Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models

Mor Shpigel Nacson [1] Suriya Gunasekar [2] Jason D. Lee [3] Nathan Srebro [2] Daniel Soudry [1]

Abstract

With an eye toward understanding complexity control in deep learning, we study how infinitesimal regularization or gradient descent optimization lead to margin maximizing solutions in both homogeneous and *non homogeneous* models, extending previous work that focused on infinitesimal regularization only in homogeneous models. To this end we study the limit of loss minimization with a diverging norm constraint (the "constrained path"), relate it to the limit of a "margin path" and characterize the resulting solution. For non-homogeneous ensemble models, which output is a sum of homogeneous sub-models, we show that this solution discards the shallowest sub-models if they are unnecessary. For homogeneous models, we show convergence to a "lexicographic max-margin solution", and provide conditions under which max-margin solutions are also attained as the limit of unconstrained gradient descent.

1. Introduction

Inductive bias introduced through the learning process plays a crucial role in training deep neural networks and in the generalization properties of the learned models (Neyshabur et al., 2015b;a; Zhang et al., 2017; Keskar et al., 2017; Neyshabur et al., 2017; Wilson et al., 2017; Hoffer et al., 2017). Deep neural networks used in practice are typically highly overparameterized, i.e., have far more trainable parameters than training examples. Thus, using these models, it is usually possible to fit the data perfectly and obtain zero training error (Zhang et al., 2017). However, simply minimizing the training loss does not guarantee good generalization to unseen data – many global minima of the training loss indeed have very high test error (Wu et al., 2017). The inductive bias introduced in our learning process affects which specific global minimizer is chosen as the predictor. Therefore, it is essential to understand the nature of this inductive bias to understand why overparameterized models, and particularly deep neural networks, exhibit good generalization abilities.

A common way to introduce an additional inductive bias in overparameterized models is via small amounts of regularization, or loose constraints . For example, Rosset et al. (2004b;a); Wei et al. (2018) show that, in overparameterized classification models, a vanishing amount of regularization, or a diverging norm constraint can lead to max-margin solutions, which in turn enjoy strong generalization guarantees.

A second and more subtle source of inductive bias is via the optimization algorithm used to minimize the underdetermined training objective (Gunasekar et al., 2017; Soudry et al., 2018b). Common algorithms used in neural network training, such as stochastic gradient descent, iteratively refine the model parameters by making incremental local updates. For different algorithms, the local updates are specified by different geometries in the space of parameters. For example, gradient descent uses an Euclidean ℓ_2 geometry, while coordinate descent updates are specified in the ℓ_1 geometry. The minimizers to which such local search based optimization algorithms converge to are indeed very special and are related to the geometry of the optimization algorithm (Gunasekar et al., 2018b) as well as the choice of model parameterization (Gunasekar et al., 2018a).

In this work we similarly investigate the connection between margin maximization and the limits of

- The "optimization path" of unconstrained, unregularized gradient descent.
- The "constrained path", where we optimize with a diverging (increasingly loose) constraint on the norm of the parameters.
- The closely related "regularization path", of solutions with decreasing penalties on the norm.

[1]Technion, Israel [2]TTI Chicago, USA [3]USC Los Angeles, USA. Correspondence to: Mor Shpigel Nacson <morshpigel@google.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

To better understand the questions we tackle in this paper, and our contribution toward understanding the inductive bias introduced in training, let us briefly survey prior work.

Equivalence of the regularization or constrained paths and margin maximization: Rosset et al. (2004b;a); Wei et al. (2018) investigated the connection between the regularization and constrained paths and the max-margin solution. Rosset et al. (2004a;b) considered linear (hence homogeneous) models with monotone loss and explicit norm regularization or constraint, and proved convergence to the max-margin solution for certain loss functions (e.g., logistic loss) as the regularization vanishes or the norm constraint diverges. Wei et al. (2018) extended the regularization path result to non-linear but positive-homogeneous prediction functions,

Definition 1 (α-positive homogeneous function)**.** *A function $g(\boldsymbol{\theta}) : \mathbb{R}^d \rightarrow \mathbb{R}$ is α-positive homogeneous if $\forall \rho > 0$ and $\forall \theta \in \mathbb{R}^d :\ g(\rho\boldsymbol{\theta}) = \rho^\alpha g(\boldsymbol{\theta})$.*

e.g. as obtained by a ReLU network with uniform depth.

These results are thus limited to only positive homogeneous predictors, and do not include deep networks with bias parameters, ensemble models with different depths, ResNets, or other models with skip connections. Here, we extend this connection beyond positive homogeneous predictors.

Furthermore, even for homogeneous or linear predictors, there might be multiple margin maximizing solutions. For linear models, Rosset et al. (2004b) alluded to a refined set of maximum margin classifiers that in addition to maximizing the distance to the closest data point (max-margin), also maximize the distance to the second closest data point, and so on. We formulate such special maximum margin solutions as "lexicographic max-margin" classifiers which we introduce in Section 4.2. We show that for general continuous homogeneous models, the constrained path with diverging norm constraint converges to these more refined "lexicographic max-margin" classifiers.

Equivalence of the optimization path and margin maximization: Another line of works studied the connection between unconstrained, unregularized optimization with a specific algorithm (i.e., the limit of the "optimization path"), and the max-margin solution. For linear prediction with the logistic loss (or other exponential tail losses), we now know gradient descent (Soudry et al., 2018b; Ji & Telgarsky, 2018) as well as SGD (Nacson et al., 2019b) converges in direction to the max-margin solution, while steepest descent with respect to an arbitrary norm converges to the max-margin w.r.t. the corresponding norm (Gunasekar et al., 2018b). All the above results are for linear prediction. Gunasekar et al. (2018a); Nacson et al. (2019a); Ji & Telgarsky (2019) obtained results establishing convergence to margin maximizing solutions also for certain uniform-depth linear networks (including fully connected networks and convolutional networks), which still implement linear model. Separately, Xu et al. (2019) analyzed a single linear unit with ReLU activation—a limited non-linear but still positive homogeneous model. Lastly, Soudry et al. (2018a) analyzed a non-linear ReLU network where only a single weight layer is optimized.

Here, we extend this relationship to general, non-linear and positive homogeneous predictors for which the loss can be minimized only at infinity. We establish a connection between the limit of unregularized unconstrained optimization and the max-margin solution.

Problems with finite minimizers: We note that the connection between regularization path and optimization path was previously considered in a different settings, where a finite (global) minimum exists. In such settings the questions asked are different than the ones we consider here, and are not about the limit of the paths. E.g., Ali et al. (2018) showed for gradient flow a multiplicative relation between the risk for the gradient flow optimization path and the ridge-regression regularization path. Also, Suggala et al. (2018) showed that for gradient flow and strongly convex and smooth loss function – gradient descent iterates on the unregularized loss function are pointwise close to solutions of a corresponding regularized problem.

