Nonlinear Spectral Methods for Nonconvex Optimization with Global Optimality

Quynh N. Nguyen, Antoine Gautier and Matthias Hein
Department of Mathematics and Computer Science
Saarland University, Saarbrücken Informatics Campus, Germany
quynh,ag,hein@cs.uni-sb.de

Abstract

We present a method for solving a class of nonconvex optimization problems over the product of nonnegative ℓ^p spheres with global optimality guarantees and linear convergence rate. We apply our results and algorithm to training feed-forward generalized polynomial neural networks on real-world datasets.

1 Introduction and main results

Deep learning [13] is currently the state of the art machine learning technique in many application areas such as computer vision, 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 [9, 5, 4]. In particular, the problem is even for a single hidden layer in general NP hard, see [16] and references therein. This implies that to achieve global optimality certain conditions on the problem have to be imposed. 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 [11, 8]. The latter two papers are up to our knowledge the first results which provide a global method for training neural networks. However, these approaches turn out to be difficult to apply in practice. In [6], we develop a new method, namely the Nonlinear Spectral Method, for training a certain class of generalized polynomial networks with global optimality guarantees and linear convergence rate. A considerable advantage over other approaches with similar guarantees is that our conditions can be checked easily without running the algorithm. It turns out that this approach can be applied to a wider class of nonconvex optimization problems which we discuss in this paper. We present this approach from a more optimization-based point of view and provide some insights into the involved assumptions. 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]. In the experiments, we apply our theory to train a certain class of polynomial networks with one/two hidden layers as in [6]. All proofs can be found in the appendix.

Notations. Let $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x \geq 0\}$ and $\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n \mid x > 0\}$, where the inequalities are meant component wise. For $\alpha > 0$ and $z \in \mathbb{R}^n_{++}$ we write $z^\alpha = (z_1^\alpha, \ldots, z_n^\alpha)$. Let $\delta \in \mathbb{N}$, $[\delta] = \{1, \ldots, \delta\}$, $n_1, \ldots, n_\delta \in \mathbb{N}$, $V_+ = \mathbb{R}^{n_1}_+ \times \ldots \times \mathbb{R}^{n_\delta}_+$ and $V_{++} = \mathbb{R}^{n_1}_{++} \times \ldots \times \mathbb{R}^{n_\delta}_{++}$. For $p_1, \ldots, p_\delta \in (1, \infty)$ let $\|\cdot\|_{p_i}$ be the usual p_i-norm on $\mathbb{R}^{n_i}$ and $p_i' = p_i/(p_i - 1)$. Let $\rho_1, \ldots, \rho_\delta > 0$, we consider the product of nonnegative unit spheres/balls defined as $S_+ = \{x \in V_+ \mid \|x^i\|_{p_i} = \rho_i, \forall i \in [\delta]\}$, $B_+ = \{x \in V_+ \mid \|x^i\|_{p_i} \leq \rho_i, \forall i \in [\delta]\}$, and the product of positive spheres/balls defined as $S_{++} = S_+ \cap V_{++}$ and $B_{++} = B_+ \cap V_{++}$. Let

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

$F\colon V_+ \to \mathbb{R}^m$ and $i \in [\delta]$, we write $D_iF(x)$ (resp. $\nabla_i F(x)$ when $m = 1$) to denote the Jacobian (resp. the gradient) of $z^i \in \mathbb{R}^{n_i} \mapsto F(x^1, \ldots, z^i, \ldots, x^\delta)$.

In this paper, we consider the maximization of a certain class of twice-differentiable nonconvex functions $\Phi\colon V_+ \to \mathbb{R}$ over the product of nonnegative unit spheres S_+, that is,

$$\max\left\{\Phi(x^1, \ldots, x^\delta) \;\middle|\; (x^1, \ldots, x^\delta) \in S_+\right\} \tag{1}$$

Our main result derived for this problem is the following.

Theorem 1. *Let $\Phi \in C^1(B_+, \mathbb{R}) \cap C^2(B_{++}, \mathbb{R})$ be such that $\nabla\Phi(S_+) \subset V_{++}$ and there exists $M \in \mathbb{R}_+^{\delta\times\delta}$ such that $|D_j\nabla_i\Phi(x)|x^j \le M_{i,j}\nabla_i\Phi(x)$ for every $i, j \in [\delta]$ and $x \in B_{++}$. Let $A = 2\operatorname{diag}(p_1' - 1, \ldots, p_\delta' - 1)M$. If there exists $\lambda \in (0, 1)$ and $\gamma \in \mathbb{R}_{++}^d$ such that $A^T\gamma = \lambda\gamma$, then $\lambda = \rho(A)$ and (1) has a unique global maximizer $\overline{x} \in S_{++}$. Moreover, let $G\colon S_{++} \to S_{++}$,*

$$G(x) = \left(\rho_1\|\nabla_1\Phi(x)\|_{p_1'}^{1-p_1'}\big(\nabla_1\Phi(x)\big)^{p_1'-1}, \ldots, \rho_\delta\|\nabla_\delta\Phi(x)\|_{p_\delta'}^{1-p_\delta'}\big(\nabla_\delta\Phi(x)\big)^{p_\delta'-1}\right). \tag{2}$$

