^k) - (w^*, u^*)\|_\infty < \tau$.

<center>With G^Φ defined as in (3). The method for two hidden layers is similar: consider G^Φ
as in (13) instead of (3) and assume that the model satisfies Theorem 5.</center>

(respectively $(n_1, n_2, \alpha, \beta, \rho_w, \rho_v, \rho_u) \in [2, 10]^2 \times [1, 4]^2 \times (0, 1]^2)$ and pick the best one based on 5-fold cross-validation error. We use Equation (4) (resp. Equation (14)) to choose p_u, p_w (resp. p_u, p_v, p_w) so that every generated model satisfies the conditions of Theorem 1 (resp. Theorem 5), i.e. $\rho(A) < 1$. Thus, global optimality is guaranteed in all our experiments. For comparison, we use the nonlinear RBF-kernel SVM and implement two versions of the Rectified-Linear Unit network - one for one hidden layer networks (ReLU1) and one for two hidden layers networks (ReLU2). To train ReLU, we use a stochastic gradient descent method which minimizes the sum of logistic loss and L_2 regularization term over weight matrices to avoid over-fitting. All parameters of each method are jointly cross validated. More precisely, for ReLU the number of hidden units takes values from 2 to 20, the step-sizes and regularizers are taken in $\{10^{-6}, 10^{-5}, \ldots, 10^2\}$ and $\{0, 10^{-4}, 10^{-3}, \ldots, 10^4\}$ respectively. For SVM, the hyperparameter C and the kernel parameter γ of the radius basis function $K(x^i, x^j) = \exp(-\gamma \|x^i - x^j\|^2)$ are taken from $\{2^{-5}, 2^{-4} \ldots, 2^{20}\}$ and $\{2^{-15}, 2^{-14} \ldots, 2^3\}$ respectively. Note that ReLUs allow negative weights while our models do not. The results presented in Table 1 show that overall our nonlinear spectral methods achieve slightly worse performance than kernel SVM while being competitive/slightly better than ReLU networks. Notably in case of Cancer, Haberman and Pima, NLSM2 outperforms all the other models. For Iris and Banknote, we note that without any constraints ReLU1 can easily find an architecture which achieves zero test error while this is difficult for our models as we impose constraints on the architecture in order to prove global optimality.

We compare our algorithms with Batch-SGD in order to optimize (2) with batch-size being 5% of the training data while the step-size is fixed and selected between 10^{-2} and 10^2. At each iteration of our spectral method and each epoch of Batch-SGD, we compute the objective and test error of each method and show the results in Figure 2. One can see that our method is much faster than SGDs, and has a linear convergence rate. We noted in our experiments that as α is large and our data lies between $[0, 1]$, all units in the network tend to have small values that make the whole objective function relatively small. Thus, a relatively large change in (w, u) might cause only small changes in the objective function but performance may vary significantly as the distance is large in the parameter space. In other words, a small change in the objective may have been caused by a large change in the parameter space, and thus, largely influences the performance - which explains the behavior of SGDs in Figure 2.

The magnitude of the entries of the matrix A in Theorems 1 and 5 grows with the number of hidden units and thus the spectral radius $\rho(A)$ also increases with this number. As we expect that the number of required hidden units grows with the dimension of the datasets we have limited ourselves in the experiments to low-dimensional datasets. However, these bounds are likely not to be tight, so that there might be room for improvement in terms of dependency on the number of hidden units.

Acknowledgment

The authors acknowledge support by the ERC starting grant NOLEPRO 307793.

<center>19</center>

References

[1] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, New York, 1999.

[2] S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. In *ICML*, 2014.

[3] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* SIAM, Philadelphia, 1994.

[4] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Mass., 1999.

[5] A. Choromanska, M. Hena, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

[6] A Daniely, R. Frostigy, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity, 2016. arXiv:1602.05897v1.

[7] A. Gautier, F. Tudisco, and M. Hein. The Perron-Frobenius Theorem for Multi-Homogeneous Maps. in preparation, 2016.

[8] B. D. Haeffele and Rene Vidal. Global optimality in tensor factorization, deep learning, and beyond, 2015. arXiv:1506.07540v1.

[9] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, 2016.

[10] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, New York, second edition, 2013.

[11] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity:guaranteed training of neural networks using tensor methods, 2015. arXiv:1506.08473v3.

[12] W. A. Kirk and M. A. Khamsi. *An Introduction to Metric Spaces and Fixed Point Theory.* John Wiley, New York, 2001.

[13] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521, 2015.

[14] B. Lemmens and R. D. Nussbaum. *Nonlinear Perron-Frobenius theory.* Cambridge University Press, New York, general edition, 2012.

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

[16] J. Schmidhuber. Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61:85–117, 2015.

[17] J. Sima. Training a single sigmoidal neuron is hard. *Neural Computation*, 14:2709–2728, 2002.

[18] A. C. Thompson. On certain contraction mappings in a partially ordered vector space. *Proceedings of the American Mathematical Society*, 14:438–443, 1963.

20