Global Optimality in Neural Network Training

Benjamin D. Haeffele and René Vidal
Johns Hopkins University, Center for Imaging Science, Baltimore, MD 21218, USA.
bhaeffele@jhu.edu rvidal@cis.jhu.edu

Abstract

The past few years have seen a dramatic increase in the performance of recognition systems thanks to the introduction of deep networks for representation learning. However, the mathematical reasons for this success remain elusive. A key issue is that the neural network training problem is nonconvex, hence optimization algorithms may not return a global minima. This paper provides sufficient conditions to guarantee that local minima are globally optimal and that a local descent strategy can reach a global minima from any initialization. Our conditions require both the network output and the regularization to be positively homogeneous *functions of the network parameters, with the regularization being designed to control the* network size. *Our results apply to networks with one hidden layer, where size is measured by the number of neurons in the hidden layer, and multiple deep subnetworks connected in parallel, where size is measured by the number of subnetworks.*

1. Introduction

As a broad definition, feed-forward deep networks are a collection of feature extraction layers, where each layer applies some form of linear transformation (e.g., convolution, dot-product), followed by a nonlinearity (e.g., rectification, max-pooling) to the output of the preceding layer. Modern networks are similar to classical neural networks, except that they involve many more layers and typically replace classical sigmoid nonlinearities by linear rectification (i.e., ReLU units). These modifications, together with the availability of massive amounts of data for training, have led to dramatic improvements in classification performance for various applications in computer vision, speech and natural language processing. However, the mathematical reasons for this success remain elusive.

An important mathematical challenge is that the problem of learning the parameters of a deep network is non-convex, hence optimization algorithms can get stuck in non-global minima. In contrast, local minimizers of a convex optimization problem are also global minimizers, so convex formulations of learning problems are often preferable as they facilitate the analysis of the properties of the learning algorithm. This is one of the reasons for the popularity of classical learning algorithms such as linear regression, support vector machines, ℓ_1 minimization, and nuclear norm minimization, all of which involve solving a *convex optimization problem* of the form:

$$\min_X \ell(Y, \Phi(X, S)) + \lambda\Theta(X). \tag{1}$$

For classification problems, $\ell(Y, \Phi(X, S))$ is a *loss function* that measures the agreement between the true labels, Y, and the predicted labels, $\Phi(X, S)$, where S is the input data to be classified and X represents the classifier parameters, while $\Theta(X)$ is a *regularization function* designed to prevent overfitting. Convex formulations require both the loss function and the regularization function to be convex on X, e.g., $\ell(Y, \Phi(X, S)) = \|Y - S^\top X\|_F^2$ and $\Theta(X) = \|X\|_F^2$.

Unfortunately, in practice many learning algorithms – and particularly those that seek to learn an appropriate representation of features directly from the data, such as principal component analysis (PCA), low-rank matrix completion, nonnegative matrix factorization, sparse dictionary learning, tensor factorization and deep learning – involve solving a *non-convex optimization problem* such as:

$$\min_{\{W^i\}_{i=1}^K} \ell(Y, \Phi(W^1, \dots, W^K)) + \lambda\Theta(W^1, \dots, W^K), \tag{2}$$

where Φ is an arbitrary, convexity destroying mapping.[1] For example, in deep neural network training, the output of the network is typically generated by applying an alternating series of linear and non-linear functions to the input data:

$$\begin{aligned}\Phi(W^1, \dots, W^K) = \psi_K(\psi_{K-1}(\cdots \\ \psi_2(\psi_1(S^\top W^1)W^2)\cdots W^{K-1})W^K),\end{aligned} \tag{3}$$

where each W^i is an appropriately sized matrix that contains the connection weights between layers $i-1$ and i of the network, and the $\psi_i(\cdot)$ functions apply some form of

[1] For the sake of notational simplicity, we will omit the dependency of Φ on the data, S, from now on.

non-linearity after each matrix multiplication, e.g., a sigmoid function, rectification, max-pooling.[2]

For a very small number of non-convex problems, e.g., PCA, one is fortunate, and a global minimizer can be found in closed form. For other problems, e.g., ℓ_0 minimization, rank minimization, and low-rank matrix completion, one can replace the non-convex objective by a convex surrogate and show that under certain conditions the solutions to both problems are the same [9, 5]. In most cases, however, the optimal solutions cannot be computed in closed form, and a good convex surrogate may not be easy to find. This presents significant challenges to existing optimization algorithms – including (but certainly not limited to) alternating minimization, gradient descent, stochastic gradient descent, block coordinate descent, back-propagation, and quasi-Newton methods – which are typically only guaranteed to converge to a critical point of the objective function [16, 20, 25, 26]. However, for non-convex problems, the set of critical points includes not only global minima, but also local minima, local maxima, saddle points and saddle plateaus, as illustrated in Figure 1. As a result, the non-convexity of the problem leaves the model somewhat ill-posed in the sense that it is not just the model formulation that is important but also implementation details, such as how the model is initialized and particulars of the optimization algorithm, which can have a significant impact on the performance of the model.

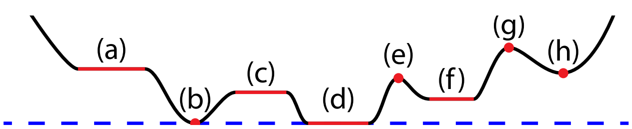

Figure 1. Example critical points of a non-convex function (shown in red). (a,c) Plateaus. (b,d) Global minima. (e,g) Local maxima. (f,h) Local minima.

To address the issue of non-convexity, a common strategy used in deep learning is to initialize the network weights, $\{W^k\}$, at random, update these weights using local descent, check if the training error decreases sufficiently fast, and if not, choose another initialization. In practice, it has been observed that if the size of the network is large enough, this strategy can lead to markedly different solutions for the network weights, which give nearly the same objective values and classification performance [6]. It has also been observed that when the size of the network is large enough and the nonlinearity is chosen to be a Rectified Linear Unit (ReLU), $\psi^+(x) = \max(x, 0)$, in lieu of a sigmoid function, many weights are zero, a phenomenon known as *dead neurons*, and the classification performance significantly improves [7, 15, 14, 27]. While this empirically suggests that when the size of the network is large enough and ReLU nonlinearities are used *all local minima could be global*, there is currently no rigorous theory that provides a precise mathematical explanation for these experimentally observed phenomena.

[2] Here we have shown the linear operations to be simple matrix multiplications to simplify notation, but this easily generalizes to other linear operators (e.g., convolution).

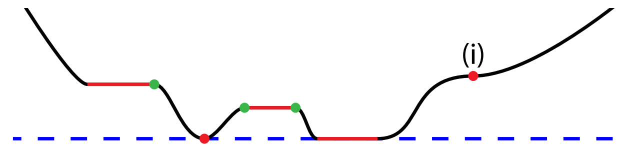

Figure 2. Guaranteed properties of the proposed framework. Starting from any initialization, a non-increasing path exists to a global minimizer. Starting from points on a plateau, a simple "sliding" method exists to find the edge of the plateau (green points).

Paper Contributions. In this paper, we study conditions under which the optimization landscape for the non-convex optimization problem in (2) is such that *all critical points are either global minimizers or saddle points/plateaus*, as shown in Figure 2. We show that if the network size is large enough and the functions Φ and Θ are *sums of positively homogeneous functions of the same degree*, any local minimizer such that *some of its entries are zero* is also a global minimizer. Interestingly, ReLU and max-pooling nonlinearities are positively homogeneous, while sigmoids are not, which could provide a possible explanation for the improved performance of ReLU and max pooling. Furthermore, many state-of-the-art networks are not trained with *classical regularization*, such as an ℓ_1 or ℓ_2 norm penalty on the weight parameters but instead rely on techniques such as dropout. Our results also provide strong guidance on the design of network regularization to ensure the non-existence of spurious local minima, showing that traditional weight decay is not appropriate for deep networks. However, more recently proposed forms of regularization such as Path-SGD [18] or batch normalization [12] can be easily incorporated into our analysis framework, and stochastic regularization methods, such as a dropout [23], also have strong similarities to our framework.

Related Work. Prior work on global optimality of neural network training [3] showed that for neural networks with a single hidden layer, if the number of neurons in the hidden layer is not fixed, but instead fit to the data through a sparsity inducing regularization, then the process of training a globally optimal neural network is analogous to selecting a finite number of hidden units from a potentially infinite dimensional space of all possible hidden units. A weighted sum of the selected hidden units is then taken to produce the output. The specific optimization problem is of the form

$$\min_x \ell(Y, \sum_i h_i(S)x_i) + \lambda \|x\|_1, \tag{4}$$

where $h_i(S)$ represents one possible hidden unit activation in response to the training data S from an infinite dimensional space $h_i(S) \in \mathcal{H}$ of all possible hidden unit activations. Clearly (4) is a convex optimization problem on x (assuming $\ell(Y, X)$ is convex on X) and straightforward to solve for a finite set of $h_i(S)$ activations. However, since $\mathcal{H}$ is an infinite dimensional space the primary difficulty lies in how to select the appropriate hidden unit activations. Nonetheless, by using arguments from gradient boosting, it is possible to show that problem (4) can be globally optimized by sequentially adding hidden units to the network until one can no longer find a hidden unit whose addition will decrease the objective function [3, 10, 17].

Several recent works have also explored the error surface of multilayer neural networks using tools derived from random matrix theory and statistical physics. Applying ideas from random matrix theory to high-dimensional nonconvex optimization, the authors of [8] argue that, under certain assumptions, for high-dimensional optimization problems if one is given a particular critical point, it is vastly more likely that the critical point will be a saddle point rather than a local minimizer and thus avoiding saddle points is the key difficulty in high-dimensional, non-convex optimization. Using arguments from statistical physics, the authors of [6] show that, under certain assumed distributions of the training data and the network weight parameters, as the number of hidden units in a network increases, the distribution of local minima becomes increasingly concentrated in a small band of objective function values near the global optimum (and thus all local minima become increasingly close to being global minima).

Additional recent work has analyzed the problem of training neural networks with a single hidden layer by estimating high order statistical moments of the network mapping using tensor decomposition methods and show that, with sufficient assumptions on the loss and data distribution, polynomial-time training is possible [13]. Further, the authors of [21] study the problem of when a given initialization of a neural network is likely to be within the basin of attraction of a global minimizer and provide conditions that ensure a random initialization will be within the basin of a global minimizer with high probability.

Our results will largely echo ideas from the above work, but we take a markedly different approach. Specifically, we will analyze the problem using a purely deterministic approach which does not make any assumptions regarding the distribution of the input data, the network weight parameter statistics, or the network initialization. With this approach, we will show that saddle points and plateaus are the *only* critical points that one needs to be concerned with due to the fact that for networks of sufficient size, local minima that require one to climb the objective surface to escape from, such as (f) and (h) in Figure 1, are guaranteed not to exist, as illustrated in Figure 2.