Then, for every $x^{(0)} \in B_{++}$, the sequence $x^{(k)} = G(x^{(k-1)})$ satisfies $\lim_{k\to\infty} x^{(k)} = \overline{x}$ and

$$\|x^{(k)} - \overline{x}\|_\infty \le \rho(A)^k\left(\frac{\mu_\gamma\big(x^{(1)}, x^{(0)}\big)}{(1-\rho(A))\min_{i\in[\delta]}\gamma_i/\rho_i}\right) \qquad \forall k \in \mathbb{N}, \tag{3}$$

where μ_γ is the weighted Thompson metric defined as $\mu_\gamma(x, y) = \sum_{j=1}^\delta \gamma_j\,\|\ln(x^j) - \ln(y^j)\|_\infty$.

Proof Strategy We first show that the global maximizer of our optimization problem (1) is attained in the "interior" of S_+, that is S_{++}. Moreover, we prove that any critical point of (1) in S_{++} is a fixed point of the mapping G. Then we proceed to show that there exists a unique fixed point of G in S_{++} and thus Φ has a unique critical point in S_{++}. As the global maximizer of (1) exists and is attained in the interior, this fixed point has to be the global maximizer. To prove that G has a unique fixed point, we first note that G maps B_{++} into B_{++} and B_{++} is a complete metric space w.r.t. the Thompson metric. Next, we provide a characterization of the Lipschitz constant of G and derive conditions under which G is a contraction. Finally, the application of Banach fixed point theorem yields the uniqueness of the fixed point of G and linear convergence rate to the global maximum of (1).

By Theorem 1, the only condition one needs to check is $\rho(A) < 1$. Unfortunately a closed form expression for $\rho(A)$ is not always available. However, since $\lim_{p_1,\ldots,p_\delta\to\infty}\rho(A) = 0$ there always exists $p_1, \ldots, p_\delta$ such that $\rho(A) < 1$ and (1) can be solved globally optimally. Such p_i's can be easily found for a given problem. Indeed, if $p_j - 1 > 2\sum_{i=1}^\delta M_{i,j}$ then $A^T\gamma < \gamma$ with $\gamma = (p_1 - 1, \ldots, p_\delta - 1)$. Hence, $\rho(A) < 1$ by Corollary 8.3.3 [10].

2 Application to generalized polynomial neural networks

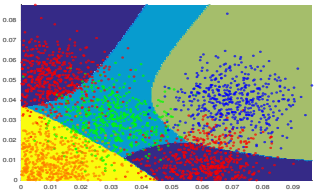

Figure 1: Classification decision boundaries in $\mathbb{R}^2$. (Best viewed in colors.)

In this section, we apply our results to neural networks for K-class classification and present the algorithm with optimality and convergence guarantees. We use the negative cross-entropy loss defined for label $y \in [K]$ and $f : \mathbb{R}^d \to \mathbb{R}^K$ as

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

One-hidden-layer network. Our function class is a feed-forward neural network with n_1 hidden units. In particular, for $\alpha \in \mathbb{R}^{n_1}$ with $\alpha_i \ge 1, i \in [n_1]$, the function class at the r^{th} output unit is defined as

$$f_r(x) = f_r(w, u)(x) = \sum_{l=1}^{n_1} w_{rl}\Big(\sum_{m=1}^d u_{lm}x_m\Big)^{\alpha_l}, \tag{4}$$

where $w \in \mathbb{R}_+^{K \times n_1}$ and $u \in \mathbb{R}_+^{n_1 \times d}$ are the parameters of the network which we optimize. We assume that $\alpha_i \neq \alpha_j$ for $i \neq j$. The reason for this lies in certain invariance properties of the network. Suppose σ is a component-wise activation function, then $\sigma(Px) = P\sigma(x)$ for every permutation matrix P. Let A, B be the optimal weight matrices, then for any permutation matrix P it holds $A\sigma(Bx) = AP^T P\sigma(Bx) = AP^T\sigma(PBx)$, 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 must hold that $A = AP^T$ and $B = PB$ for all permutation matrices P. This implies A and B have both rank one which leads to trivial classifiers. This is the reason why we have to use different activation functions for different units.