Contributions

We examine overparameterized realizable problems (i.e., where it is possible to perfectly classify the training data), when training using monotone decreasing classification loss functions. For simplicity, we focus on the exponential loss. However, using similar techniques as in Soudry et al. (2018a) our results should extend to other exponential-tailed loss functions such as the logistic loss and its multi-class generalization. This is indeed the common setting for deep neural networks used in practice.

We show that in any model,

- As long as the margin attainable by a (unregularized, unconstrained) model is unbounded, then the margin of the constrained path converges to the max-margin. See Corollary 1.
- If additional conditions hold, the constrained path also converges to the "margin path" in parameter space (the path of minimal norm solutions attaining increasingly large margins). See section 3.1.

We then demonstrate that

- If the model is a sum of homogeneous functions of different orders (i.e., it is not homogeneous itself), then we can still characterize the asymptotic solution of both the constrained path and the margin path. See Theorem 3.2.
- This solution implies that in an ensemble of homogeneous neural networks, the ensemble will aim to discard the most *shallow* network. This is in contrast to what we would expect from considerations of optimization difficulty (since deeper networks are typically harder to train (He et al., 2016)).
- This also allows us to represent hard-margin SVM problems **with unregularized bias** using such models. This is in contrast to previous approaches which fail to do so, as pointed out recently (Nar et al., 2019).

Finally, for homogeneous models,

- We find general conditions under which the optimization path converges to stationary points of the margin path or the constrained path. See section 4.1.
- We show that the constrained path converges to a specific type max-margin solution, which we term the "lexicographic max-margin". [1] See Theorem 4.

2. Preliminaries and Basic Results

In this paper, we will study the following exponential tailed loss function

$$\mathcal{L}\left(\boldsymbol{\theta}\right)\triangleq\sum_{n=1}^{N}\exp\left(-f_{n}\left(\boldsymbol{\theta}\right)\right)\,, \tag{1}$$

where $f_n:\mathbb{R}^d\rightarrow\mathbb{R}$ is a continuous function, and N is the number of samples. Also, for any norm $\|\cdot\|$ in $\mathbb{R}^d$ we define $\mathbb{S}^{d-1}$ as the unit norm ball in $\mathbb{R}^d$.

We will use in our results the following basic lemma

Lemma 1. *Let f and g be two functions from $\mathbb{R}^d$ to $\mathbb{R}$, such that*

$$\phi(\rho)=\min_{\mathbf{w}\in\mathbb{R}^d}f(\mathbf{w})\text{ s.t. }g(\mathbf{w})\leq\rho \tag{2}$$

exists and is strictly monotonically decreasing in ρ, $\forall\rho\geq\rho_0$, for some ρ_0. Then, $\forall\rho\geq\rho_0$, the optimization problem in eq. 2 has the same set of solutions $(\mathbf{w})$ as

$$\min_{\mathbf{w}\in\mathbb{R}^d}g(\mathbf{w})\text{ s.t. }f(\mathbf{w})\leq\phi(\rho)\,, \tag{3}$$

whose minimum is obtained at $g(\mathbf{w})=\rho$.

Proof. See Appendix A. □

2.1. The Optimization Path

The optimization path in the Euclidean norm $\boldsymbol{\theta}\left(t\right)$, is given by the direction of iterates of gradient descent algorithm with initialization $\boldsymbol{\theta}\left(0\right)$ and learning rates $\{\eta_t\}_{t=1}^{\infty}$,

Optimization path: $\bar{\boldsymbol{\theta}}(t)=\dfrac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|}$, where

$$\boldsymbol{\theta}\left(t\right)=\boldsymbol{\theta}\left(t-1\right)-\eta_t\nabla\mathcal{L}\left(\boldsymbol{\theta}\left(t-1\right)\right). \tag{4}$$

2.2. The Constrained Path

The constrained path for the loss in eq. 1 is given by minimizer of the loss at a given norm value $\rho>0$, i.e.,

Constrained path: $\Theta_c\left(\rho\right)\triangleq\arg\min\limits_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}}\mathcal{L}\left(\rho\boldsymbol{\theta}\right)\,.$ (5)

The constrained path was previously considered for linear models (Rosset et al., 2004a). However, most previous works (e.g. Rosset et al. (2004b); Wei et al. (2018)) focused on the regularization path, which is the minimizer of the regularized loss. These two paths are closely linked, as we discuss in more detail in Appendix F.

Denote the constrained minimum of the loss as follows:

$$\mathcal{L}^*\left(\rho\right)\triangleq\min_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}}\mathcal{L}\left(\rho\boldsymbol{\theta}\right).$$

$\mathcal{L}^*\left(\rho\right)$ exists for any finite ρ as the minimum of a continuous function on a compact set.

By Lemma 1, the Assumption

Assumption 1. *There exists ρ_0 such that $\mathcal{L}^*\left(\rho\right)$ is strictly monotonically decreasing to zero for any $\rho\geq\rho_0$.*

enables an alternative form of the constrained path

$$\Theta_c\left(\rho\right)=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\left\|\boldsymbol{\theta}\right\|^2\text{ s.t. }\mathcal{L}\left(\rho\boldsymbol{\theta}\right)\leq\mathcal{L}^*\left(\rho\right)\,.$$

In addition, in the next lemma we show that under this assumption the constrained path minimizers are obtained on the boundary of $\mathbb{S}^{d-1}$.

Lemma 2. *Under assumption 1, for all $\rho>\rho_0$ and for all $\theta_c\in\Theta_c\left(\rho\right)$, we have $\|\boldsymbol{\theta}_c\|=1$.*

Proof. Let $\rho>0$. We assume, in contradiction, that $\exists\theta_c\in\Theta_c\left(\rho\right)$ so that $\|\boldsymbol{\theta}_c\|=b<1$. This implies that $\mathcal{L}^*\left(\rho\right)=\mathcal{L}^*\left(\rho b\right)$ which contradicts our assumption that $\mathcal{L}^*\left(\rho\right)$ is strictly monotonically decreasing. □

2.3. The Margin Path

For prediction functions $f_n:\mathbb{R}^d\rightarrow\mathbb{R}$ on data points indexed as $n=1,2,\ldots,N$, we define the margin path as:

Margin path: $\Theta_m\left(\rho\right)\triangleq\arg\max\limits_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}}\min\limits_{n}f_n\left(\rho\boldsymbol{\theta}\right)\,.$ (6)

[1] The authors thank Rob Shapire for the suggestion of the nomenclature during initial discussions.

For $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$, we denote the margin at scaling $\rho > 0$ as

$$\gamma\left(\rho, \boldsymbol{\theta}\right) = \min_n f_n\left(\rho\boldsymbol{\theta}\right) ,$$

and the max-margin at scale of $\rho > 0$ as

$$\gamma^*\left(\rho\right) = \max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \min_n f_n\left(\rho\boldsymbol{\theta}\right) .$$

Note that for all ρ, this maximum exists as the maximum of a continuous function on a compact set.