2. Problem Formulation

In this paper, we will study the non-convex problem:

$$\min_{r \in \mathbb{N}^+} \min_{\{W^i\}_{i=1}^K} \ell(Y, \Phi_r(W^1, \dots, W^K)) + \lambda \Theta_r(W^1, \dots, W^K). \tag{5}$$

For classification problems, Y typically contains the labels of the training examples, but in general this could be any arbitrary set of desired network outputs. K describes the number of weight layers in the network; the $\{W^k\}$ variables describe the parameters of different layers; and Φ_r defines the output of the network as a function of the network parameters. As an example, $K = 2$ is a network with one hidden layer where W^1 defines the weights from the input to the hidden layer, and W^2 defines the weights from the hidden layer to the output (Figure 3, left). The integer r defines the *size* of the network. For example, r could be the number of neurons in the hidden layer of a two-layer neural network (Figure 3, left), or the number of parallel sub-networks in a deep network (Figure 3, right). Note that a key feature of our formulation is that we optimize over the size of the network r, so as a result we need a means to control the network size and prevent overfitting. This is accomplished through a *regularization function* $\Theta_r(W^1, \dots, W^K)$, while the *loss function* ℓ measures how well Y is approximated by the network output, $\Phi_r(W^1, \dots, W^K)$, and $\lambda > 0$ balances the trade-off between regularization and loss.

3. Motivation: Matrix Factorization

Before considering neural networks, as a motivating example consider the following matrix factorization problem: Given a matrix $Y \in \mathbb{R}^{d_1 \times d_2}$, find factors $W^1 \in \mathbb{R}^{d_1 \times r}$ and $W^2 \in \mathbb{R}^{d_2 \times r}$ of small size, r, and small Frobenius norm that approximate Y as $W^1 W^{2\top}$, i.e.:

$$\min_{r \in \mathbb{N}^+} \min_{W^1, W^2} \frac{1}{2}\|Y - W^1 W^{2\top}\|_F^2 + \frac{\lambda}{2}(\|W^1\|_F^2 + \|W^2\|_F^2). \tag{6}$$

Notice that this problem is a particular case of (5) where $K = 2$; the loss function is chosen as the squared loss $\ell(Y, X) = \frac{1}{2}\|Y - X\|_F^2$; the factorization map Φ_r in (3) reduces to matrix multiplication - i.e., $\Phi_r(W^1, W^2) = W^1 W^{2\top}$, $S = I$, $\psi_i(Z) = Z$; and the regularization function is chosen as ℓ_2 (Tykhonov) regularization, i.e., $\Theta(W^1, W^2) = \frac{1}{2}(\|W^1\|_F^2 + \|W^2\|_F^2)$.

For the matrix factorization problem above, the regularizer is convex on (W^1, W^2), but the overall objective is not due to the product $W^1 W^{2\top}$. While this makes the optimization problem in (6) non-convex, we can still analyze it by recalling the variational form of the nuclear norm $\|X\|_*$ (the sum of the singular values of the matrix X), which is given by [22]:

$$\|X\|_* = \min_r \min_{W^1,W^2:W^1W^{2\top}=X} \frac{1}{2}(\|W^1\|_F^2+\|W^2\|_F^2). \quad (7)$$

The strong similarity between (7) and the regularizer in (6) suggests looking at the *convex* problem:

$$\min_X \frac{1}{2}\|Y-X\|_F^2 + \lambda\|X\|_*, \quad (8)$$

which is a classical nuclear norm minimization problem, whose solution can be found in closed form from the SVD of Y. Additionally, well-known results from positive semidefinite optimization have shown that although (6) is a non-convex optimization problem, all local minima of (6) will be globally optimal provided r is initialized to be sufficiently large [4, 2, 19, 11]. However, while for the particular case of (6) the problem can be easily recast as a semi-definite optimization problem [19], for alternative choices of regularization functions, Θ, this quickly becomes a non-trivial problem. Further, results from semi-definite optimization apply to problems of the form

$$\min_{r\in\mathbb{N}^+} \min_{W^1,W^2} \ell(Y, W^1W^{2\top}) + \lambda\bar{\Theta}(W^1W^{2\top}) \quad (9)$$

which are subtly (but critically) different from the problem we consider here, as we consider regularization on the factors directly, $\Theta(W^1, W^2)$, instead of on the product of the factors, $\bar{\Theta}(W^1W^{2\top})$, resulting in problems that are typically considerably more challenging [1].

Nevertheless, we build on these ideas from matrix factorization to analyze a wide range of non-convex optimization problems, including neural network training and significant generalizations of the matrix factorization problem given in (6), and present a framework where a convex optimization problem, e.g., (8), provides an achievable lower bound of the non-convex problem of interest, e.g., (6). From this strong coupling between the convex and non-convex problems we then derive sufficient conditions to verify if a local-minimizer is a global-minimizer and show that if the size of the variables in the non-convex problem is initialized to be sufficiently large then from *any* initialization it is possible to reach a global-minimizer using purely local descent.

4. Neural Networks with One Hidden Layer

The above discussion on matrix factorization can be extended to neural networks with one hidden layer by properly adjusting the definitions of the maps Φ_r and Θ_r. In matrix factorization, Φ_r can be re-written as $\Phi_r(W^1, W^2) = W^1W^{2\top} = \sum_{i=1}^r W_i^1 W_i^{2\top}$, where W_i^1 and W_i^2 are the ith columns of W^1 and W^2, respectively. Likewise, Θ_r can be re-written as $\Theta_r(W^1, W^2) = \frac{1}{2}(\|W^1\|_F^2 + \|W^2\|_F^2) = \sum_{i=1}^r \frac{1}{2}(\|W_i^1\|_2^2 + \|W_i^2\|_2^2)$. This motivates the following more general definitions for Φ_r and Θ_r:

$$\Phi_r(W^1, W^2) = \sum_{i=1}^r \phi(W_i^1, W_i^2) \quad \text{and}$$
$$\Theta_r(W^1, W^2) = \sum_{i=1}^r \theta(W_i^1, W_i^2), \quad (10)$$

where ϕ and θ are positively homogeneous of degree 2, i.e., $\phi(\alpha w^1, \alpha w^2) = \alpha^2\phi(w^1, w^2)$ for all $\alpha \geq 0$. Clearly, $\phi(w^1, w^2) = w^1 w^{2\top}$ and $\theta(w^1, w^2) = \|w^1\|_2^2 + \|w^2\|_2^2$ satisfy this property. But notice that it is also satisfied, for example, by the map $\phi(w^1, w^2) = \psi^+(S^\top w^1)w^{2\top}$, where recall $\psi^+(x, 0) = \max(x, 0)$ is a ReLU applied to each entry of $S^\top w^1$. The fundamental observation is that both linear transformations and ReLU nonlinearities[3] are positively homogeneous functions of degree one, and so the output of a two-layer network is positively homogeneous of degree two.

With these definitions, it is easy to see that the output of a two-layer neural network with nonlinearity ψ^+ on the hidden units, such as the one illustrated in the left panel of Figure 3, can be expressed by the map Φ_r in (10), where r now represents the number of neurons in the hidden layer. Therefore, we can write the training problem for a two-layer network as:

$$\min_{r\in\mathbb{N}^+} \min_{W^1,W^2} \ell(Y, \Phi_r(W^1, W^2)) + \lambda\Theta_r(W^1, W^2). \quad (11)$$

This problem is non-convex due to the mapping Φ_r, so to analyze this non-convex problem, we define a generalization of the nuclear norm in (7) for two-layer neural networks as:

$$\Omega_{\phi,\theta}(X) = \min_{r\in\mathbb{N}^+} \min_{W^1,W^2:\Phi_r(W^1,W^2)=X} \Theta_r(W^1, W^2). \quad (12)$$

The intuition behind the above problem is that, given an output X generated by the network for some input S, we wish to find the network size r and weights (W^1, W^2) that produced the output X. Among all possible sizes and weights, we prefer those that minimize $\Theta_r(W^1, W^2)$.

While $\Omega_{\phi,\theta}$ is no longer necessarily a norm, we will show in Proposition 1 that, under some additional conditions on θ, $\Omega_{\phi,\theta}$ is still convex. Therefore, if the loss ℓ is convex on X, so is the problem

$$\min_X \ell(Y, X) + \Omega_{\phi,\theta}(X). \quad (13)$$

As shown in Theorem 1, the convex problem (13) gives an achievable lower bound to the non-convex problem (11), and a local minimizer of the non-convex problem such that one of the columns of W^1 and W^2 is equal to zero gives a global minimizer for both the convex and non-convex problems.

[3]Notice that many other neural network operators such as max-pooling and convolution are also positively homogeneous.

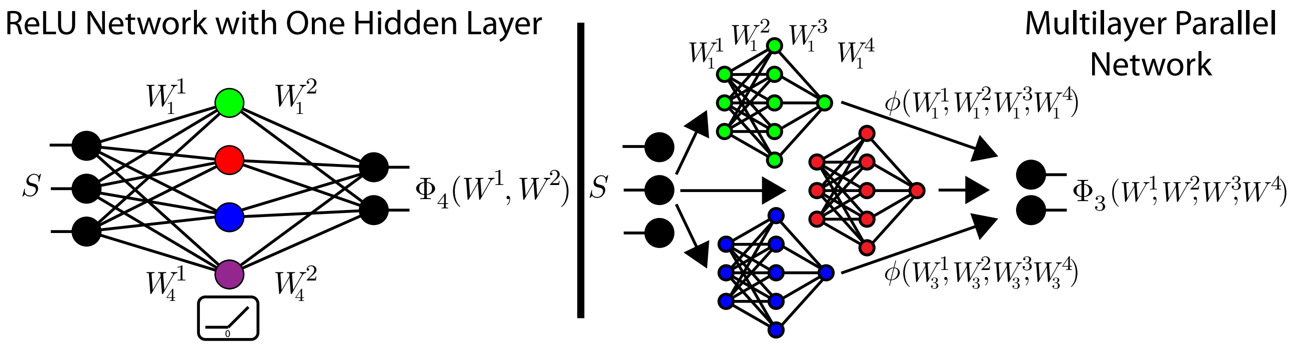

Figure 3. Example networks. (Left panel) ReLU network with a single hidden layer with the mapping Φ_r described by the equation in (10) with ($r = 4$). Each color corresponds to one element of the elemental mapping $\phi(W_i^1, W_i^2)$. The colored hidden units have rectifying non-linearities, while the black units are linear. (Right panel) Multilayer ReLU network with 3 fully connected parallel subnetworks ($r = 3$), where each color corresponds to the subnetwork described the elemental mapping $\phi(W_i^1, W_i^2, W_i^3, W_i^4)$.