The function class in (4) can be seen as a generalized polynomial. Polynomial neural networks have been recently analyzed in [15]. Note that ReLU activation functions are meaningless in our setting as the data as well as the weights are restricted to be nonnegative. Even though the nonnegativity is a strong constraint, we show in our experiments that it can model quite complex decision boundaries (see Figure 1).

To simplify notation, let $w = (w_1, \ldots, w_K)$ where $w_i \in \mathbb{R}_+^{n_1}$ are weight vectors of K output units. All weights are normalized. In particular, our constraint set is defined as

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

We aim to solve the following optimization problem for one-hidden layer network

$$\max\big\{\Phi(w,u) \mid (w,u) \in S_+\big\} \quad \text{with} \tag{5}$$

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

where $(x^i, y_i) \in \mathbb{R}_+^d \times [K]$, $i \in [n]$ is the training data. Note that we use minus the loss in the objective to obtain a maximization problem as we want to apply our Theorem 1. Note that $\epsilon > 0$ can be chosen arbitrarily small and is added out of technical reasons.

Now, we derive the matrix $M \in \mathbb{R}_{++}^{(K+1)\times(K+1)}$ from Theorem 1 for one-hidden layer network. Let $\Psi_{p,q}^{\alpha} \colon \mathbb{R}_{++}^{n_1} \times \mathbb{R}_{++} \to \mathbb{R}_{++}$ be a function defined for every $p, q \in (1,\infty)$ and $\alpha \in \mathbb{R}_{++}^{n_1}$ as

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

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

Theorem 2. *Let Φ as in (5). Let $(x^i, y_i) \in \mathbb{R}_+^d \times [K]$, $i \in [n]$ and $p_w, p_u \in (1,\infty), \rho_w, \rho_u > 0, n_1 \in \mathbb{N}$ and $\alpha \in \mathbb{R}^{n_1}$ with $\alpha \geq \mathbf{1}$. Let $\rho_x = \max_{i \in [n]} \|x^i\|_{p_u'}$ and $M \in \mathbb{R}_{++}^{(K+1)\times(K+1)}$ with*

$$M_{w_a, w_b} = 2C_w,\ M_{w_a, u} = 2C_u + \|\alpha\|_\infty,\ M_{u, w_b} = 2C_w + 1,\ M_{u,u} = 2C_u + \|\alpha\|_\infty - 1,$$

where $C_w = \rho_w\, \Psi_{p_w', p_u}^{\alpha}(\mathbf{1}, \rho_u\rho_x)$, $C_u = \rho_w\, \Psi_{p_w', p_u}^{\alpha}(\alpha, \rho_u\rho_x)$. Then Φ and M satisfy the conditions of Theorem 1.

Two-hidden-layer network. For $\alpha \in \mathbb{R}^{n_1}, \beta \in \mathbb{R}^{n_2}$ $(\alpha, \beta \geq \mathbf{1})$ our new function class is: $f_r(x) = f_r(w,v,u)(x) = \sum_{b=1}^{n_2} w_{rb}\Big(\sum_{a=1}^{n_1} v_{ba}\Big(\sum_{s=1}^{d} u_{as} x_s\Big)^{\alpha_a}\Big)^{\beta_b}$. Let $V_+ = \mathbb{R}_+^{K \times n_2} \times \mathbb{R}_+^{n_2 \times n_1} \times \mathbb{R}_+^{n_1 \times d}$, the problem becomes

$$\max\big\{\Phi(w,v,u) \;|(w,v,u) \in V_+, \|w_i\|_{p_w} = \rho_w,\ \|v\|_{p_v} = \rho_v, \|u\|_{p_u} = \rho_u\big\} \tag{6}$$

where

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

The application of Theorem 1 now yields the following result for 2-hidden layers.