Again, we make a simplifying assumption

Assumption 2. *There exist ρ_0 such that $\gamma^*\left(\rho\right)$ is strictly monotonically increasing to ∞ for any $\rho \geq \rho_0$.*

Many common prediction functions satisfy this assumption, including the sum of positive-homogeneous prediction functions.

Using Lemma 1 with Assumption 2, we have:

$$\begin{aligned}\Theta_m\left(\rho\right) &= \arg\max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \min_n f_n\left(\rho\boldsymbol{\theta}\right) &(7)\\ &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{d}} \left\|\boldsymbol{\theta}\right\|^2 \text{ s.t. } \min_n f_n\left(\rho\boldsymbol{\theta}\right) \geq \gamma^*\left(\rho\right) .\end{aligned}$$

3. Non-Homogeneous Models

We first study the constrained path in non-homogeneous models, and relate it to the margin path. To do so, we need to first define the ϵ-ball surrounding a set $\mathcal{A} \subset \mathbb{R}^d$

$$\mathcal{B}_\epsilon\left(\mathcal{A}\right) \triangleq \left\{\boldsymbol{\theta} \in \mathbb{R}^d \mid \exists\boldsymbol{\theta}' \in \mathcal{A} : \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\| < \epsilon\right\} ,$$

and the notion of set convergence

Definition 2 (Set convergence)**.** *A sequence of sets $\mathcal{A}_t \subset \mathbb{R}^d$ converges to another sequence of sets $\mathcal{A}'_t \subset \mathbb{R}^d$ if $\forall\epsilon > 0\ \exists t_0$ such that $\forall t > t_0\ \mathcal{A}_t \subset \mathcal{B}_\epsilon\left(\mathcal{A}'_t\right)$.*

3.1. Margin of Constrained Path Converges to Maximum Margin

For all ρ, the constrained path margin deviation from the max-margin is bounded, as we prove next.

Lemma 3. *For all ρ, and every $\boldsymbol{\theta}_c\left(\rho\right)$ in $\Theta_c\left(\rho\right)$*

$$\gamma^*\left(\rho\right) - \gamma\left(\rho, \boldsymbol{\theta}_c\left(\rho\right)\right) \leq \log N . \quad (8)$$

Proof. Note that $\forall\boldsymbol{\theta}$:

$$e^{-\gamma(\rho,\boldsymbol{\theta})} \leq \sum_{n=1}^N \exp\left(-f_n\left(\rho\boldsymbol{\theta}\right)\right) \leq Ne^{-\gamma(\rho,\boldsymbol{\theta})}. \quad (9)$$

Since, $\forall\boldsymbol{\theta} \in \mathbb{S}^{d-1}, \mathcal{L}\left(\rho\boldsymbol{\theta}_c\left(\rho\right)\right) \leq \mathcal{L}\left(\rho\boldsymbol{\theta}\right)$, we have, $\forall\boldsymbol{\theta}_c\left(\rho\right) \in \Theta_c\left(\rho\right)$ and $\forall\boldsymbol{\theta}_m\left(\rho\right) \in \Theta_m\left(\rho\right)$,

$$\begin{aligned}1 \leq \frac{\mathcal{L}\left(\rho\boldsymbol{\theta}_m\left(\rho\right)\right)}{\mathcal{L}\left(\rho\boldsymbol{\theta}_c\left(\rho\right)\right)} &= \frac{\sum_{n=1}^N \exp\left(-f_n\left(\rho\boldsymbol{\theta}_m\left(\rho\right)\right)\right)}{\sum_{n=1}^N \exp\left(-f_n\left(\rho\boldsymbol{\theta}_c\left(\rho\right)\right)\right)}\\ &\leq N\exp\left(-\left(\gamma^*\left(\rho\right) - \gamma\left(\rho, \boldsymbol{\theta}_c\left(\rho\right)\right)\right)\right) .\end{aligned}$$

$$\Rightarrow \gamma^*\left(\rho\right) - \gamma\left(\rho, \boldsymbol{\theta}_c\left(\rho\right)\right) \leq \log N . \qquad \square$$

Lemma 3 immediately implies that

Corollary 1. *If $\lim_{\rho\to\infty} \gamma^*\left(\rho\right) = \infty$, then for all ρ, and every $\boldsymbol{\theta}_c\left(\rho\right)$ in $\Theta_c\left(\rho\right)$*

$$\lim_{\rho\to\infty} \frac{\gamma^*\left(\rho\right)}{\gamma\left(\rho, \boldsymbol{\theta}_c\left(\rho\right)\right)} = 1 .$$

The last corollary states that the margin of the constrained path converges to the maximum margin. However, this does not necessarily imply convergence in parameter space, i.e., this result does not guaranty that $\Theta_c\left(\rho\right)$ converges to $\Theta_m\left(\rho\right)$. We analyze some positive and negative examples to demonstrate this claim.

Example 1: homogeneous models

It is straightforward to see that, for α-positive homogeneous prediction functions (Definition 1) the margin path $\Theta_m(\rho)$ in eq. 6 is the same set for any ρ, and is given by

$$\Theta^*_m = \arg\max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \min_n f_n\left(\boldsymbol{\theta}\right) .$$

Additionally, as we show next, for such models Lemma 3 implies convergence in parameter space, i.e., $\Theta_c\left(\rho\right)$ converges to $\Theta_m\left(\rho\right)$. To see this, notice that for α-positive homogeneous functions f_n, $\forall\boldsymbol{\theta}_c\left(\rho\right) \in \Theta_c\left(\rho\right)$:

$$\begin{aligned}&\gamma^*\left(\rho\right) - \gamma\left(\rho, \boldsymbol{\theta}_c\left(\rho\right)\right) =\\ &\max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \min_n f_n\left(\rho\boldsymbol{\theta}\right) - \min_n f_n\left(\rho\boldsymbol{\theta}_c\left(\rho\right)\right)\\ &\rho^\alpha\left(\max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \min_n f_n\left(\boldsymbol{\theta}\right) - \min_n f_n\left(\boldsymbol{\theta}_c\left(\rho\right)\right)\right)\\ &\leq \log N .\end{aligned}$$

For $\rho \to \infty$ we must have

$$\left(\max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \min_n f_n\left(\boldsymbol{\theta}\right) - \min_n f_n\left(\boldsymbol{\theta}_c\left(\rho\right)\right)\right) \to 0 .$$

By continuity, the last equation implies that $\Theta_c\left(\rho\right)$ converges to $\Theta_m\left(\rho\right)$. For full details see Appendix D.1.

