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, $\Phi(W^1, \dots, W^K)$, which is positively homogeneous with degree p (but which is not assumed to have the parallel network form). Now, consider a general regularization function, $\Theta(W^1, \dots, W^K)$, 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 $\Phi : \mathbb{R}^{D^1} \times \dots \times \mathbb{R}^{D^K} \to \mathbb{R}^D$ be a positively homogeneous mapping with degree p; and let $\Theta : \mathbb{R}^{D^1} \times \dots \times \mathbb{R}^{D^K} \to \mathbb{R}_+$ be a positively homogeneous function with degree $p' < p$ such that $\Theta(0, \dots, 0) = 0$ and $\Theta(W^1, \dots, W^K) > 0 \ \forall \{(W^1, \dots, W^K) : \Phi(W^1, \dots, W^K) \neq 0\}$. Then, the optimization problem given by*

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

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

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

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

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

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

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

The above proposition shows that unless we have the special case where $(W^1, \dots, W^K) = (0, \dots, 0)$ 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

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

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

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

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

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

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

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

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

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

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 $\Omega_{\phi,\theta}(X) \geq 0 \ \forall X$. Trivially, $\Omega_{\phi,\theta}(0) = 0$ since we can always take $(W^1,\dots,W^K) = (0,\dots,0)$ to achieve the infimum. For $X \neq 0$, because (ϕ,θ) is a non-degenerate pair then $\sum_{i=1}^r \theta(W_i^1,\dots,W_i^K) > 0$ for any $(W^1,\dots,W^K)_r$ s.t. $\Phi_r(W^1,\dots,W^K) = X$ 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,

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

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

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 $(W^1,\dots,W^K)_{r_x}$ be an ϵ optimal factorization of X. Specifically, $\Phi_{r_x}(W^1,\dots,W^K) = X$ and $\sum_{i=1}^{r_x} \theta(W_i^1,\dots,W_i^K) \leq \Omega_{\phi,\theta}(X) + \epsilon$. Similarly, let $(\tilde{W}^1,\dots,\tilde{W}^K)_{r_z}$ be an ϵ optimal factorization of Z. From the definition of Φ_r we have $\Phi_{r_x+r_z}([W^1\ \tilde{W}^1],\dots,[W^K\ \tilde{W}^K]) = X + Z$, so $\Omega_{\phi,\theta}(X+Z) \leq \sum_{i=1}^{r_x} \theta(W_i^1,\dots,W_i^K) + \sum_{j=1}^{r_z} \theta(\tilde{W}_j^1,\dots,\tilde{W}_j^K) \leq \Omega_{\phi,\theta}(X) + \Omega_{\phi,\theta}(Z) + 2\epsilon$. 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

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

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 $\mu_{opt} \geq 0$, $\{\beta \in \mathbb{R}^{\text{card}(X)} : \beta_i \geq 0\ \forall i,\ \sum_{i=1}^{\text{card}(X)} \beta_i = 1\}$, and $\{(Z_i^1, \dots, Z_i^K) : \phi(Z_i^1, \dots, Z_i^K) \in \Gamma\}_{i=1}^{\text{card}(X)}$ such that $X = \mu_{opt} \sum_{i=1}^{\text{card}(X)} \beta_i \phi(Z_i^1, \dots, Z_i^K)$ 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 $r \leq \text{card}(X)$,

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

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

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

and shows that a factorization of size $r \leq \text{card}(X)$ 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 $(w^1, \dots, w^K)$ 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 $X = \Phi_r(W^1, \dots, W^K)$ and a regularization function $\Omega_{\phi,\theta}(X)$, then the following conditions are equivalent:*

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

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 $\left\langle Z, \phi(W_i^1, \ldots, W_i^K)\right\rangle \leq \theta(W_i^1, \ldots, W_i^K)\ \ \forall i$. 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 $\Omega_{\phi,\theta}(X) \leq \sum_{i=1}^r \theta(W_i^1, \ldots, W_i^K) = \langle Z, X\rangle < \infty$. 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 $\Omega_{\phi,\theta}(X) = \Omega_{\phi,\theta}(X) + \Omega^*_{\phi,\theta}(Z) > \langle X, Z\rangle = \sum_{i=1}^r \theta(W_i^1, \ldots, W_i^K)$ which contradicts the definition of $\Omega_{\phi,\theta}(X)$.

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

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

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

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

Proof. Let $(Z_i^1, \ldots, Z_i^K) = (\beta_i W_i^1, \ldots, \beta_i W_i^K)$ for all $i \in \{1 \ldots r\}$ and let $\Lambda = \sum_{i=1}^r \beta_i \phi(W_i^1, \ldots, W_i^K)$. 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

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

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

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

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

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 $d\ell(Y, \Phi_r(W^1, \ldots, W^K))(p\Lambda)$, thus from the differentiability of ℓ we get

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

Noting that for $\epsilon > 0$ but sufficiently small, we also must have $f_r(W^1, \ldots, W^K) \leq f_r(W^1 - \epsilon Z^1, \ldots, W^K - \epsilon Z^K)$, 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

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

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

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

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

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

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

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

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. ■