Theorem 3. *Let Φ as in (6). Let $(x^i, y_i) \in \mathbb{R}_+^d \times [K]$, $i \in [n]$ and $p_w, p_v, p_u \in (1,\infty)$, $\rho_w, \rho_v, \rho_u > 0$, $n_1, n_2 \in \mathbb{N}$ and $\alpha \in \mathbb{R}^{n_1}, \beta \in \mathbb{R}^{n_2}$ with $\alpha, \beta \geq \mathbf{1}$. Let $M \in \mathbb{R}_{++}^{(K+2)\times(K+2)}$ with*

$$\begin{array}{lll}
M_{w_a,w_b} = 2C_w, & M_{w_a,v} = 2C_v + \|\beta\|_\infty, & M_{w_a,u} = 2C_u + \|\alpha\|_\infty\|\beta\|_\infty \\
M_{v,w_a} = 2C_w + 1, & M_{v,v} = 2C_v + \|\beta\|_\infty - 1, & M_{v,u} = 2C_u + \|\alpha\|_\infty\|\beta\|_\infty \\
M_{u,w_a} = 2C_w + 1, & M_{u,v} = 2C_v + \|\beta\|_\infty, & M_{u,u} = 2C_u + \|\alpha\|_\infty\|\beta\|_\infty - 1
\end{array}$$

for every $w_a, w_b \in [K], v = K+1, u = K+2$, *and where*
$\theta = \rho_v \Psi^{\alpha}_{p'_v,p_u}(\mathbf{1}, \rho_u \rho_x)$, $C_w = \rho_w \Psi^{\beta}_{p'_w,p_v}(\mathbf{1}, \theta)$, $C_v = \rho_w \Psi^{\beta}_{p'_w,p_v}(\beta, \theta)$, $C_u = \|\alpha\|_\infty C_v$.
Then M *and* Φ *satisfies the conditions of Theorem 1.*

3 Experiments

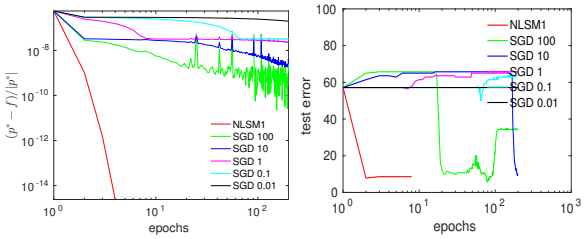

Figure 2: Training score (left) w.r.t. the optimal score p^* and test error (right) of NLSM1 and Batch-SGD with different step-sizes.

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

The shown experiments should be seen as a proof of concept. We do not have yet a good understanding of how to 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 $(n_1, \alpha, \rho_w, \rho_u) \in [2, 20] \times [1, 4] \times (0, 1]^2$ (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 pick p_u, p_w (resp. p_u, p_v, p_w) such that every generated model satisfies the conditions of Theorem 1 (resp. Theorem 3), i.e. $\rho(A) < 1$. Thus, global optimality is guaranteed in all our experiments. For comparison, we use RBF-kernel SVM, a one-hidden-layer rectified linear network (ReLU1) and a two-hidden-layers ReLu network (ReLU2). For training ReLU, we use stochastic gradient descent with logistic loss and L_2-norm regularization to prevent over-fitting. All parameters are jointly cross validated for each method. Specifically, for ReLU the number of hidden units is varied from 2 to 20, step-sizes and regularizers are taken from $\{10^{-6}, 10^{-5}, \ldots, 10^2\}$ 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 chosen from $\{2^{-5}, 2^{-4} \ldots, 2^{20}\}$ and $\{2^{-15}, 2^{-14} \ldots, 2^3\}$ respectively. Note that RBF-SVM and ReLUs allow negative weights while our models do not. The results from Table 1 show that our nonlinear spectral methods achieve overall competitive performance to other methods. In case of Cancer, Haberman and Pima, NLSM2 outperforms all other models. For Iris and Banknote, we note that ReLU1 (without any constraint) can achieve zero test error while this is difficult for our nonlinear spectral methods since we impose constraints on the architecture in order to prove global optimality.

We compare our algorithms against Batch-SGD (batch-size $\approx 0.05n$) with different fixed step-sizes. At each iteration (epoch) of our spectral method (Batch-SGD), we compute the training objective and test error. Figure 2 shows 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, which makes training objective function become relatively small. Thus, a small change in the objective can be caused by a relatively large change in parameter space which eventually leads to large influence on performance. This somehow explains the behavior of SGDs in Figure 2 .

The magnitude of the entries of the matrix A in Theorems 2 and 3 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.

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

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

[5] 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.

[6] A. Gautier, Q. Nguyen, and M. Hein. Globally optimal training of generalized polynomial neural networks with nonlinear spectral methods. *NIPS 2016*.

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

[8] B. D. Haeffele and R. 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, 2015. arXiv:1509.01240v2.

[10] R. Horn and C. Johnson, editors. *Matrix Analysis*. Cambridge University Press, 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. Nussbaum. *Nonlinear Perron-Frobenius theory*. Cambridge University Press, general edition.