Connection to previous results: For linear models, Rosset et al. (2004a) connected the L_1 constrained path and maximum L_1 margin solution. In addition, for any norm, Rosset et al. (2004b) showed that the regularization path converges to the limit of the margin path. In a recent work, Wei et al. (2018) extended this result to homogeneous models with cross-entropy loss. Here, for homogeneous models and any norm, we show a connection between the constrained path and the margin path.

Extension: Later, in Theorem 4 we prove a more refined result: the constrained path converges to a *specific subset* of the margin path set (the lexicographic max-margin set).

In contrast, in general models, 8 does not necessarily imply convergence in the parameter space. We demonstrate this result in the next example.

Example 2: log predictor: We denote $\mathbf{z}_n = y_n \mathbf{x}_n$ for some dataset $\{\mathbf{x}_n, y_n\}_{n=1}^N$, with features $\mathbf{x}_n$ and label y_n. We examine the prediction function $f_n(\rho, \boldsymbol{\theta}) = \log\left(\rho\boldsymbol{\theta}^\top \mathbf{z}_n\right)$ for $\boldsymbol{\theta}^\top \mathbf{z}_n > 0$. We focus on the loss function tail behaviour and thus only care about the loss function behaviour in $\boldsymbol{\theta}^\top \mathbf{z}_n > 0$ region. We assume that a separator which satisfy this constraint exists since we are focusing on realizable problems.

Since $\log(.)$ is strictly increasing and $\rho > 0$, we have

$$\gamma(\rho, \boldsymbol{\theta}) = \min_n f_n(\rho\boldsymbol{\theta}) = \log\left(\rho \min_n \boldsymbol{\theta}^\top \mathbf{z}_n\right) .$$

We denote $\widetilde{\gamma}(\boldsymbol{\theta}) = \min_n \boldsymbol{\theta}^\top \mathbf{z}_n$ and $\widetilde{\gamma}^* = \max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \widetilde{\gamma}(\boldsymbol{\theta})$. Note that $\gamma^*(\rho) = \log(\rho\widetilde{\gamma}^*)$. Now consider $\boldsymbol{\theta}_c(\rho) \in \Theta_c(\rho)$ such that for some ρ_0 and $\forall\rho > \rho_0$: $\widetilde{\gamma}(\boldsymbol{\theta}_c(\rho)) = \min_n \boldsymbol{\theta}_c(\rho)^\top \mathbf{z}_n \geq \frac{\widetilde{\gamma}^*}{N}$. For this case, we still have,

$$\begin{aligned}\gamma^*(\rho) - \gamma(\rho, \boldsymbol{\theta}_c(\rho)) &= \log(\rho\widetilde{\gamma}^*) - \log(\rho\widetilde{\gamma}(\boldsymbol{\theta}_c(\rho))) \\ &= \log\left(\frac{\widetilde{\gamma}^*}{\widetilde{\gamma}(\boldsymbol{\theta}_c(\rho))}\right) \leq \log N .\end{aligned}$$

but clearly, $\widetilde{\gamma}(\boldsymbol{\theta}_c(\rho)) \nrightarrow \widetilde{\gamma}^*$. Thus, Lemma 3 does not guarantee that $\widetilde{\gamma}(\boldsymbol{\theta}_c(\rho)) \to \widetilde{\gamma}^*$ as $\rho \to \infty$, or that $\Theta_c(\rho)$ converges to $\Theta_m(\rho)$.

Analogies with regularization and optimization paths: This example demonstrates that for the prediction function $\log(\rho\boldsymbol{\theta}^\top \mathbf{z})$ for $\boldsymbol{\theta}^\top \mathbf{z} > 0$, the constrained path does not necessarily converge to the margin path. This is equivalent to *setup A*: linear prediction models with loss function $\exp(-\log(u))$. Rosset et al. (2004b) and Nacson et al. (2019a) state related results for *setup A*. Both works derived conditions on the loss function that ensure convergence to the margin path from the regularization/ optimization path respectively. Rosset et al. (2004b) showed that in *setup A* the regularization path does not necessarily converge to the margin path. (Nacson et al., 2019a) showed a similar result for the optimization path, i.e., that in *setup A* the optimization path does not necessarily converge to the margin path. Both results align with our results for the constrained path.

In contrast, according to the conditions of Rosset et al. (2004b); Nacson et al. (2019a), we know that if the prediction function is $\log^{1+\epsilon}(\rho\boldsymbol{\theta}^\top \mathbf{z})$ for some $\epsilon > 0$ and $\boldsymbol{\theta}^\top \mathbf{z} > 0$, then the regularization path and optimization path *do converge* to the margin path. In the next example, we show that this is also true for the constrained path.

Example 3: $(1+\epsilon)$**-log predictor**: We examine the prediction function $f_n(\rho, \boldsymbol{\theta}) = \log^{1+\epsilon}\left(\rho\boldsymbol{\theta}^\top \mathbf{z}_n\right)$ for $\boldsymbol{\theta}^\top \mathbf{z}_n > 0$ and some $\epsilon > 0$. Since the log function is strictly increasing and $\epsilon, \rho > 0$, we have

$$\gamma(\rho, \boldsymbol{\theta}) = \min_n f_n(\rho\boldsymbol{\theta}) = \log^{1+\epsilon}\left(\rho \min_n \boldsymbol{\theta}^\top \mathbf{z}_n\right) .$$

For all $\boldsymbol{\theta}_c(\rho) \in \Theta_c(\rho)$:

$$\begin{aligned}&\gamma^*(\rho) - \gamma(\rho, \boldsymbol{\theta}_c(\rho)) \\ &= (1+\epsilon)\log^\epsilon(\rho)\left(\log(\widetilde{\gamma}^*) - \log(\widetilde{\gamma}(\boldsymbol{\theta}_c(\rho)))\right) \\ &\quad + o(\log^\epsilon(\rho)) \leq N .\end{aligned}$$

For $\rho \to \infty$ we must have $(\log(\widetilde{\gamma}^*) - \log(\widetilde{\gamma}(\boldsymbol{\theta}_c(\rho)))) \to 0$, which implies, by continuity, that $\Theta_c(\rho)$ converges to $\Theta_m(\rho)$. For details, see Appendix D.2.

3.2. Sum of Positively Homogeneous Functions

Remark: The results in this subsection are specific for the Euclidean or L_2 norm.

Let $f_n(\rho\boldsymbol{\theta})$ be functions that are a finite sum of positively homogeneous functions, i.e., for some finite K:

$$\forall n: \ f_n(\rho\boldsymbol{\theta}) = \sum_{k=1}^K f_n^{(k)}(\rho\boldsymbol{\theta}_k) \ , \tag{10}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K]$ and $f_n^{(k)}(\boldsymbol{\theta}_k)$ are α_k-positive homogeneous functions, where $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_K$.

First, we characterize the asymptotic form of the margin path in this setting.