5. Deep Networks with Parallel Structure

In this section, we extend our results to networks formed by the addition of r parallel sub-networks, with each sub-network having the same architecture (see Figure 3, right). After introducing some specialized notation for the network weights and dimensions, we generalize the maps Φ and Θ to the context of deep networks. We then introduce a regularizer on the network weights and show that it induces a convex regularizer on the output. Finally, we state our main results to give sufficient conditions to guarantee that local minima are global minima and that for sufficiently large networks all local minima are guaranteed to be global minima.

Notation. We will use capital letters as a shorthand for a set of dimensions, and individual dimensions will be denoted with lower case letters. For example, $X \in \mathbb{R}^{d_1 \times \ldots \times d_N} \equiv X \in \mathbb{R}^D$ for $D = d_1 \times \ldots \times d_N$; we also denote the cardinality of $X \in \mathbb{R}^D$ as $\text{card}(X) = \prod_{i=1}^N d_i$. Similarly, $X \in \mathbb{R}^{D \times R} \equiv X \in \mathbb{R}^{d_1 \times \ldots \times d_N \times r_1 \times \ldots \times r_M}$ for $D = d_1 \times \ldots \times d_N$ and $R = r_1 \times \ldots \times r_M$. Given an element from a tensor space, we will use a subscript to denote a slice of the tensor along the last dimension. For example, given a matrix $W \in \mathbb{R}^{d_1 \times r}$, then $W_i \in \mathbb{R}_1^d, i \in \{1, \ldots, r\}$, denotes the i'th column of W. Similarly, given a cube $W \in \mathbb{R}^{d_1 \times d_2 \times r}$ then $W_i \in \mathbb{R}^{d_1 \times d_2}, i \in \{1 \ldots, r\}$, denotes the i'th slice along the third dimension. Further, given two tensors with matching dimensions except for the last dimension, $W \in \mathbb{R}^{D \times r_w}$ and $Q \in \mathbb{R}^{D \times r_q}$, we will use $[W\ Q] \in \mathbb{R}^{D \times (r_w + r_q)}$ to denote the concatenation of the two tensors along the last dimension. We'll also need the following definitions:

1. A *size-r set of K factors* $(W^1, \ldots, W^K)_r \in \mathbb{R}^{(D^1 \times r)} \times \ldots \times \mathbb{R}^{(D^K \times r)}$ is defined to be a set of K tensors where the final dimension of each tensor is equal to r.

2. A function $\theta : \mathbb{R}^{D^1} \times \ldots \times \mathbb{R}^{D^K} \to \mathbb{R}^D$ is *positively homogeneous with degree p* if, $\forall \alpha \geq 0$, $\theta(\alpha w^1, \ldots, \alpha w^K) = \alpha^p \theta(w^1, \ldots, w^K)$. Note that this implies that $\theta(0, \ldots, 0) = 0$ for $p \neq 0$.

3. A function $\theta : \mathbb{R}^{D^1} \times \ldots \times \mathbb{R}^{D^K} \to \mathbb{R}_+$ is *positive semidefinite* if $\theta(0, \ldots, 0) = 0$ and $\theta(w^1, \ldots, w^K) \geq 0$, $\forall (w^1, \ldots, w^K)$.

Factorization and regularization maps. The maps Φ_r and Θ_r are defined as *sums of positively homogeneous elemental mappings* ϕ and θ, i.e.:

$$\Phi_r(W^1, \ldots, W^K) = \sum_{i=1}^r \phi(W_i^1, \ldots, W_i^K) \quad \text{and}$$
$$\Theta_r(W^1, \ldots, W^K) = \sum_{i=1}^r \theta(W_i^1, \ldots, W_i^K). \tag{14}$$

Note that for matrix factorization, the elemental mapping $\phi : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{d_1 \times d_2}$ is defined as $\phi(w^1, w^2) = w^1 w^{2\top}$, which is positively homogeneous of degree 2, and the factorization map $\Phi_r(W^1, W^2) = \sum_{i=1}^r W_i^1 W_i^{2\top} = W_1 W_2^\top$ is simply matrix multiplication for matrices with r columns. For neural networks, a typical elemental mapping ϕ would be defined as $\phi(w^1, \ldots, w^K) = \psi_K(\cdots \psi_2(\psi_1(S^\top w^1) w^2) \cdots w^K)$, which denotes the application of a linear transformation (dot-product, convolution) with parameters w^1 to the data S followed by a positively homogeneous nonlinearity ψ_i (ReLU, max-pooling), and so on for K weight layers. Therefore, the map Φ_r in (14) corresponds to the addition of r deep sub-networks in parallel, each one with the same number of layers and the same number of neurons per layer (see Figure 3, right). The well-known AlexNet network from [14], which consists of a series of convolutional layers, linear-rectification, max-pooling layers, response normalization layers, and fully connected layers, can be described by taking $r = 1$ and defining ϕ to be the entire transformation of the network (with slight modification of the response normalization layers, which are not positively homogenous, see supplement).

Note, however, that our results will rely on r potentially changing or being initialized to be sufficiently large, which limits the applicability of our results to current state-of-the-art network architectures (see discussion).

The *elemental regularization function* $\theta : \mathbb{R}^{D^1} \times \ldots \times \mathbb{R}^{D^K} \to \mathbb{R}_+ \cup \infty$ takes as input the parameters of one sub-network and returns a non-negative number. The requirements we place on θ are that it must be positively homogeneous and positive semidefinite.

A regularizer on the parameters of the network that induces a convex regularizer on its output. To define our regularization function on the output of the network, $X = \Phi_r(W^1, \ldots, W^K)$, it will be necessary that the elemental regularization function, θ, and the elemental mapping, ϕ, satisfy a few properties to be considered *compatible* for the definition of our regularization function. Specifically, we say that (ϕ, θ) are a *nondegenerate pair* if: 1) θ and ϕ are both positively homogeneous with degree p, for some $p > 0$ and 2) $\theta(z^1, \ldots, z^K) > 0,\ \forall (z^1, \ldots, z^K)$ such that $\phi(z^1, \ldots, z^K) \neq 0$ and for all sequences $(z^1[n], \ldots, z^K[n])$, $n = 1, \ldots, \infty$, if $\|\phi(z^1[n], \ldots, z^K[n])\| \to \infty$ then $\theta(z^1[n], \ldots, z^K[n]) \to \infty$.[4]

Notice that any norm $\|w\|$ is positively homogeneous with degree 1, so by taking products of norms or sums of norms raised to an appropriate power we can match the degree of positive homogeneity between the mapping and regularizer. A typical regularizer for a mapping of degree K could be $\theta(w^1, \ldots, w^K) = \prod_{i=1}^K \|w^i\|$ or $\theta(w^1, \ldots, w^K) = \sum_{i=1}^K \|w^1\|^K$, where the choice of norms is arbitrary. However, as we place very few requirements on θ, our framework can include an extremely wide variety of potential regularization functions. For example, indicator functions on conic sets, such as requiring the factors to be non-negative, are also positively homogeneous and can be incorporated by our framework.

Given a nondegenerate pair (ϕ, θ) of an elemental mapping ϕ and an elemental regularization function θ, we define the *factorization regularization function*, $\Omega_{\phi,\theta}(X) : \mathbb{R}^D \to \mathbb{R}_+ \cup \infty$ to be

$$\Omega_{\phi,\theta}(X) \equiv \inf_{r \in \mathbb{N}^+} \inf_{(W^1,\ldots,W^K)_r} \sum_{i=1}^r \theta(W_i^1, \ldots, W_i^K) \quad \text{s.t. } \Phi_r(W^1, \ldots, W^K) = X \tag{15}$$

with the additional condition that $\Omega_{\phi,\theta}(X) = \infty$ if $X \notin \bigcup_r \operatorname{Im}(\Phi_r)$.

The following proposition shows that $\Omega_{\phi,\theta}$ is a convex function of X and that in general the infimum in (15) can be achieved with a finitely sized network (i.e., r does not need to approach ∞)[5].

Proposition 1 *The function $\Omega_{\phi,\theta} : \mathbb{R}^D \to \mathbb{R} \cup \infty$ as defined in (15) has the following properties*

1. *$\Omega_{\phi,\theta}$ is positve definite, i.e., $\Omega_{\phi,\theta}(0) = 0$ and $\Omega_{\phi,\theta}(X) > 0\ \forall X \neq 0$.*
2. *$\Omega_{\phi,\theta}$ is positively homogeneous with degree 1.*
3. *$\Omega_{\phi,\theta}(X + Z) \leq \Omega_{\phi,\theta}(X) + \Omega_{\phi,\theta}(Z)\ \forall (X, Z)$*
4. *$\Omega_{\phi,\theta}(X)$ is convex w.r.t. $X \in \mathbb{R}^D$.*
5. *The infimum in (15) can be achieved with $r \leq \operatorname{card}(X)\ \forall X$ s.t. $\Omega_{\phi,\theta}(X) < \infty$.*

Global optimality from local minima. While $\Omega_{\phi,\theta}(X)$ is convex, unlike the nuclear norm $\|X\|_*$, it typically cannot be evaluated in polynomial time due to its complicated definition. Nonetheless, its convexity allows us to use $\Omega_{\phi,\theta}$ as an analysis tool to derive results for neural network training formulations. In particular, it allows us to consider the convex (but typically non-tractable) problem, given by

$$\min_X F(X) \equiv \ell(Y, X) + \lambda \Omega_{\phi,\theta}(X). \tag{16}$$

Here $X \in \mathbb{R}^D$ is the output of the factorization mapping $X = \Phi_r(W^1, \ldots, W^K)$, $\ell(Y, X)$ is a loss function that is assumed to be once differentiable and convex in X, $\Omega_{\phi,\theta}(X)$ is as defined by (15) where (ϕ, θ) is assumed to be a nondegenerate pair, and $\lambda > 0$. Given these assumptions, we are now ready to state our main results.

Theorem 1 *Any local minimizer of the non-convex optimization problem*

$$\min_{(W^1,\ldots,W^K)_r} f_r(W^1, \ldots, W^K) \equiv \ell(Y, \Phi_r(W^1, \ldots, W^K)) + \lambda \sum_{i=1}^r \theta(W_i^1, \ldots, W_i^K) \tag{17}$$

such that $(W_{i_0}^1, \ldots, W_{i_0}^K) = (0, \ldots, 0)$ for some $i_0 \in \{1, \ldots, r\}$ is a global minimizer of (17). Moreover, $X = \Phi_r(W^1, \ldots, W^K)$ is a global minimizer of (16).