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

A Proof of Theorem 1

The assumption $\nabla\Phi(S_+) \subset V_{++}$ guarantees that Φ attains its global maximum in S_{++}.

Lemma 1. *Let $\Phi \in C^1(B_+, \mathbb{R})$ and suppose that $\nabla\Phi(S_+) \subset V_{++}$. 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 $(x^1, \dots, x^\delta)$ of S_+ is characterized by the fact that at least one of the variables $x^1, \dots, x^\delta$ has a zero component. Suppose w.l.o.g. that the subset $J \subset [n_1]$ of components of $x^1 \in \mathbb{R}^{n_1}_+$ are zero, that is $x^1_J = 0$. The normal vector of the p_1-sphere at x^1 is given by $\nu = (x^1)^{p_1 - 1}$. 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 $(x^1, \dots, x^\delta)$ is a local maximum, then

$$\left\langle \nabla_1 \Phi(x^1, \dots, x^\delta), t \right\rangle \leq 0, \quad \forall t \in T_+ = \{v \in \mathbb{R}^{n_1}_+ \mid \langle \nu, v \rangle = 0\}, \tag{7}$$

where T_+ is the set of "positive" tangent directions, that are pointing inside the set $\{x^1 \in \mathbb{R}^{n_1}_+ \mid \|x^1\|_{p_1} = \rho_1\}$. Otherwise there would exist a direction of ascent which leads to a feasible point. Now note that ν has non-negative components as $x^1 \in \mathbb{R}^{n_1}_+$. Thus

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

However, by assumption $\nabla_1 \Phi(x^1, \ldots, x^\delta)$ is a vector with strictly positive components and thus (7) 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 in S_{++}.

Lemma 2. *Let $\Phi \in C^1(B_+, \mathbb{R})$ and suppose that $\nabla\Phi(S_+) \subset V_{++}$. Then x is a critical point of Φ in S_{++} if and only if it is a fixed point of G.*

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

$$\mathcal{L}(x, \xi) = \Phi(x) - \sum_{j=1}^{\delta} \xi_j (\rho_j - \|x^j\|_{p_j}) \qquad \forall x \in B_+,\ \xi \in \mathbb{R}^d.$$

From $\nabla\mathcal{L}(x, \xi) = 0$, one can easily show that the necessary condition [3] for $x = (x^1, \ldots, x^\delta) \in S_{++}$ being a critical point of Φ is the existence of $\lambda_1, \ldots, \lambda_\delta$ such that

$$\nabla_j \Phi(x) = \lambda_j (x^j)^{p_j - 1} \qquad \forall j \in [\delta]. \tag{8}$$

Since $x^j > 0$ and $\nabla_j \Phi(x) > 0$ by our assumption, it must hold that $\lambda_i > 0$ for every $i \in [\delta]$. By noting that $(z^{p-1})^{p'-1} = z$ for $z \geq 0$, we get

$$\nabla_j \Phi(x)^{p_j' - 1} = \lambda_j^{p_j' - 1} x^j \qquad \forall j \in [\delta]. \tag{9}$$

Diving both sides of (9) by the norms, one obtains

$$\frac{\nabla_j \Phi(x)^{p_j' - 1}}{\|\nabla_j \Phi(x)^{p_j' - 1}\|_{p_j}} = \frac{\lambda_j^{p_j' - 1} x^j}{\lambda_j^{p_j' - 1} \|x^j\|_{p_j}} = \frac{x^j}{\rho_j} \qquad \forall j \in [\delta]$$

Combining with the fact that $\|z^{p'-1}\|_p = \|z\|_{p'}^{p'-1}$ for $z > 0$, we get

$$(x^1, \ldots, x^\delta) = \left(\rho_1 \frac{\nabla_1 \Phi(x)^{p_1' - 1}}{\|\nabla_1 \Phi(x)\|_{p_1'}^{p_1' - 1}}, \ldots, \rho_\delta \frac{\nabla_\delta \Phi(x)^{p_\delta' - 1}}{\|\nabla_\delta \Phi(x)\|_{p_\delta'}^{p_\delta' - 1}} \right) = G(x) \in S_{++},$$

as the gradient is strictly positive on S_{++} and thus the mapping G from (2) is well-defined. We have proved that if x is a critical point then $G(x) = x$. Assume now that $G(x) = x$, then

$$\rho_j \frac{\nabla_j \Phi(x)^{p_j' - 1}}{\|\nabla_j \Phi(x)\|_{p_j'}^{p_j' - 1}} = x^j, \qquad \forall j \in [\delta]$$

and thus there exists $\lambda_j = \rho_j^{-1} \|\nabla_j \Phi(x)\|_{p_j'}^{p_j' - 1}$, $j \in [\delta]$ such that (9) holds implying that x is a critical point of Φ in S_{++}. □