Lemma 4. *Let $f_n(\boldsymbol{\theta})$ be a sum of positively homogeneous functions as in eq. 10. Then, the set of solutions of*

$$\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|^2 \ \text{ s.t. } \forall n: f_n(\rho\boldsymbol{\theta}) \geq \gamma^*(\rho) \ . \tag{11}$$

can be written as

$$\boldsymbol{\theta}_k^* = \frac{1}{\rho}\left(\boldsymbol{w}_k + o(1)\right)\left(\gamma^*(\rho)\right)^{\frac{1}{\alpha_k}} \tag{12}$$

where the $o(1)$ term is vanishing as $\gamma^*(\rho) \to \infty$, and

$$\boldsymbol{w}^* = [\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_K^*] \in \mathcal{W},$$

where

$$\mathcal{W} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{w}_1\|^2 \ \text{ s.t. } \forall n: f_n(\boldsymbol{w}) \geq 1. \tag{13}$$

Proof. We write the original optimization problem

$$\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{k=1}^K \|\boldsymbol{\theta}_k\|^2 \ \text{ s.t. } \forall n: \sum_{k=1}^K f_n^{(k)}(\rho\boldsymbol{\theta}_k) \geq \gamma^*(\rho) \ .$$

Dividing by $\gamma^*(\rho)$, using the α_k positive homogeneity of $f_n^{(k)}$, and changing the variables as $\boldsymbol{\theta}_k = \frac{1}{\rho}\boldsymbol{w}_k\left(\gamma^*(\rho)\right)^{\frac{1}{\alpha_k}}$, we obtain an equivalent optimization problem

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d}\sum_{k=1}^{K}\gamma^*(\rho)^{\frac{2}{\alpha_k}}\left\|\boldsymbol{w}_k\right\|^2 \text{ s.t. } \forall n: f_n(\boldsymbol{w})\geq 1\,. \tag{14}$$

We denote the set of solutions of eq. 14 as $\mathcal{W}(\gamma^*(\rho))$. Taking the limit of $\gamma^*(\rho)\to\infty$ of this optimization problem we find that any solution $\boldsymbol{w}\in\mathcal{W}(\gamma^*(\rho))$ must minimize the first term in the sum $\left\|\boldsymbol{w}_1\right\|^2$, and only then the other terms. Therefore the asymptotic solution is of the form of eqs. 12 and 13. We prove this reasoning formally in Appendix B, i.e., we show that

Claim 1. *The solution of eq. 14 is the same solution described in Lemma 4, i.e., eqs. 12 and 13.* □

The following Lemma will be used to connect the constrained path to the characterization of the margin path.

Lemma 5. *Let $f_n(\rho\boldsymbol{\theta})$ be a sum of positively homogeneous functions as in eq. 10. Any path $\boldsymbol{\theta}(\rho)$ such that*

$$\gamma^*(\rho)-\gamma(\rho,\boldsymbol{\theta}(\rho))<C\,. \tag{15}$$

is of the form described in eqs. 12 and 13.

Proof. See Appendix C. □

Combining Lemma 3, 4 and Lemma 5 we obtain the following Theorem

Theorem 1. *Under Assumption 1 and 2, any solution in* $\operatorname{argmin}_{\|\boldsymbol{\theta}\|\leq 1}\mathcal{L}(\rho\boldsymbol{\theta})$ *converges to*

$$\boldsymbol{\theta}_k^* = \frac{1}{\rho}\left(\boldsymbol{w}_k^*+o(1)\right)\left(\gamma^*(\rho)\right)^{\frac{1}{\alpha_k}} \tag{16}$$

where the $o(1)$ term is vanishing as $\gamma^*(\rho)\to\infty$, and

$$\boldsymbol{w}^* = [\boldsymbol{w}_1^*,\dots,\boldsymbol{w}_K^*]\in\mathcal{W},$$

where

$$\mathcal{W} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d}\left\|\boldsymbol{w}_1\right\|^2 \text{ s.t. } \forall n: f_n(\boldsymbol{w})\geq 1. \tag{17}$$

Theorem 1 implications: An important implication of Theorem 1 is that an ensemble on neural networks will aim to discard the shallowest network in the ensemble. Consider the following setting: for each $k\in\{1,\dots,K\}$, the function $\forall n: f_n^{(k)}(\rho\boldsymbol{\theta}_k)$ represents a prediction function of some feedforward neural network with no bias, all with the same positive-homogeneous activation function $\sigma(\cdot)$ of some degree α (*e.g.*, ReLU activation is positive-homogeneous of degree 1). Note that in this setup, each of the k prediction functions $f_n^{(k)}(\rho\boldsymbol{\theta}_k)$ is also a positive-homogeneous function. In particular, network k with depth d_k is positive homogeneous with degree $\alpha_k=\alpha d_k$ where α is the activation function degree. Since all the networks have the same activation function, deeper networks will have larger degree. We assume WLOG that $d_1<d_2<\cdots<d_K$. This implies that $\alpha_1<\alpha_2<\cdots<\alpha_K$. In this setting, $\forall n: f_n(\rho\boldsymbol{\theta})=\sum_{k=1}^{K}f_n^{(k)}(\rho\boldsymbol{\theta}_k)$ represents an ensemble of these networks. From Theorem 1, the solution of the constrained path will satisfy

$$\begin{aligned}
f_n(\rho\boldsymbol{\theta}^*) &= \sum_{k=1}^{K}f_n^{(k)}(\rho\boldsymbol{\theta}_k^*)\\
&= \sum_{k=1}^{K}f_n^{(k)}\left(\left(\boldsymbol{w}_k^*+o(1)\right)\left(\gamma^*(\rho)\right)^{\frac{1}{\alpha_k}}\right)\\
&= \gamma^*(\rho)\sum_{k=1}^{K}f_n^{(k)}\left(\boldsymbol{w}_k^*+o(1)\right)\,,
\end{aligned}$$

where $\boldsymbol{w}^*\in\mathcal{W}$ and $\mathcal{W}$ is calculated using eq. 13. Examining equation 13, we observe that the network aims to minimize the $\boldsymbol{w}_1$ norm. In particular, if the network ensemble can satisfy the constraints $\forall n: f_n(\boldsymbol{w})\geq 1$ with $\boldsymbol{w}_1=\mathbf{0}$, then the first equation obtained solutions will satisfy $\boldsymbol{w}_1=\mathbf{0}$. Thus the ensemble will discard the shallowest network if it is ”unnecessary” to satisfy the constraint.

Furthermore, from eq. 14 we conjecture that after discarding the shallowest “unnecessary” network, the ensemble will tend to minimize $\|\boldsymbol{w}_2\|$, i.e., to discard the second shallowest ”unnecessary” network. This will continue until there are no more ”unnecessary” shallow networks. In other words, we conjecture that the an ensemble of neural networks will aim to discard the shallowest “unnecessary” networks.