Proof Sketch. The proofs of all our results are available in the supplement, but here we outline the sketch of the argument. First, note that from the definition of

[4] Property 1 from the definition of a nondegenerate pair will be critical to our formulation. Property 2 is typically satisfied for most 'interesting' choices of (ϕ, θ) and is designed to avoid 'pathological' $\Omega_{\phi,\theta}$ functions (such as $\Omega_{\phi,\theta}(X) = 0\ \forall X$).

[5] In particular, the largest r needs to be is $\operatorname{card}(X)$, and we note that $\operatorname{card}(X)$ is a worst case upper bound on the size of the factorization. In certain cases the bound can be shown to be lower. As an example, $\Omega_{\phi,\theta}(X) = \|X\|_*$ when $\phi(u, v) = uv^T$ and $\theta(u, v) = \|u\|_2 \|v\|_2$. In this case the infimum can be achieved with $r \leq \operatorname{rank}(X) \leq \min\{\operatorname{card}(u), \operatorname{card}(v)\}$.

$\Omega_{\phi,\theta}$ the convex optimization problem (16) globally lower bounds the non-convex factorization problem (17) for any $X = \Phi_r(W^1, \ldots, W^K)$, and because (16) is a convex function of X, the conditions for global optimality are easily derived. The result is completed by showing that a local minimium of (17) which satisfies the statement of the Theorem also satisfies the conditions to be global minimum of (16) at $X = \Phi_r(W^1, \ldots, W^K)$, and due to the fact that (16) globally lower bounds (17) this implies that the local minimum of (17) is a global minimum. ■

From this result, we can then test the global optimality of any local minimum from the immediate corollary:

Corollary 1 *Given a function $f_r(W^1, \ldots, W^K)$ of the form given in* (17)*, any local minimizer of the optimization problem*

$$\min_{(W^1,\ldots,W^K)_r} f_r(W^1, \ldots, W^K) \tag{18}$$

is a global minimizer if $f_{r+1}([W^1\ 0], \ldots, [W^K\ 0])$ is a local minimizer of f_{r+1}.

Global minima can be found by local descent. From the results of Theorem 1, we are now also able to show that if we let the number of subnetworks (r) become large enough, then from any initialization we can always find a global minimizer of $f_r(W^1, \ldots, W^K)$ using a purely local descent strategy. Specifically, we have the following result, whose proof gives a meta-algorithm for solving the optimization problem.

Theorem 2 *Given a function $f_r(W^1, \ldots, W^K)$ as defined by* (17)*, if $r >$ card(X) then from any point $(Z^1, \ldots, Z^K)$ such that $f_r(Z^1, \ldots, Z^K) < \infty$ there must exist a non-increasing path from $(Z^1, \ldots, Z^K)$ to a global minimizer of $f_r(W^1, \ldots, W^K)$.*

Proof Sketch. The proof is done in a constructive manner and defines a meta-algorithm that can be combined with any local-descent algorithm to reach a global minimum.

1. Perform local descent until arriving at a local minimum.
2. If one of the parallel networks is all 0 - i.e., $\exists i_0 \in \{1, \ldots, r\}$ such that $(W^1_{i_0}, \ldots, W^K_{i_0}) = (0, \ldots, 0)$ - then we are at a global minimum due to Theorem 1.
3. Else if there exists a nonzero $\beta \in \mathbb{R}^r$ such that $\sum_{i=1}^r \beta_i \phi(W^1_i, \ldots, W^K_i) = 0$ then scale β so that $\min_i \beta_i = -1$ and set $W^k_i \leftarrow (1 + \beta_i)^{1/p} W^k_i$ for $k = 1, \ldots, K$. Such a β is guaranteed to exist if $r >$ card(X), and the operation of scaling the variables by the $(1+\beta_i)^{1/p}$ terms is shown to traverse a flat surface of the objective function until arriving at a point where one of the parallel networks is all 0. From there, if a local descent direction exists continue local-descent. If no local descent direction exists then we are again at a global minimum due to Theorem 1.
4. Otherwise, if $r \leq$ card(X) and no β exists in the prior step then increment r by appending a subnetwork in parallel initialized as all-zeros. If a local-descent direction exists continue local-descent. Otherwise, we are at a global minimum due to Corollary 1.

■

6. Discussion: Limitations and Implications

While the framework described so far is very general, and provides guarantees of global optimality for various forms of neural network problems, we pause to note a few practical limitations and then discuss implications of our results in the design of neural networks.

First, note that the size of the network is controlled by a single parameter: r. While a single parameter is sufficient for matrix factorization, where r is the size of the factors, and two-layer neural networks, where r is the number of neurons in the hidden layer, this is insufficient to model deep networks where we want to control also the number of layers K as well as the number of neurons in each layer. In other words, while one could naively assume that the results so far apply to current deep networks by setting $r = 1$ (i.e., the network is not assumed to have a parallel structure), this is *not* the case because the assumption that the network needs to be "large enough" (i.e., r must be sufficiently large) is essential to prove that local minima are global. Therefore, to analyze current deep networks without this parallel structure, it is essential that we extend the framework to optimize over additional network size parameters, such as the number of layers and the numbers of neurons per layer.

Additionally, the maximum upper-bound size for r in Theorem 2 is typically much too large to be of practical use. Note, however, that the bounds we have shown here are for the most general case of mapping and regularizer, (ϕ, θ), and that a very interesting line of future work is to explore sufficient conditions on these functions that allows the size of r to be greatly reduced. For example, in the nuclear norm case of matrix factorization, it is well known that the largest r will need to be is equal to the rank of the final solution. As an example from neural networks, if the architecture of a parallel sub-network (as defined by ϕ) is sufficiently rich to span the output space (i.e., Im$(\phi) = \mathbb{R}^D$), then if θ satisfies the triangle inequality it is easily seen from the definition of $\Omega_{\phi,\theta}$ that the infimum in (15) can always be achieved with a single parallel network (i.e., $r = 1$).

A final limitation to note is that the meta-algorithm used to construct the proof of Theorem 2 relies on using local descent, which is to be contrasted with typical optimization algorithms such as gradient descent. While gradient descent is certainly a form of local descent, we remind the reader that in general finding a local descent direction of a non-convex function can be a NP-hard problem in general (for example, at a saddle-point), so our results do not necessarily imply the existence of polynomial time algorithms that can solve all of the potential formulations captured within our framework. Again, however, we emphasize that our analysis is a worst-case analysis (i.e., choose *any* possible initialization) for any potential mapping and regularizer, and significant potential exists to strengthen our results by considering specific families of mapping and regularizers, initialization strategies, or statistical distributions of the training data that allow for polynomial time guarantees.

Implications for Neural Networks. Despite the limitations discussed above, our analysis suggests several significant guiding principles regarding the training of neural networks which can facilitate more efficient optimization. The first is that balancing the degree of positive homogeneity between the regularization function and the mapping function is crucial. In fact, it can be shown (see supplement, Section 8) that if the degrees of positive homogeneity do not match between the mapping and the regularization function, then it either becomes impossible to make guarantees regarding the global optimality of a local minimum, or it becomes possible that the regularization function will do nothing to limit the size of the network, so the degrees of freedom in the model are largely determined by the user defined choice of r. In practice, this issue often arises in the context of weight decay, where the regularization function is typically chosen as $\Theta_r(W^1, \dots, W^K) = \sum_i \|W^i\|_F^2$ or $\Theta_r(W^1, \dots, W^K) = \sum_i \|W^i\|_1$. Since these functions are only positively homogeneous of degree 2 and 1, respectively, the mapping of a deep network will typically not be balanced with the regularization and one is guaranteed that non-optimal local minima will exist regardless of the size of the network (supplement, Proposition 2). We note that this is a potential explanation of the noted inferior performance of using weight decay versus dropout regularization [23, 14, 24] and that several more recently proposed successful regularization strategies are compatible with balanced degrees of positive homogeneity. For example, the Path-SGD regularizer proposed in [18] takes a product of weights along a path through the network and then calculates a norm of all possible paths through the network. Note that the product of weights along a path through the network will typically have a degree of positive homogeneity equal to K and thus will be balanced with the degree of positive homogeneity of the network output for typical network architectures. Likewise, Batch Normalization proposed in [12] essentially adds a whitening operation to the input of a layer which is a positively homogeneous transformation similar to contrast normalization but across training examples (see Section 9 of the supplement). Taken together, our results suggest at several key properties one should account for in the design of network regularization and provide many interesting opportunities for experimental exploration in future work.

A final implication of our analysis is that neural networks which generate the output by taking the sum of multiple parallel subnetworks are highly conducive to efficient optimization. This idea, of linearly combining the outputs of multiple subnetworks, has clear analogies to ensemble methods like boosting and bagging and was a large motivation in the development of techniques such as dropout, which stochastically approximates the average output of an exponential number of subnetworks [23]. The framework we present here is not an exact analogy to dropout, as dropout enforces equality in network weights across parallel networks with a shared parameterization (i.e., if a neuron is present in a given subnetwork all of its input and output weights must be equal to the same neuron in the other subnetworks), but many interesting questions for future work can be asked regarding the concept of summing multiple subnetworks combined with considering more general forms of network mappings which allow for common parametrization of the subnetworks.

7. Conclusions

Here we have presented a general framework which allows for a wide variety of non-convex optimization problems, including certain forms of neural network training, to be analyzed with tools from convex analysis and induces a convex regularizer on the output of the non-convex mapping. In particular, we have shown sufficient conditions to guarantee that any local minimum is a global minimum of the non-convex factorization problem and that if the non-convex factorization problem is done with factors of sufficient size, then from any feasible initialization it is always possible to find a global minimizer using a purely local descent algorithm. Additionally, our results suggest that balancing the degrees of positive homogeneity between the network mapping and the regularization function is critical for preventing non-optimal local minima in the loss surface of modern neural network architectures and offer guidance for the design of network architectures and regularizers.

Acknowledgments. This work was supported by NSF grants 1447822, 1618485 and 1618637. We thank Laurent Younes for insightful discussions.