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

Theorem 4 (Banach fixed point theorem, e.g. Theorem 3.1 [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 $\overline{x}$ in X, that is $T(\overline{x}) = \overline{x}$ and the sequence defined as $x^{(n+1)} = T(x^{(n)})$ with $x^{(0)} \in X$ converges $\lim_{n\to\infty} x^{(n)} = \overline{x}$ with linear convergence rate*

$$d(x^{(n)}, \overline{x}) \ \leq\ \frac{q^n}{1 - q}\, d(x^{(1)}, x^{(0)}). \tag{10}$$

The following lemma shows that (B_{++}, μ_γ) is a complete metric space.

Lemma 3. *(B_{++}, μ_γ) is a complete metric space for every $\gamma \in \mathbb{R}^\delta_{++}$.*

Proof. We first prove that for $p \in (1,\infty)$ and $\rho > 0$, $\big(\{z \in \mathbb{R}^n_{++} \mid \|z\|_p \le \rho\}, d\big)$ is a complete metric space. Let $(z^k)_k \subset \{z \in \mathbb{R}^n_{++} \mid \|z\|_p \le \rho\}$ be a Cauchy sequence w.r.t. 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, meaning $\lim_{k\to\infty} z^k = z^*$ w.r.t. the norm topology. Finally, since $z \mapsto \|z\|_p$ is a continuous function, we get $\|z^*\|_p = \lim_{k\to\infty} \|z^k\|_p \le \rho$, i.e. $z^* \in \{z \in \mathbb{R}^n_{++} \mid \|z\|_p \le 1\}$ which proves our claim.

Now, the idea is to see B_{++} as a product of such metric spaces. For $i \in [\delta]$, let $B^i_{++} = \{x^i \in \mathbb{R}^{n_i}_{++} \mid \|x^i\|_{p_i} \le \rho_i\}$ and $d_i(x^i, \tilde{x}^i) = \gamma_i \|\ln(x^i) - \ln(\tilde{x}^i)\|_\infty$ for some constant $\gamma_i > 0$. Then (B^i_{++}, d_i) is a complete metric space for every $i \in [\delta]$ and $B_{++} = B^1_{++} \times \ldots \times B^\delta_{++}$. It follows that (B_{++}, μ_γ) is a complete metric space with $\mu_\gamma \colon B_{++} \times B_{++} \to \mathbb{R}_+$ defined as

$$\mu_\gamma\big(x, \tilde{x}\big) = \sum_{i=1}^{\delta} \gamma_i \|\ln(x^i) - \ln(\tilde{x}^i)\|_\infty. \qquad \square$$

Suppose that $\rho(A)$ is a Lipschitz constant of G w.r.t. μ_γ, i.e.

$$\mu_\gamma(G(x), G(y)) \le \rho(A)\mu_\gamma(x, y) \qquad \forall x, y \in B_{++}. \tag{11}$$

If $\rho(A) < 1$, then, as (B_{++}, μ_γ) is a complete metric space by Lemma 3, we can apply the Banach fixed point theorem 4 to G. As an implication, G has a unique fixed point in S_{++} and therefore Φ has a unique maximizer $\overline{x} \in S_{++}$. Moreover, the sequence $(x^{(k)})_{k\in\mathbb{N}} \subset S_{++}$ converges to $\overline{x}$ for every starting point $x^{(0)} \in B_{++}$ and the linear convergence rate (10) holds for every $k \in \mathbb{N}$.

To prove (11), we first prove two lemmas. Lemma 4 establishes the connection between the matrices M and A defined in Theorem 1. Meanwhile, Lemma 5 shows how to use the property of these matrices in order to get a Lipschitz constant for G w.r.t. μ_γ.