Additionally, using Theorem 1 we can now represent hard-margin SVM problems **with unregularized bias**. Previous results only focused on linear prediction functions without bias. Trying to extend these results to SVM with bias by extending all the input vectors $\mathbf{x}_n$ with an additional $'1'$ component would fail since the obtained solution in the original $\mathbf{x}$ space is the solution of

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d,b\in\mathbb{R}}\left\|\mathbf{w}\right\|^2+b^2 \text{ s.t. } y_n\left(\mathbf{w}^\top\mathbf{x}_n+b\right)\geq 1,$$

which is not the L_2 max-margin (SVM) solution, as pointed out by (Nar et al., 2019). However, we can now achieve this goal using Theorem 1. For some dataset $\{\mathbf{x}_n,y_n\}_{n=1}^{N}$, $\mathbf{x}_n\in\mathbb{R}^d$, $y_n\in\{-1,1\}$, we use the following prediction function $f_n(\boldsymbol{\theta})=y_n\left(\boldsymbol{\theta}_1^\top\mathbf{x}_n+b^2\right)$ where $\boldsymbol{\theta}=[\boldsymbol{\theta}_1,b]$. From eqs. 12, 13 the asymptotic solution will satisfy $\arg\min_{\boldsymbol{\theta}_1,b}\left\|\boldsymbol{\theta}_1\right\|^2$ s.t. $\forall n: y_n\left(\boldsymbol{\theta}_1^\top\mathbf{x}_n+b^2\right)\geq 1$.

4. Homogeneous Models

In the previous section we connected the constrained path to the margin path. We would like to refine this characterization and also understand the connection to the optimization path. In this section we are able to do so for prediction functions $f_n(\boldsymbol{\theta})$ which are α-positive homogeneous functions (definition 1).

In the homogeneous case, eq. 7 is equivalent, $\forall\rho$, to

$$\Theta_m^* = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|^2 \text{ s.t. } \min_n f_n(\boldsymbol{\theta}) \geq \gamma^*(1) \qquad (18)$$

since f_n is homogeneous.

4.1. Optimization Path Converges to Stationary Points of the Margin Path and Constrained Path

Remark: The results in this subsection are specific for the Euclidean or L_2 norm, as opposed to many of the results in this paper which are stated for any norm.

In this section, we link the optimization path to the margin path and the constrained path. These results require the following smoothness assumption:

Assumption 3 (Smoothness). *We assume $f_n(\cdot)$ is a $\mathcal{C}^2$ function.*

Relating optimization path and margin path. The limit of the margin path for homogeneous models is given by eq. 18. In this section we first relate the optimization path to this limit of margin path.

Note that for general homogeneous prediction functions f_n, eq. 18 is a non-convex optimization problem, and thus it is unlikely for an optimization algorithm such as gradient descent to find the global optimum. We can relax the set to $\boldsymbol{\theta}$ that are first-order stationary, i.e., critical points of 18. For $\boldsymbol{\theta}\in\mathbb{S}^{d-1}$, denote the set of support vectors of $\boldsymbol{\theta}$ as

$$S_m(\boldsymbol{\theta}) = \{n : f_n(\boldsymbol{\theta}) = \gamma^*(1)\}. \qquad (19)$$

Definition 3 (First-order Stationary Point). *The first-order optimality conditions of 18 are:*

1. *$\forall n$, $f_n(\boldsymbol{\theta}) \geq \gamma^*(1)$*

2. *There exists $\boldsymbol{\lambda}\in\mathbb{R}_+^N$ such that $\boldsymbol{\theta} = \sum_n \lambda_n \nabla f_n(\boldsymbol{\theta})$ and $\lambda_n = 0$ for $n \notin S_m(\boldsymbol{\theta})$.*

We denote by Θ_m^s the set of first-order stationary points.

Let $\boldsymbol{\theta}(t)$ be the iterates of gradient descent. Define $\ell_n(t) = \exp(-f_n(\boldsymbol{\theta}(t)))$ and $\boldsymbol{\ell}(t)$ be the vector with entries $\ell_n(t)$. The following two assumptions assume that the limiting direction $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|}$ exist and the limiting direction of the losses $\frac{\boldsymbol{\ell}(t)}{\|\boldsymbol{\ell}(t)\|_1}$ exist. Such assumptions are natural in the context of max-margin problems, since we want to argue that $\boldsymbol{\theta}(t)$ converges to a max-margin direction, and also the losses $\boldsymbol{\ell}(t)/\|\boldsymbol{\ell}(t)\|_1$ converges to an indicator vector of the support vectors. The first step to argue this convergence is to ensure the limits exist.

Assumption 4 (Asymptotic Formulas). *Assume that $\mathcal{L}(\boldsymbol{\theta}(t)) \to 0$, that is we converge to a global minimizer. Further assume that $\lim_{t\to\infty}\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ and $\lim_{t\to\infty}\frac{\boldsymbol{\ell}(t)}{\|\boldsymbol{\ell}(t)\|_1}$ exist. Equivalently,*

$$\ell_n(t) = h(t)a_n + h(t)\epsilon_n(t) \qquad (20)$$

$$\boldsymbol{\theta}(t) = g(t)\bar{\boldsymbol{\theta}} + g(t)\boldsymbol{\delta}(t), \qquad (21)$$

with $\|\boldsymbol{a}\|_1 = 1$, $\|\bar{\boldsymbol{\theta}}\|_2 = 1$, $\lim_{t\to\infty} h(t) = 0$, $\lim_{t\to\infty}\epsilon_n(t) = 0$, and $\lim_{t\to\infty}\boldsymbol{\delta}(t) = 0$.

Assumption 5 (Linear Independence Constraint Qualification). *Let $\boldsymbol{\theta}\in\mathbb{S}^{d-1}$ be a unit vector. LICQ holds at $\boldsymbol{\theta}$ if the vectors $\{\nabla f_n(\boldsymbol{\theta})\}_{n\in S_m(\boldsymbol{\theta})}$ are linearly independent.*

Remark 1. *Constraint qualifications allow the first-order optimality conditions of Definition 3 to be a necessary condition for optimality. Without constraint qualifications, the global optimum need not satisfy the optimality conditions.*

LICQ is the simplest among many constraint qualification conditions identified in the optimization literature (Nocedal & Wright, 2006).

For example, in linear SVM, LICQ is ensured if the set of support vectors is linearly independent. Consider $f_n(\boldsymbol{\theta}) = \mathbf{x}_n^\top\boldsymbol{\theta}$ and $\mathbf{x}_n$ be the support vectors. Then $\nabla f_n(\bar{\boldsymbol{\theta}}) = \mathbf{x}_n$, and so linear independence of the support vectors implies LICQ. For data sampled from an absolutely continuous distribution, the SVM solution will always have linearly independent support vectors (Soudry et al., 2018b, Lemma 12), but LICQ may fail when the data is degenerate.