References

[1] F. Bach. Convex relaxations of structured matrix factorizations. *arXiv:1309.3117v1*, 2013. 4

[2] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. *arXiv:0812.1869v1*, 2008. 4

[3] Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. In *Neural Information Processing Systems*, pages 123–130, 2005. 2, 3

[4] S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, Series A(103):427—444, 2005. 4

[5] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. 2

[6] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *International Conference on Artificial Intelligence and Statistics*, pages 192–204, 2015. 2, 3

[7] G. E. Dahl, T. N. Sainath, and G. E. Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8609–8613, 2013. 2

[8] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Neural Information Processing Systems*, pages 2933–2941, 2014. 3

[9] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006. 2

[10] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001. 3

[11] B. Haeffele, E. Young, and R. Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International Conference on Machine Learning*, pages 2007–2015, 2014. 4

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 2, 8

[13] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015. 3

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, pages 1097–1105, 2012. 2, 5, 8

[15] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, volume 30, 2013. 2

[16] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010. 2

[17] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean. Boosting algorithms as gradient descent. In *Neural Information Processing Systems*, pages 512–518, 2000. 3

[18] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Neural Information Processing Systems*, pages 2422–2430, 2015. 2, 8

[19] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. 4

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive Modeling*, 5, 1988. 2

[21] I. Safran and O. Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782, 2016. 3

[22] N. Srebro, J. D. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Neural Information Processing Systems*, volume 17, pages 1329–1336, 2004. 4

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 2, 8

[24] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066, 2013. 8

[25] S. J. Wright and J. Nocedal. *Numerical Optimization*, volume 2. Springer New York, 1999. 2

[26] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013. 2

[27] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton. On rectified linear units for speech processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3517–3521, 2013. 2

8. The Importance of the Degrees of Homogeneity

Here we briefly explore the importance of the relative degrees of positive homogeneity between the mapping function and the regularization function as we briefly mentioned in the main text. Currently our results are limited to the study of networks with parallel architectures due to our choice of the Φ_r mapping, but we conjecture our results can likely be generalized to include additional positively homogeneous factorization mappings and regularizers. However, even for more general mappings and regularization functions, it will be necessary to carefully consider the degrees of positive homogeneity between the regularization function and the mapping function to show results similar to those we present here. In general, if the degrees of positive homogeneity do not match between the factorization mapping and the regularization function, then it either becomes impossible to make guarantees regarding the global optimality of a local minimum, or it becomes possible that the regularization function does nothing to limit the size of the factorization, so the degrees of freedom in the model become largely determined by the user defined choice of r.

As a demonstration of these phenomena, first consider the case where we have a general mapping, $\Phi(W^1, \ldots, W^K)$, which is positively homogeneous with degree p (but which is not assumed to have the parallel network form). Now, consider a general regularization function, $\Theta(W^1, \ldots, W^K)$, which is positively homogeneous with degree $p' < p$, then the following proposition provides a simple counter-example demonstrating that in general it is not possible to guarantee that a global minimum can be found from local descent from an arbitrary initialization.

Proposition 2 *Let $\ell : \mathbb{R}^D \to \mathbb{R}$ be a convex function with $\partial\ell(0) \neq \emptyset$; let $\Phi : \mathbb{R}^{D^1} \times \ldots \times \mathbb{R}^{D^K} \to \mathbb{R}^D$ be a positively homogeneous mapping with degree p; and let $\Theta : \mathbb{R}^{D^1} \times \ldots \times \mathbb{R}^{D^K} \to \mathbb{R}_+$ be a positively homogeneous function with degree $p' < p$ such that $\Theta(0, \ldots, 0) = 0$ and $\Theta(W^1, \ldots, W^K) > 0 \;\; \forall\{(W^1, \ldots, W^K) : \Phi(W^1, \ldots, W^K) \neq 0\}$. Then, the optimization problem given by*

$$\min_{(W^1,\ldots,W^K)} \tilde{f}(W^1, \ldots, W^K) = \ell(\Phi(W^1, \ldots, W^K)) + \Theta(W^1, \ldots, W^K) \tag{19}$$

has a local minimum at $(W^1, \ldots, W^K) = (0, \ldots, 0)$. Additionally, $\forall(W^1, \ldots, W^K)$ such that $\Phi(W^1, \ldots, W^K) \neq 0$ there exists a δ such that $\forall\epsilon \in (0, \delta)$ $\tilde{f}(\epsilon W^1, \ldots, \epsilon W^K) > \tilde{f}(0, \ldots, 0)$.

Proof. Consider $\tilde{f}(\epsilon W^1, \ldots, \epsilon W^K) - \tilde{f}(0, \ldots, 0)$. This gives

$$\ell(\Phi(\epsilon W^1, \ldots, \epsilon W^K)) + \Theta(\epsilon W^1, \ldots, \epsilon W^K) - \ell(0) - \Theta(0, \ldots, 0) = \tag{20}$$

$$\ell(\epsilon^p \Phi(W^1, \ldots, W^K)) - \ell(0) + \epsilon^{p'}\Theta(W^1, \ldots, W^K) \geq \tag{21}$$

$$\epsilon^p \left\langle \partial\ell(0), \Phi(W^1, \ldots, W^K)\right\rangle + \epsilon^{p'}\Theta(W^1, \ldots, W^K), \tag{22}$$

where the inequality is simply due to the definition of the subgradient of a convex function. Recall that $p > p'$ and $\Phi(W^1, \ldots, W^K) \neq 0 \iff \Theta(W^1, \ldots, W^K) > 0$, so $\forall(W^1, \ldots, W^K)$, $\tilde{f}(\epsilon W^1, \ldots, \epsilon W^K) - \tilde{f}(0, \ldots, 0) \geq 0$ for $\epsilon > 0$ and sufficiently small, with equality iff $\Theta(W^1, \ldots, W^K) = 0 \iff \Phi(W^1, \ldots, W^K) = 0$, giving the result. ■

The above proposition shows that unless we have the special case where $(W^1, \ldots, W^K) = (0, \ldots, 0)$ happens to be a global minimizer, then there will always exist a local minimum at the origin, and from the origin it will always be necessary to take an increasing path along the objective function surface to escape the local minimum. The case described above, where $p > p'$, is arguably the more common situation for mismatched degrees of homogeneity (as opposed to $p < p'$), and a typical example might be an objective function such as

$$\ell(\Phi(W^1, \ldots, W^K)) + \lambda \sum_{i=1}^{K} \|W^i\|_{(i)}^{p'}, \tag{23}$$

where Φ is a positively homogeneous mapping with degree $K > 2$ (e.g., the mapping of a deep neural network) but p' is typically taken to be only 1 or 2 depending on the particular choice of norms (e.g., $\|W^i\|_F^2$ or $\|W^i\|_1$).

Conversely, in the situation where $p' > p$, then it is often the case that the regularization function is not sufficient to "limit" the size of the network, in the sense that the objective function can always be decreased by allowing additional subnetworks to be added in parallel. As a simple example, consider the case of matrix factorization with the objective function

$$\min_{U,V} \ell(UV^T) + \lambda(\|U\|^{p'} + \|V\|^{p'}). \tag{24}$$

If the size of the factorization doubles, then we can always take $[\frac{\sqrt{2}}{2}U\ \frac{\sqrt{2}}{2}U][\frac{\sqrt{2}}{2}V\ \frac{\sqrt{2}}{2}V]^T = UV^T$, so if $(\frac{\sqrt{2}}{2})^{p'}(\|[U\ U]\|^{p'} + \|[V\ V]\|^{p'}) < \|U\|^{p'} + \|V\|^{p'}$, then the objective function can always be decreased by simply duplicating and scaling the existing factorization. It is easily verified that the above inequality is satisfied for many choices of norms (for example, all the l_q norms with $q \geq 1$) when $p' > 2$. As a result, this implies that the degrees of freedom in the model will be largely dependent on the particular choice of the number of columns in (U, V), since in general the objective function is typically decreased by having all entries of (U, V) be non-zero. Likewise, in neural network training, the degrees of freedom are largely determined by the user defined choice of network size and not fit to the data via regularization. We note, however, that in some cases having $p' > p$ does induce a meaningful convex regularization function of the form $\Omega^q_{\phi,\theta}$ for some $q > 1$, but we save a full characterization of such cases for future work.

9. Extentions to Other Types of Network Layers

For the types of network layers discussed in the main text (i.e., a linear operation followed by a non-linear function which is positively homogeneous with degree 1), it is clear that adding an extra layer to the network typically increases the overall degree of the mapping by 1, but there are a few points to consider that can complicate the overall positive homogeneity of a network mapping when other types of network layers are used which we briefly discuss here. The first is contrast normalization. This is typically used in convolutional networks and takes the form of applying a transformation such as $g_i = z_i / f(N(z_i))$, where g_i denotes the i^{th} output of the normalization layer, z_i denotes the i^{th} input to the normalization layer, and $f(N(z_i))$ denotes a function of the inputs to the normalization layer in a neighborhood surrounding z_i. If $f(N(z_i))$ is positively homogeneous with degree p', such as a norm raised to p', then the normalization layer is also a positively homogeneous transformation[6], but it "resets" the degree of positive homogeneity to be $1 - p'$ at that stage in the network. As a result, care must be taken to ensure that sufficiently many layers exist following the normalization layer so that the overall degree of the network mapping becomes larger that 0. The second issue to consider with regards to staying strictly within the positively homogeneous framework is the use of bias terms. For example, the output of a fully connected ReLU layer with bias terms is given by $G = \psi^+(ZW + B)$, where again G denotes the output of the layer, Z denotes the input to the layer, W denotes the connection weights, and B denotes the bias terms. If the input, Z, comes from lower layers of the network then it can already be a positively homogeneous function of the weight parameters in the lower layers, so B must be raised to an appropriate power to preserve the overall homogeneity of the mapping with respect to all the variables we are optimizing over (including B). For example, if Z is positively homogeneous of degree 3, then we could instead use bias terms of the form $G = \psi^+(ZW + B_p^{(4)} - B_n^{(4)})$, where $B^{(4)}$ denotes raising each element to the 4'th power entry-wise, and the use of both the B_p and B_n terms allows for negative bias terms. This then results in a mapping which is positively homogeneous with respect to all of the connection weights and bias terms in the network. Note that in this case, the θ regularization should also include the bias parameters as input.

10. Proofs

Here we will formally present proofs for all of our Propositions, Corollaries, and Theorems. In addition, we will also introduce a few additional intermediate Propositions and Lemmas which will be necessary for our argument. We begin by first deriving the Fenchel dual of the $\Omega_{\phi,\theta}$ function. Recall, that the Fenchel dual of a function $g(x)$ is defined as $g^*(z) \equiv \sup_x \langle z, x \rangle - g(x)$.

Proposition 3 *The Fenchel dual of $\Omega_{\phi,\theta}(X)$ is given by*

$$\Omega^*_{\phi,\theta}(Z) = \begin{cases} 0 & \Omega^\circ_{\phi,\theta}(Z) \leq 1 \\ \infty & \text{otherwise} \end{cases} \tag{25}$$

where

$$\Omega^\circ_{\phi,\theta}(Z) \equiv \sup_{(w^1,\ldots,w^K)} \left\langle Z, \phi(w^1, \ldots, w^K) \right\rangle \ \text{ s.t. } \ \theta(w^1, \ldots, w^K) \leq 1. \tag{26}$$

[6] Usually, most response normalization layers are not strictly positively homogeneous as they add a small non-zero constant to the denominator to avoid division by 0, but if the constant is significantly smaller than the value of $f(N(z_i))$ it is a very close approximation of a positively homogeneous transformation.