Lemma 4. *Let $F \in C^1(\mathbb{R}^n_{++}, \mathbb{R}^m_{++})$, $p' \in (1,\infty)$ and $H(x) = \|F(x)\|_{p'}^{1-p'} F(x)^{p'-1}$. If $c \ge 0$ and $x > 0$ satisfy $|DF(x)|x \le cF(x)$, then we have $|DH(x)|x \le 2c(p'-1)H(x)$.*

Proof. Let $F = (F_1, \ldots, F_m)$, $H = (H_1, \ldots, H_m)$ and $l \in [m]$. It holds

$$\begin{aligned}\nabla H_l(x) &= (p'-1)\left(\frac{F_l(x)^{p'-2}\nabla F_l(x)}{\|F(x)\|_{p'}^{p'-1}} - \frac{F_l(x)^{p'-1}\|F(x)\|_{p'}^{-1}\sum_{k=1}^m F_k(x)^{p'-1}\nabla F_k(x)}{\|F(x)\|_{p'}^{2p'-2}}\right)\\ &= (p'-1)H_l(x)\left(\frac{\nabla F_l(x)}{F_l(x)} - \frac{\sum_{k=1}^m F_k(x)^{p'-1}\nabla F_k(x)}{\|F(x)\|_{p'}^{p'}}\right).\end{aligned}$$

Using the triangle inequality and $|DF(x)|x \le cF(x)$ one gets

$$\frac{\langle|\nabla H_l(x)|, x\rangle}{H_l(x)} \le (p'-1)\Big(\frac{\langle|\nabla F_l(x)|, x\rangle}{F_l(x)} + \frac{\sum_{i=1}^m F_i(x)^{p'-1}\langle|\nabla F_i(x)|, x\rangle}{\|F(x)\|_{p'}^{p'}}\Big) \le 2(p'-1)c.$$

As this is true for every $l \in [m]$, we get $|DH(x)|x \le 2c(p'-1)H(x)$. $\square$

Lemma 5. *Let $F \in C^1(B_{++}, B_{++})$, $F = (F_1, \ldots, F_\delta)$ with $F_i \colon B_{++} \to \mathbb{R}^{n_i}$, and $Q \in \mathbb{R}^{\delta\times\delta}_+$ be such that $|D_j F_i(x)|x^j \le Q_{i,j}F_i(x)$ for every $x \in B_{++}$ and every $i, j \in [\delta]$. Then, for any $\gamma \in \mathbb{R}^\delta_{++}$, we have*

$$\mu_\gamma(F(x), F(y)) \le C\mu_\gamma(x, y) \quad \forall x, y \in B_{++} \qquad \textit{with} \qquad C = \max_{i\in[\delta]} \frac{(Q^T\gamma)_i}{\gamma_i}.$$

Proof. Let $i \in [\delta]$, $F_i = (F_{i,1}, \ldots, F_{i,n_i})$, $l_i \in [n_i]$ and $f(x) = \ln(F_{i,l_i}(\exp(x)))$ where $\exp(\cdot)$ and $\ln(\cdot)$ are taken component wise. Let $x, y \in B_{++}, x \ne y$. Set $\tilde{x} = \ln(x)$ and $\tilde{y} = \ln(y)$. By the mean value theorem, there exists $t \in (0,1)$ such that $f(\tilde{x}) - f(\tilde{y}) = \langle\nabla f(\tilde{z}), \tilde{x} - \tilde{y}\rangle$ where $\tilde{z} = t\tilde{x} + (1-t)\tilde{y}$. Note that $z = \exp(\tilde{z}) \in B_{++}$ because the exponential is convex and $x, y \in B_{++}$, indeed for every $i \in [\delta]$, we have

$$\|z^i\|_{p_i} = \|\exp\big(t\ln(x^i) + (1-t)\ln(y^i)\big)\|_{p_i} \le \|tx^i + (1-t)y^i\| \le t\|x^i\|_{p_i} + (1-t)\|y^i\|_{p_i} \le \rho_i.$$

Hence, with the Hölder inequality, we get

$$\begin{aligned}|\ln(F_{i,l_i}(x)) - \ln(F_{i,l_i}(y))| &= \frac{|\langle \nabla F_{i,l_i}(z) \circ z, \tilde{x} - \tilde{y}\rangle|}{F_{i,l_i}(z)} \\ &\leq \sum_{k=1}^{\delta} \frac{|\left\langle \nabla_k F_{i,l_i}(z) \circ z^k, \tilde{x}^k - \tilde{y}^k\right\rangle|}{F_{i,l_i}(z)} \\ &\leq \sum_{k=1}^{\delta} \frac{\left\langle |\nabla_k F_{i,l_i}(z)|, z^k\right\rangle}{F_{i,l_i}(z)} \|\tilde{x}^k - \tilde{y}^k\|_\infty \\ &\leq \sum_{k=1}^{\delta} Q_{i,k} \|\ln(x^k) - \ln(y^k)\|_\infty\end{aligned}$$

where $u \circ v = (u_1 v_1, \ldots, u_n v_n)$ for all $u, v \in \mathbb{R}^n$. Taking the maximum over $l_i \in [n_i]$ shows that

$$\|\ln(F_i(x)) - \ln(F_i(y))\|_\infty \leq \sum_{k=1}^{\delta} Q_{i,k} \|\ln(x^k) - \ln(y^k)\|_\infty.$$