Theorem 2. *Define $\bar{\boldsymbol{\theta}} = \lim_{t\to\infty}\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$. Under Assumptions 3, 4, and constraint qualification at $\bar{\boldsymbol{\theta}}$ (Assumption 5), $\bar{\boldsymbol{\theta}}$ is a first-order stationary point of 18.*

The proof of Theorem 2 can be found in Appendix E.1.

Optimization path and constrained path. Next, we study how the optimization path as $t\to\infty$ converges to stationary points of the constrained path with $\rho\to\infty$.

The first-order optimally conditions of the constrained path $\min_{\|\boldsymbol{\theta}\|\leq 1}\mathcal{L}(\rho\boldsymbol{\theta})$, require that the constraints hold, and the gradient of the Lagrangian of the constrained path

$$\rho\nabla_{\boldsymbol{\theta}}\mathcal{L}(\rho\boldsymbol{\theta}) + \lambda(\rho)\boldsymbol{\theta} \qquad (22)$$

is equal to zero. In other words,

Remark 2. *Under Assumption 1, $\boldsymbol{\theta}$ is first-order optimal for the problem $\min_{\|\boldsymbol{\theta}\|\leq 1}\mathcal{L}(\rho\boldsymbol{\theta})$ if it satisfies:*

- $-\frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\rho\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\rho\boldsymbol{\theta})\|} = \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}$,
- $\|\boldsymbol{\theta}\| = 1$.

On many paths the gradient of the Lagrangian goes to zero as $\rho \to \infty$. However, we have a faster vanishing rate for the specific optimization paths that follow Definition 4 below. Therefore, these paths better approximate true stationary points:

Definition 4 (First-order optimal for $\rho \to \infty$)**.** *A sequence $\widetilde{\boldsymbol{\theta}}(t)$ is first-order optimal for $\min_{\|\boldsymbol{\theta}\|\leq 1}\mathcal{L}(\rho\boldsymbol{\theta})$ with $\rho \to \infty$ if*

- $\lim\limits_{t\to\infty} -\frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\rho\widetilde{\boldsymbol{\theta}}(t))}{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\rho\widetilde{\boldsymbol{\theta}}(t)\|} = \lim\limits_{t\to\infty}\frac{\widetilde{\boldsymbol{\theta}}(t)}{\|\widetilde{\boldsymbol{\theta}}(t)\|}$,
- $\|\widetilde{\boldsymbol{\theta}}(t)\| = 1$.

To relate the limit points of gradient decent to the constrained path, we will focus on stationary points of the constrained path that minimize the loss.

Theorem 3. *Let $\bar{\boldsymbol{\theta}} = \lim\limits_{t\to\infty}\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|}$ be the limit direction of gradient descent. Under Assumptions 1, 3, 4, and constraint qualification at $\bar{\boldsymbol{\theta}}$ (Assumption 5), the sequence $\boldsymbol{\theta}(t)/\|\boldsymbol{\theta}(t)\|$ is a first-order optimal point for $\rho \to \infty$ (Definition 4).*

The proof of Theorem 3 can be found in Appendix E.2.

4.2. Lexicographic Max-Margin

Recall that for positive homogeneous prediction functions, the margin path $\Theta_m(\rho)$ in eq. 11 is the same set for any ρ and is given by

$$\Theta_m^* = \arg\max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}}\min_n f_n(\boldsymbol{\theta}) \ .$$

For non-convex functions f_n or non-Euclidean norms $\|.\|$, the above set need not be unique. In this case, we define the following refined set of maximum margin solution set

Definition 5 (Lexicographic maximum margin set)**.** *The lexicographic margin set denoted by $\Theta^*_{m,N}$ is given by the following iterative definition of $\Theta^*_{m,k}$ for $k = 1, 2, \dots, N$:*

$$\Theta^*_{m,0} = \mathbb{S}^{d-1}\,,$$

$$\Theta^*_{m,k} = \arg\max_{\boldsymbol{\theta}\in\Theta^*_{m,k-1}}\left(\min_{\{n_\ell\}_{\ell=1}^k}\max_{\ell\in[k]} f_{n_\ell}(\boldsymbol{\theta})\right) \subseteq \Theta^*_{m,k-1}\,.$$

In the above definition, $\Theta^*_{m,1} = \Theta^*_m$ denotes the set of maximum margin solutions, $\Theta^*_{m,1}$ denotes the subset of $\Theta^*_{m,1}$ with second smallest margin, and so on.

For an alternate representation of $\Theta^*_{m,k}$, we introduce the following notation: for $\boldsymbol{\theta}\in\mathbb{S}^{d-1}$, let $n^*_\ell(\boldsymbol{\theta})\in[N]$ denote the index corresponding to the ℓ^{th} smallest margin of $\boldsymbol{\theta}$ as defined below by breaking ties in the $\arg\min$ arbitrarily:

$$\begin{aligned} n^*_1(\boldsymbol{\theta}) &= \arg\min_n f_n(\boldsymbol{\theta}) \\ n^*_k(\boldsymbol{\theta}) &= \arg\min_{n\notin\{n^*_\ell(\boldsymbol{\theta})\}_{l=1}^{k-1}} f_n(\boldsymbol{\theta}) \quad \text{for } k\geq 2. \end{aligned} \tag{23}$$

Using this notation, we can rewrite $\Theta^*_{m,k+1}$ as

$$\Theta^*_{m,k+1} = \arg\max_{\boldsymbol{\theta}\in\Theta^*_{m,k}} f_{n^*_{k+1}(\boldsymbol{\theta})}(\boldsymbol{\theta})\ .$$

We also define the limit set of constrained path as follows:

Definition 6 (Limit set of constrained path)**.** *The limit set of constrained path is defined as follows:*

$$\Theta_c^\infty = \left\{\boldsymbol{\theta} : \begin{matrix}\exists\{\rho_i, \boldsymbol{\theta}_{\rho_i}\}_{i=1}^\infty \textit{ with } \rho_i\to\infty, \boldsymbol{\theta}_{\rho_i}\in\Theta_c(\rho_i) \\ \textit{such that } \boldsymbol{\theta}_{\rho_i}\to\boldsymbol{\theta}\end{matrix}\right\}.$$

Theorem 4. *For α-positive homogeneous prediction functions the limit set of constrained path is contained in the lexicographic maximum margin set, i.e., $\Theta_c^\infty \subseteq \Theta^*_{m,N}$.*

The proof of the above Theorem follows from adapting the arguments of (Rosset et al., 2004a) (Theorem 7 in Appendix B.2) for general homogeneous models. We show the complete proof in Appendix E.3.