Proof. $\Omega^*_{\phi,\theta}(Z) \equiv \sup_X \langle Z, X\rangle - \Omega_{\phi,\theta}(X)$, so for X to approach the supremum we must have $X \in \bigcup_r \text{Im}(\Phi_r)$. As result, the problem is equivalent to

$$\Omega^*_{\phi,\theta}(Z) = \sup_{r\in\mathbb{N}^+} \sup_{(W^1,\ldots,W^K)_r} \left\langle Z, \Phi_r(W^1,\ldots,W^K)\right\rangle - \sum_{i=1}^r \theta(W_i^1,\ldots,W_i^K) \tag{27}$$

$$= \sup_{r\in\mathbb{N}^+} \sup_{(W^1,\ldots,W^K)_r} \sum_{i=1}^r \left[\left\langle Z, \phi(W_i^1,\ldots,W_i^K)\right\rangle - \theta(W_i^1,\ldots,W_i^K)\right]. \tag{28}$$

If $\Omega^\circ_{\phi,\theta}(Z) \le 1$ then all the terms in the summation of (28) will be non-positive, so taking $(W^1,\ldots,W^K) = (0,\ldots,0)$ will achieve the supremum. This can be seen by noting that because of the balanced degrees of homogeneity between ϕ and θ, if $\Omega^\circ_{\phi,\theta}(Z) \le 1$ then we will always have $\left\langle Z, \phi(w^1,\ldots,w^K)\right\rangle \le \theta(w^1,\ldots,w^K)$ since we can always rescale the $(w^1,\ldots,w^K)$ terms by a positive constant α so that $\theta(\alpha w^1,\ldots,\alpha w^K) = 1$. To make this point explicit, consider any $(w^1,\ldots,w^K)$ and $\alpha > 0$ such that $\theta(\alpha w^1,\ldots,\alpha w^K) = 1$, giving

$$\alpha^p \left\langle Z, \phi(w^1,\ldots,w^K)\right\rangle = \left\langle Z, \phi(\alpha w^1,\ldots,\alpha w^K)\right\rangle \le 1 = \theta(\alpha w^1,\ldots,\alpha w^K) = \alpha^p \theta(w^1,\ldots,w^K). \tag{29}$$

The inequality above comes from the fact that $\Omega^\circ_{\phi,\theta}(Z) \le 1$, and since $\alpha^p > 0$ we can cancel it from both sides of the inequality to give $\left\langle Z, \phi(w^1,\ldots,w^K)\right\rangle \le \theta(w^1,\ldots,w^K)$.

Conversely, if $\Omega^\circ_{\phi,\theta}(Z) > 1$, then $\exists(w^1,\ldots,w^K)$ such that $\left\langle Z, \phi(w^1,\ldots,w^K)\right\rangle > \theta(w^1,\ldots,w^K)$. This result, combined with the positive homogeneity of ϕ and θ gives that (28) is unbounded by considering $(\alpha w^1,\ldots,\alpha w^K)$ as $\alpha \to \infty$. ■

Having established this result, we are now ready to present a proof of Proposition 1.

Proof. (Proposition 1) For brevity of notation, we will notate the optimization problem in (15) as

$$\Omega_{\phi,\theta}(X) \equiv \inf_{\Phi_r(W^1,\ldots,W^K)=X} \sum_{i=1}^r \theta(W_i^1,\ldots,W_i^K), \tag{30}$$

where recall that r is variable although it is not explicitly notated.

1. By definition and the fact that θ is positive semidefinite, we always have $\Omega_{\phi,\theta}(X) \ge 0\ \forall X$. Trivially, $\Omega_{\phi,\theta}(0) = 0$ since we can always take $(W^1,\ldots,W^K) = (0,\ldots,0)$ to achieve the infimum. For $X \neq 0$, because (ϕ,θ) is a non-degenerate pair then $\sum_{i=1}^r \theta(W_i^1,\ldots,W_i^K) > 0$ for any $(W^1,\ldots,W^K)_r$ s.t. $\Phi_r(W^1,\ldots,W^K) = X$ and r finite. Property 5 shows that the infimum can be achieved with r finite, completing the result.

2. The result is easily seen from the positive homogeneity of ϕ and θ,

$$\begin{aligned}\Omega_{\phi,\theta}(\alpha X) &= \inf_{\Phi_r(W^1,\ldots,W^K)=\alpha X} \sum_{i=1}^r \theta(W_i^1,\ldots,W_i^K)\\ &= \inf_{\Phi_r(\alpha^{-1/p}W^1,\ldots,\alpha^{-1/p}W^K)=X} \sum_{i=1}^r \theta(W_i^1,\ldots,W_i^K)\\ &= \inf_{\Phi_r(Z^1,\ldots,Z^K)=X} \alpha \sum_{i=1}^r \theta(Z_i^1,\ldots,Z_i^K) = \alpha\Omega_{\phi,\theta}(X),\end{aligned} \tag{31}$$

   where the equality between the middle and final lines is simply due to the change of variables $(Z^1,\ldots,Z^K) = (\alpha^{-1/p}W^1,\ldots,\alpha^{-1/p}W^K)$.

3. If either $\Omega_{\phi,\theta}(X) = \infty$ or $\Omega_{\phi,\theta}(Z) = \infty$ then the inequality is trivially satisfied. Considering any (X,Z) pair such that $\Omega_{\phi,\theta}$ is finite for both X and Z, for any $\epsilon > 0$ let $(W^1,\ldots,W^K)_{r_x}$ be an ϵ optimal factorization of X. Specifically, $\Phi_{r_x}(W^1,\ldots,W^K) = X$ and $\sum_{i=1}^{r_x} \theta(W_i^1,\ldots,W_i^K) \le \Omega_{\phi,\theta}(X) + \epsilon$. Similarly, let $(\tilde{W}^1,\ldots,\tilde{W}^K)_{r_z}$ be an ϵ optimal factorization of Z. From the definition of Φ_r we have $\Phi_{r_x+r_z}([W^1\ \tilde{W}^1],\ldots,[W^K\ \tilde{W}^K]) = X + Z$, so $\Omega_{\phi,\theta}(X+Z) \le \sum_{i=1}^{r_x} \theta(W_i^1,\ldots,W_i^K) + \sum_{j=1}^{r_z} \theta(\tilde{W}_j^1,\ldots,\tilde{W}_j^K) \le \Omega_{\phi,\theta}(X) + \Omega_{\phi,\theta}(Z) + 2\epsilon$. Letting ϵ tend to 0 completes the result.

4. Convexity is given by the combination of properties 2 and 3. Further, note that properties 2 and 3 also show that $\{X \in \mathbb{R}^D : \Omega_{\phi,\theta}(X) < \infty\}$ is a convex set.

5. Let $\Gamma \subset \mathbb{R}^D$ be defined as

$$\Gamma = \{X : \exists (w^1, \ldots, w^K),\ \phi(w^1, \ldots, w^K) = X,\ \theta(w^1, \ldots, w^K) \leq 1\}. \tag{32}$$

Note that because (ϕ, θ) is a nondegenerate pair, for any non-zero $X \in \Gamma$ there exists $\alpha \in [1, \infty)$ such that αX is on the boundary of Γ, so Γ and its convex hull are compact sets.

Further, note that Γ contains the origin by definition of ϕ and θ, so as a result, we can define σ_Γ to be a gauge function on the convex hull of Γ,

$$\sigma_\Gamma(X) = \inf_\mu \{\mu : \mu \geq 0,\ X \in \mu \operatorname{conv}(\Gamma)\}. \tag{33}$$

Since the infimum w.r.t. μ is linear and constrained to a compact set, it must be achieved. Therefore, there must exist $\mu_{opt} \geq 0$, $\{\beta \in \mathbb{R}^{\operatorname{card}(X)} : \beta_i \geq 0\ \forall i,\ \sum_{i=1}^{\operatorname{card}(X)} \beta_i = 1\}$, and $\{(Z_i^1, \ldots, Z_i^K) : \phi(Z_i^1, \ldots, Z_i^K) \in \Gamma\}_{i=1}^{\operatorname{card}(X)}$ such that $X = \mu_{opt} \sum_{i=1}^{\operatorname{card}(X)} \beta_i \phi(Z_i^1, \ldots, Z_i^K)$ and $\sigma_\Gamma(X) = \mu_{opt}$.

Combined with positive homogeneity, this gives that σ_Γ can be defined identically to $\Omega_{\phi,\theta}$, but with the additional constraint $r \leq \operatorname{card}(X)$,

$$\sigma_\Gamma(X) \equiv \inf_{r \in [1, \operatorname{card}(X)]} \inf_{(W^1, \ldots, W^K)_r} \sum_{i=1}^r \theta(W^1, \ldots, W^K) \ \text{ s.t. } \ \Phi_r(W^1, \ldots, W^K) = X. \tag{34}$$

This is seen by noting that we can take $(W_i^1, \ldots, W_i^K) = ((\mu_{opt}\beta_i)^{1/p} Z_i^1, \ldots, (\mu_{opt}\beta_i)^{1/p} Z_i^K)$ to give

$$\mu_{opt} = \sigma_\Gamma(X) \leq \sum_{i=1}^{\operatorname{card}(X)} \theta(W_i^1, \ldots, W_i^K) = \mu_{opt} \sum_{i=1}^{\operatorname{card}(X)} \beta_i \theta(Z_i^1, \ldots, Z_i^K) \leq \mu_{opt} \sum_{i=1}^{\operatorname{card}(X)} \beta_i = \mu_{opt}, \tag{35}$$

and shows that a factorization of size $r \leq \operatorname{card}(X)$ which achieves the infimum $\mu_{opt} = \sigma_\Gamma(X)$ must exist. Clearly from (34) σ_Γ is very similar to $\Omega_{\phi,\gamma}$. To show that the two functions are, in fact, the same function, recall that the proof of the Fenchel dual of $\Omega_{\phi,\theta}$ given in Proposition 3 does not depend on the size of r but only on the existence (or non-existence) of a single $(w^1, \ldots, w^K)$ element. As a result, using an identical series of arguments to derive the Fenchel dual of σ_Γ, one finds that $\sigma_\Gamma^* = \Omega_{\phi,\theta}^*$, and since both σ_Γ and $\Omega_{\phi,\theta}$ are convex function, the one-to-one correspondence between convex functions and their Fenchel duals gives that $\sigma_\Gamma(X) = \Omega_{\phi,\theta}(X)$, completing the result.