It follows that

$$\begin{aligned}\mu_\gamma(F(x), F(y)) &\leq \sum_{i=1}^{\delta}\sum_{j=1}^{\delta} \gamma_i Q_{i,j} \|\ln(x^j) - \ln(y^j)\|_\infty = \sum_{j=1}^{\delta} (Q^T\gamma)_j \|\ln(x^j) - \ln(y^j)\|_\infty \\ &= \sum_{j=1}^{\delta} \frac{(Q^T\gamma)_j}{\gamma_j} \gamma_j \|\ln(x^j) - \ln(y^j)\|_\infty \leq \Big(\max_{i\in[\delta]} \frac{(Q^T\gamma)_i}{\gamma_i}\Big) \mu_\gamma(x, y)\end{aligned}$$

which proves the claim. □

We have now all the tools to conclude the proof of Theorem 1. Indeed, the assumption

$$|D_j \nabla_i \Phi(x)| x^j \leq M_{i,j} \nabla_i \Phi(x) \qquad \forall i, j \in [\delta],\ x \in B_{++}$$

and Lemma 4 (with $F(x) = \nabla_i \Phi(x)$, $i \in [\delta]$) imply that

$$|D_j G_i(x)| x^j \leq 2(p_i' - 1) M_{i,j}\, G_i(x) = A_{i,j}\, G_i(x) \qquad \forall i, j \in [\delta],\ x \in B_{++},$$

where $G_i(x) = \|\nabla_i \Phi(x)\|_{p_i'}^{1-p_i'} \nabla_i \Phi(x)^{p_i'-1}$ for $i \in [\delta]$. Now, Lemma 5 implies

$$\mu_\gamma(G(x), G(y)) \leq C \mu_\gamma(x, y) \quad \forall x, y \in B_{++} \qquad \text{with} \qquad C = \max_{i\in[\delta]} \frac{(Q^T\gamma)_i}{\gamma_i}.$$

We have $A^T\gamma = \lambda\gamma$, and thus $C = \lambda$. Theorem 8.3.4 [10] implies that $\lambda = \rho(A)$ and, as $\rho(A) < 1$, G is a strict contraction. Thus, we can apply the Banach fixed point theorem 4 to G. It follows that G has a unique fixed point $\overline{x} \in S_{++}$ which is also the global maximizer of Φ on S_+ by Lemmas 1 and 2. Moreover, it holds

$$\mu_\gamma(x^{(k)}, \overline{x}) \leq \frac{\rho(A)^k}{1 - \rho(A)}\, \mu_\gamma(x^{(1)}, x^{(0)}) \qquad k = 1, 2, \ldots \tag{12}$$

Now, let $k \in \mathbb{N}$ be fixed and $z = x^{(k)}$. The mean value theorem implies that for every $r \in \mathbb{R}$, we have

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

In particular, we have

$$\ln\left(z_{j_i}^i\right) \in (-\infty, \ln(\rho_i)] \qquad \forall i \in [\delta], j_i \in [n_i]$$

It follows that

$$\begin{aligned}\mu\big(z, \overline{x}\big) &= \sum_{i=1}^{\delta} \gamma_i \|\ln(z^i) - \ln(\overline{x}^i)\|_\infty \geq \sum_{i=1}^{\delta} \frac{\gamma_i}{\rho_i} \|z^i - \overline{x}^i\|_\infty \\ &\geq \max_{i\in[\delta]} \frac{\gamma_i}{\rho_i} \|z^i - \overline{x}^i\|_\infty \geq \|z - \overline{x}\|_\infty \Big(\min_{i\in[\delta]} \frac{\gamma_i}{\rho_i}\Big)\end{aligned}$$

and thus

$$\|z - \overline{x}\|_\infty \leq \frac{\mu\big(z, \overline{x}\big)}{\min_{i\in[\delta]} \frac{\gamma_i}{\rho_i}} \leq \rho(A)^k \left(\frac{\mu\big(x^{(1)}, x^{(0)}\big)}{(1 - \rho(A)) \min_{i\in[\delta]} \frac{\gamma_i}{\rho_i}}\right).$$

B Proof of Theorem 2

Proof. Similar to the proof of Theorem 4 in [6] where the matrix M is referred to as Q. □

C Proof of Theorem 3

Proof. Similar to the proof of Theorem 5 in [6]. □