■

Having established the convexity and Fenchel dual of $\Omega_{\phi,\theta}$ we are now ready to characterize the subgradient of $\Omega_{\phi,\theta}$ via the following result.

Proposition 4 *The subgradient of $\Omega_{\phi,\theta}(X)$ is given by*

$$\partial \Omega_{\phi,\theta}(X) = \{Z : \langle X, Z \rangle = \Omega_{\phi,\theta}(X),\ \Omega_{\phi,\theta}^\circ(Z) \leq 1\}. \tag{36}$$

Proof. Recall that because $\Omega_{\phi,\theta}(X)$ is convex then from Fenchel duality theory we have $W \in \partial\Omega_{\phi,\theta}(X) \iff \langle X, Z \rangle = \Omega_{\phi,\theta}(X) + \Omega_{\phi,\theta}^*(Z)$. From Proposition 3 we have that $\Omega_{\phi,\theta}^*(Z)$ is just the indicator function on the set $\{Z : \Omega_{\phi,\theta}^\circ(Z) \leq 1\}$, which gives the stated result. ■

From this simple result, we now have the basis for the following two lemmas which will be used in our main results.

Lemma 1 *Given a factorization $X = \Phi_r(W^1, \ldots, W^K)$ and a regularization function $\Omega_{\phi,\theta}(X)$, then the following conditions are equivalent:*

1. *$(W^1, \ldots, W^K)_r$ is an optimal factorization of X; i.e., $\sum_{i=1}^r \theta(W_i^1, \ldots, W_i^K) = \Omega_{\phi,\theta}(X)$.*
2. *$\exists Z$ such that $\Omega_{\phi,\theta}^\circ(Z) \leq 1$ and $\langle Z, \Phi_r(W^1, \ldots, W^K) \rangle = \sum_{i=1}^r \theta(W_i^1, \ldots, W_i^K)$.*
3. *$\exists Z$ such that $\Omega_{\phi,\theta}^\circ(Z) \leq 1$ and $\forall i \in \{1, \ldots, r\}$, $\langle Z, \phi(W_i^1, \ldots, W_i^K) \rangle = \theta(W_i^1, \ldots, W_i^K)$.*

Further, any Z which satisfies condition 2 or 3 satisfies both conditions 2 and 3 and $Z \in \partial\Omega_{\phi,\theta}(X)$.

Proof. $2 \iff 3$) 3 trivially implies 2 from the definition of Φ_r. For the opposite direction, recall from the proof of Proposition 3 that because $\Omega^\circ_{\phi,\theta}(Z) \leq 1$ we have $\langle Z, \phi(W_i^1,\dots,W_i^K)\rangle \leq \theta(W_i^1,\dots,W_i^K)\ \forall i$. Taking the sum over i, we can only achieve equality in 2 if we have equality $\forall i$ in condition 3. This also shows that any Z which satisfies condition 2 or 3 must also satisfy the other condition.

We next show that if W satisfies conditions 2/3 then $Z \in \partial\Omega_{\phi,\theta}(X)$. First, from condition 2/3 and the definition of $\Omega_{\phi,\theta}$, we have $\Omega_{\phi,\theta}(X) \leq \sum_{i=1}^r \theta(W_i^1,\dots,W_i^K) = \langle Z, X\rangle < \infty$. Thus, recall that because $\Omega_{\phi,\theta}(X)$ is convex and finite at X, we have $\langle Z, X\rangle \leq \Omega_{\phi,\theta}(X) + \Omega^*_{\phi,\theta}(Z)$ with equality iff $Z \in \partial\Omega_{\phi,\theta}(X)$. Now, by contradiction assume Z satisfies conditions 2/3 but $Z \notin \partial\Omega_{\phi,\theta}(X)$. From condition 2/3 we have $\Omega^*_{\phi,\theta}(Z) = 0$, so $\Omega_{\phi,\theta}(X) = \Omega_{\phi,\theta}(X) + \Omega^*_{\phi,\theta}(Z) > \langle X, Z\rangle = \sum_{i=1}^r \theta(W_i^1,\dots,W_i^K)$ which contradicts the definition of $\Omega_{\phi,\theta}(X)$.

$1 \implies 2$) Any $Z \in \partial\Omega_{\phi,\theta}(X)$ satisfies $\langle X, Z\rangle = \Omega_{\phi,\theta}(X) + \Omega^*_{\phi,\theta}(Z) = \sum_{i=1}^r \theta(W_i^1,\dots,W_i^K)$.

$2 \implies 1$) By contradiction, assume $(W^1,\dots,W^K)_r$ was not an optimal factorization of X. This gives, $\Omega_{\phi,\theta}(X) < \sum_{i=1}^r \theta(W_i^1,\dots,W_i^K) = \langle Z, X\rangle = \Omega_{\phi,\theta}(X) + \Omega^*_{\phi,\theta}(Z) = \Omega_{\phi,\theta}(X)$, producing the contradiction. ■

Lemma 2 *If $(W^1,\dots,W^K)$ is a local minimum of $f_r(W^1,\dots,W^K)$ as given in (17), then for any $\beta \in \mathbb{R}^r$*

$$\left\langle -\tfrac{1}{\lambda}\nabla_X \ell(Y, \Phi_r(W^1,\dots,W^K)), \sum_{i=1}^r \beta_i \phi(W_i^1,\dots,W_i^K)\right\rangle = \sum_{i=1}^r \beta_i \theta(W_i^1,\dots,W_i^K). \tag{37}$$

Proof. Let $(Z_i^1,\dots,Z_i^K) = (\beta_i W_i^1,\dots,\beta_i W_i^K)$ for all $i \in \{1\dots r\}$ and let $\Lambda = \sum_{i=1}^r \beta_i \phi(W_i^1,\dots,W_i^K)$. From positive homogeneity and the fact that we have a local minimum, then $\exists \delta > 0$ such that $\forall \epsilon \in (0,\delta)$ we must have

$$f_r(W^1,\dots,W^K) \leq f_r(W^1 + \epsilon Z^1,\dots,W^K + \epsilon Z^K) \implies \tag{38}$$

$$\begin{aligned}&\ell(Y, \Phi_r(W^1,\dots,W^K)) + \lambda \sum_{i=1}^r \theta(W_i^1,\dots,W_i^K) \leq \\ &\ell\left(Y, \sum_{i=1}^r (1+\epsilon\beta_i)^p \phi(W_i^1,\dots,W_i^K)\right) + \lambda \sum_{i=1}^r (1+\epsilon\beta_i)^p \theta(W_i^1,\dots,W_i^K).\end{aligned} \tag{39}$$

Taking the first order approximation $(1+\epsilon\beta_i)^p = 1 + p\epsilon\beta_i + O(\epsilon^2)$ and rearranging the terms of (39), we arrive at

$$0 \leq \ell\left(Y, \Phi_r(W^1,\dots,W^K) + p\epsilon\Lambda + O(\epsilon^2)\right) - \ell(Y, \Phi_r(W^1,\dots,W^K)) + p\epsilon\lambda \sum_{i=1}^r \beta_i \theta(W_i^1,\dots,W_i^K) + O(\epsilon^2), \tag{40}$$

After dividing by ϵ and taking $\lim_{\epsilon \searrow 0}[\frac{(40)}{\epsilon}]$, we note that the difference in the $\ell(\cdot,\cdot)$ terms gives the one-sided directional derivative $d\ell(Y, \Phi_r(W^1,\dots,W^K))(p\Lambda)$, thus from the differentiability of ℓ we get

$$0 \leq \langle \nabla_X \ell(Y, \Phi_r(W^1,\dots,W^K)), p\Lambda\rangle + p\lambda \sum_{i=1}^r \beta_i \theta(W_i^1,\dots,W_i^K). \tag{41}$$

Noting that for $\epsilon > 0$ but sufficiently small, we also must have $f_r(W^1,\dots,W^K) \leq f_r(W^1 - \epsilon Z^1,\dots,W^K - \epsilon Z^K)$, using identical steps as before and taking the first order approximation $(1-\epsilon\beta_i)^p = 1 - p\epsilon\beta_i + O(\epsilon^2)$, we get

$$0 \leq \ell(Y, \Phi_r(W^1,\dots,W^K) - p\epsilon\Lambda + O(\epsilon^2)) - \ell(Y, \Phi_r(W^1,\dots,W^K)) - p\epsilon\lambda \sum_{i=1}^r \beta_i \theta(W_i^1,\dots,W_i^K) + O(\epsilon^2). \tag{42}$$

Dividing by ϵ and taking the limit $\lim_{\epsilon \searrow 0}[\frac{(42)}{\epsilon}]$, we arrive at

$$0 \leq \langle \nabla_X \ell(Y, \Phi_r(W^1,\dots,W^K)), -p\Lambda\rangle - p\lambda \sum_{i=1}^r \beta_i \theta(W_i^1,\dots,W_i^K) \tag{43}$$

Combining (41) and (43) and rearranging terms gives the result. ■

With these preliminary results, we are now ready present the proof of our first Theorem.

Proof. (Theorem 1) We begin by noting that from the definition of $\Omega_{\phi,\theta}(X)$, for any factorization $X = \Phi_r(W^1,\dots,W^K)$

$$F(X) = \ell(Y,X) + \lambda\Omega_{\phi,\theta}(X) \leq \ell(Y,\Phi_r(W^1,\dots,W^K)) + \lambda\sum_{i=1}^r \theta(W_i^1,\dots,W_i^K) = f_r(W^1,\dots,W^K) \tag{44}$$

with equality at any factorization which achieves the infimum in (15). We will show that a local minimum of $f_r(W^1,\dots,W^K)$ satisfying the conditions of the theorem also satisfies the conditions for $(\Phi_r(W^1,\dots,W^K))$ to be a global minimum of the convex function $F(X)$, which implies a global minimum of $f_r(W^1,\dots,W^K)$ due to the global bound in (44).

First, because (16) is a convex function, a simple subgradient condition gives that X is a global minimum of $F(X)$ iff the following condition is satisfied

$$-\tfrac{1}{\lambda}\nabla_X \ell(Y,X) \in \partial\Omega_{\phi,\theta}(X) \tag{45}$$

where $\nabla_X \ell(Y,X)$ denotes the gradient of $\ell(Y,X)$ w.r.t. X.

Turning to the factorization objective, if $(W^1,\dots,W^K)$ is a local minimum of $f_r(W^1,\dots,W^K)$, then $\forall (Z^1,\dots,Z^K)_r$ there exists $\delta > 0$ such that $\forall \epsilon \in (0,\delta)$ we have $f_r(W^1+\epsilon^{1/p}Z^1,\dots,W^K+\epsilon^{1/p}Z^K) \geq f_r(W^1,\dots,W^K)$. If we now consider search directions $(Z^1,\dots,Z^K)_r$ of the form

$$(Z_j^1,\dots,Z_j^K) = \begin{cases} (0,\dots,0) & j \neq i_0 \\ (z^1,\dots,z^K) & j = i_0 \end{cases}, \tag{46}$$

where i_0 is the index such that $(W_{i_0}^1,\dots,W_{i_0}^K) = (0,\dots,0)$, then for $\epsilon \in (0,\delta)$, we have

$$\ell(Y,\Phi_r(W^1,\dots,W^K)) + \lambda\sum_{i=1}^r \theta(W_i^1,\dots,W_i^K) \leq \tag{47}$$

$$\ell(Y,\Phi_r(W^1+\epsilon^{1/p}Z^1,\dots,W^K+\epsilon^{1/p}W^K)) + \lambda\sum_{i=1}^r \theta(W_i^1+\epsilon^{1/p}Z_i^1,\dots,W_i^K+\epsilon^{1/p}Z_i^K) = \tag{48}$$

$$\begin{aligned}&\ell(Y,\sum_{i\neq i_0}\phi(W_i^1,\dots,W_i^K) + \phi(W_{i_0}^1+\epsilon^{1/p}Z_{i_0}^1,\dots,W_{i_0}^K+\epsilon^{1/p}Z_{i_0}^K)) + \\ &\quad\lambda\sum_{i\neq i_0}\theta(W_i^1,\dots,W_i^K) + \lambda\theta(W_{i_0}^1+\epsilon^{1/p}Z_{i_0}^1,\dots,W_{i_0}^K+\epsilon^{1/p}Z_{i_0}^K) =\end{aligned} \tag{49}$$

$$\ell(Y,\Phi_r(W^1,\dots,W^K) + \epsilon\phi(z^1,\dots,z^K)) + \lambda\sum_{i=1}^r \theta(W_i^1,\dots,W_i^K) + \epsilon\lambda\theta(z^1,\dots,z^K). \tag{50}$$

The equality between (49) and (50) comes from the special form of Z given by (46), the fact that $(W_{i_0}^1,\dots,W_{i_0}^K) = (0,\dots,0)$, and the positive homogeneity of ϕ and θ. Rearranging terms, we now have

$$\epsilon^{-1}[\ell(Y,\Phi_r(W^1,\dots,W^K) + \epsilon\phi(z^1,\dots,z^K)) - \ell(Y,\Phi_r(W^1,\dots,W^K))] \geq -\lambda\theta(z^1,\dots,z^K). \tag{51}$$

Taking the limit of (51) as $\epsilon \searrow 0$, we note that the left side of the inequality is simply the definition of the one-sided directional derivative of $\ell(Y,\Phi_r(W^1,\dots,W^K))$ in the direction $(\phi(z^1,\dots,z^K))$, which combined with the differentiability of $\ell(Y,X)$, gives

$$\langle \phi(z^1,\dots,z^K), \nabla_X\ell(Y,\Phi_r(W^1,\dots,W^K))\rangle \geq -\lambda\theta(z^1,\dots,z^K). \tag{52}$$

Because $(z^1,\dots,z^K)$ was arbitrary, we have established that

$$\begin{aligned}&\langle \phi(z^1,\dots,z^K), -\tfrac{1}{\lambda}\nabla_X\ell(Y,\Phi_r(W^1,\dots,W^K))\rangle \leq \theta(z^1,\dots,z^K) \;\; \forall (z^1,\dots,z^K) \\ &\iff \Omega_{\phi,\theta}^\circ(-\tfrac{1}{\lambda}\nabla_X\ell(Y,\Phi_r(W^1,\dots,W^K))) \leq 1,\end{aligned} \tag{53}$$

where the equivalence is seen by identical arguments to those used in the proof of Proposition 3. Further, if we choose β to be vector of all ones in Lemma 2, we get

$$\sum_{i=1}^{r} \theta(W_i^1, \dots, W_i^K) = \left\langle \Phi_r(W^1, \dots, W^K), -\tfrac{1}{\lambda}\nabla_X \ell(Y, \Phi_r(W^1, \dots, W^K)) \right\rangle . \tag{54}$$

This fact, combined with (53), Lemma 1, and Proposition 4 shows that $-\frac{1}{\lambda}\nabla_X \ell(Y, \Phi_r(W^1, \dots, W^K)) \in \partial\Omega_{\phi,\theta}(\Phi_r(W^1, \dots, W^K))$, completing the result. ■

With this result, we next show the proof of Corollary 1.

Proof. (Corollary 1) Note that from the structure of $f_r(W^1, \dots, W^K)$ the following two problems are equivalent

$$\begin{aligned} &\min_{(W^1,\dots,W^K)_r} f_r(W^1, \dots, W^K) \equiv \\ &\min_{([W^1\ w^1],\dots,[W^K\ w^K])} f_{r+1}([W^1\ w^1], \dots, [W^K\ w^k]) \text{ s.t. } (w^1, \dots, w^K) = (0, \dots, 0). \end{aligned} \tag{55}$$

If we remove the equality constraint we then have that $\min f_{r+1} \leq \min f_r$, and if the condition of the corollary is satisfied, then $([W^1\ 0], \dots, [W^K\ 0])$ is a global minimizer for f_{r+1} due to Theorem 1. This then implies that $(W^1, \dots, W^K)$ is global minimizer of f_r due to the equivalence in (55). ■

Finally, we conclude our results with a proof of the second Theorem.

Proof. (Theorem 2) Clearly if $(Z^1, \dots, Z^K)$ is not a local minimum, then we can follow a decreasing path until we reach a local minimum. Having arrived at a local minimum, $(\tilde{W}^1, \dots, \tilde{W}^K,)$, if $(\tilde{W}_i^1, \dots, \tilde{W}_i^K) = (0, \dots, 0)$ for any $i \in \{1, \dots, r\}$ then from Theorem 1 we must be at a global minimum. Similarly, if for any $i_0 \in \{1, \dots, r\}$ we have $\phi(\tilde{W}_{i_0}^1, \dots, \tilde{W}_{i_0}^K) = 0$ then we can scale the slice $(\alpha\tilde{W}_{i_0}^1, \dots, \alpha\tilde{W}_{i_0}^K)$ as α goes from $1 \to 0$ without increasing the objective function. Once $\alpha = 0$ we will then have an all zero slice in the factor tensors, so from Theorem 1 we are either at a global minimum or a local descent direction must exist from that point. We are thus left to show that a non-increasing path to a global minimizer must exist from any local minima such that $\phi(\tilde{W}_i^1, \dots, \tilde{W}_i^K) \neq 0$ for all $i \in \{1, \dots, r\}$.

First, note that because $r > \text{card}(X)$ there must exist $\hat{\beta} \in \mathbb{R}^r$ such that $\hat{\beta} \neq 0$ and $\sum_{i=1}^r \hat{\beta}_i \phi(\tilde{W}_i^1, \dots, \tilde{W}_i^K) = 0$. Further, from Lemma 2 we must have that $\sum_{i=1}^r \hat{\beta}_i \theta(\tilde{W}_i^1, \dots, \tilde{W}_i^K) = \left\langle -\frac{1}{\lambda}\nabla_X \ell(Y, \Phi_r(W^1, \dots, W^K)), \sum_{i=1}^r \hat{\beta}_i \phi(W_i^1, \dots, W_i^K) \right\rangle = 0$. Due to the non-degeneracy of the (ϕ, θ) pair we must have $\theta(\tilde{W}_i^1, \dots, \tilde{W}_i^K) > 0,\ \ \forall i \in \{1, \dots, r\}$, which implies that at least one entry of $\hat{\beta}$ must be strictly less than zero.

Without loss of generality, assume $\hat{\beta}$ is scaled so that $\min_i \hat{\beta}_i = -1$. Now, for all $(\gamma, i) \in \{[0, 1]\} \times \{1, \dots, r\}$, let us define

$$(R_i^1(\gamma), \dots, R_i^K(\gamma)) \equiv ((1 + \gamma\hat{\beta}_i)^{1/p}\tilde{W}_i^1, \dots, (1 + \gamma\hat{\beta}_i)^{1/p}\tilde{W}_i^K) \tag{56}$$

where p is the degree of positive homogeneity of (ϕ, θ). Note that by construction $(R^1(0), \dots, R^K(0)) = (\tilde{W}^1, \dots, \tilde{W}^K)$ and that for $\gamma = 1$ there must exist $i_0 \in \{1, \dots, r\}$ such that $(R_{i_0}^1(1), \dots, R_{i_0}^K(1)) = (0, \dots, 0)$.

Further, from the positive homogeneity of (ϕ, θ) we have $\forall\gamma \in [0, 1]$

$$\begin{aligned} f_r(R^1(\gamma), \dots, R^K(\gamma)) =& \ell\left(Y, \sum_{i=1}^r \phi(\tilde{W}_i^1, \dots, \tilde{W}_i^K) + \gamma\sum_{i=1}^r \hat{\beta}_i \phi(\tilde{W}_i^1, \dots, \tilde{W}_i^K), \right) + \\ & \lambda\gamma\sum_{i=1}^r \hat{\beta}_i\theta(\tilde{W}_i^1, \dots, \tilde{W}_i^K) + \lambda\sum_{i=1}^r \theta(\tilde{W}_i^1, \dots, \tilde{W}_i^K) \end{aligned} \tag{57}$$

$$= \ell(Y, \Phi_r(\tilde{W}^1, \dots, \tilde{W}^K)) + \lambda\sum_{i=1}^r \theta(\tilde{W}_i^1, \dots, \tilde{W}_i^K) \tag{58}$$

$$= f_r(\tilde{W}^1, \dots, \tilde{W}^K), \tag{59}$$

where the equality between (57) and (58) is seen by recalling that $\sum_{i=1}^r \hat{\beta}_i \phi(\tilde{W}_i^1, \dots, \tilde{W}_i^K) = 0$ and $\sum_{i=1}^r \hat{\beta}_i \theta(\tilde{W}_i^1, \dots, \tilde{W}_i^K) = 0$.

As a result, as γ goes from $0 \to 1$ we can traverse a path from $(\tilde{W}^1, \dots, \tilde{W}^K) \to (R^1(1), \dots, R^K(1))$ without changing the value of f_r. Also recall that by construction $(R_{i_0}^1(1), \dots, R_{i_0}^K(1)) = (0, \dots, 0)$, so if $(R^1(1), \dots, R^K(1))$ is a local minimizer of f_r then it must be a global minimizer due to Theorem 1. If $(R^1(1), \ldots, R^K(1))$ is not a local minimizer then there must exist a descent direction and we can iteratively apply this result until we reach a global minimizer, completing the proof. ■