5. Summary

In this paper we characterized the connections between the constrained, margin and optimization paths. First, in Section 3, we examined general non-homogeneous models. We showed that the margin of the constrained path solution converges to the maximum margin. We further analyzed this result and demonstrated how it implies convergence in parameters, i.e., $\Theta_c(\rho)$ converges to $\Theta_m(\rho)$, for some models. Then, we examined functions that are a finite sum of positively homogeneous functions. These prediction function can represent an ensemble of neural networks with positive homogeneous activation functions. For this model, we characterized the asymptotic constrained path and margin path solution. This implies a surprising result: ensembles of neural networks will aim to discard the most shallow network. In the future work we aim to analyze sum of homogeneous functions with shared variables, such as ResNets.

Second, in Section 4 we focus on homogeneous models. For such models we link the optimization path to the margin and constrained paths. Particularly, we show that the optimization path converges to stationary points of the constrained path and margin path. In future work, we aim to extend this to non-homogeneous models. In addition, we give a more refined characterization of the constrained path limit. It will be interesting to find whether this characterization be further refined to answer whether the weighting of the data point can have any effect on the selection of the asymptotic solution — as (Byrd & Lipton, 2018) observed empirically that it did not.

Acknowledgements

The authors are grateful to C. Zeno, and N. Merlis for helpful comments on the manuscript. This research was supported by the Israel Science foundation (grant No. 31/1031), and by the Taub foundation. SG and NS were partially supported by NSF awards IIS-1302662 and IIS-1764032.

References

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. *arXiv preprint arXiv:1810.10082*, 2018.

Byrd, J. and Lipton, Z. C. Weighted risk minimization & deep learning. *arXiv preprint arXiv:1812.03372*, 2018.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6152–6160, 2017.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *NIPS*, 2018a.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. *ICML*, 2018b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *NIPS*, 2017.

Ji, Z. and Telgarsky, M. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ICLR*, pp. 1–16, 2017.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H., Srebro, N., and Soudry, D. Convergence of gradient descent on separable data. *AISTATS*, 2019a.

Nacson, M. S., Srebro, N., and Soudry, D. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *AISTATS*, 2019b.

Nar, K., Ocal, O., Sastry, S. S., and Ramchandran, K. Cross-entropy loss leads to poor margins, 2019.

Neyshabur, B., Salakhutdinov, R. R., and Srebro, N. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2015a.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*, 2015b.

Neyshabur, B., Tomioka, R., Salakhutdinov, R., and Srebro, N. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Nocedal, J. and Wright, S. Numerical optimization. *Springer Science*, 35(67-68), 2006.

Rosset, S., Zhu, J., and Hastie, T. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 2004a.

Rosset, S., Zhu, J., and Hastie, T. J. Margin maximizing loss functions. In *Advances in neural information processing systems*, pp. 1237–1244, 2004b.

Soudry, D., Hoffer, E., Shpigel Nacson, M., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *JMLR*, 2018a.

Soudry, D., Hoffer, E., and Srebro, N. The implicit bias of gradient descent on separable data. *ICLR*, 2018b.

Suggala, A., Prasad, A., and Ravikumar, P. K. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pp. 10631–10641, 2018.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369v1*, pp. 1–34, 2018.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. *arXiv preprint arXiv:1705.08292*, 2017.

Wu, L., Zhu, Z., and E, W. Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes. *arXiv*, 2017.

Xu, T., Zhou, Y., Ji, K., and Liang, Y. When will gradient methods converge to max-margin classifier under reLU models?, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Appendix

A. Proof of Lemma 1

Proof. Let $\mathbf{w}^*(\rho)$ be a solution of the optimization problem in eq. 2. Then, $g(\mathbf{w}^*(\rho)) = \rho$, since otherwise we could have decreased ρ without changing $\mathbf{w}^*(\rho)$ or $\phi(\rho)$ — and this is impossible, since $\phi(\rho)$ is strictly monotonically decreasing. Therefore, we cannot decrease $g(\mathbf{w})$ below ρ without increasing $f(\mathbf{w})$ above $\phi(\rho)$. This implies that $\mathbf{w}^*(\rho)$ is a solution of the optimization problem in eq. 3 with $\phi(\rho)$. Next, all that is left to show that eq. 3 has no additional solutions. Suppose by contradiction there were such solutions $\mathbf{w}'(\rho)$. Since they are also minimizers of eq. 3, like $\mathbf{w}^*(\rho)$, they have the same minimum value $g(\mathbf{w}'(\rho)) = \rho$. Since they are not solutions of eq. 2, we have $f(\mathbf{w}) > \phi(\rho)$. However, this means they are not feasible for eq. 3, and therefore cannot be solutions. □

B. Proof of Claim 1

Proof. Recall we denoted the set of solutions of eq. 14 as $\mathcal{W}(\gamma^*(\rho))$, and recall $\mathcal{W}$ from eq. 13. To simplify notations we omit the dependency on ρ from the notation, i.e., we replace $\gamma^*(\rho)$ with γ. Suppose the claim was not correct. Then, there would have existed $\epsilon > 0$ such that $\forall\gamma$, $\exists\gamma' > \gamma$ such that $\exists\boldsymbol{w}^*(\gamma') \in \mathcal{W}(\gamma') \setminus \mathcal{B}_\epsilon(\mathcal{W})$. Note that $\boldsymbol{w}^*(\gamma') \in \mathcal{W}(\gamma')$ is feasible in both optimization problems (eq. 13 and 14), since both problems have the same constraints. Moreover, since $\boldsymbol{w}^*(\gamma') \notin \mathcal{B}_\epsilon(\mathcal{W})$ it must be sub-optimal in comparison to the solution of eq. 13. Therefore, $\exists\epsilon' > 0$ such that for any γ', $\|\boldsymbol{w}_1^*(\gamma')\|^2 > \min_{\boldsymbol{w}\in\mathcal{W}} \|\boldsymbol{w}_1\|^2 + \epsilon'$. Then we can write (from eq. 14)

$$\mathcal{W}(\gamma') = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\|\boldsymbol{w}_1\|^2 + \sum_{k=2}^{K} (\gamma')^{\frac{2}{\alpha_k} - \frac{2}{\alpha_1}} \|\boldsymbol{w}_k\|^2 \right] \text{ s.t. } \forall n : f_n(\boldsymbol{w}) \geq 1 \,. \tag{24}$$

From Assumption 2 we know that $\exists c > 0$ such that $\forall\gamma > c$ a solution of the margin path exists. Therefore, $\forall\gamma \geq c$, eq. 11 is feasible. We assume, WLOG, that $c < \gamma'$. This implies that there exist a feasible finite solution $\widetilde{\boldsymbol{w}}$ to eq. 24 which does not depend on γ'. Therefore, $\forall\gamma'$, $\forall\boldsymbol{w} \in \mathcal{W}(\gamma')$, and $\forall k \in [K]$ the values of $\|\boldsymbol{w}_k\|^2$ are respectively bounded below the values of $\|\widetilde{\boldsymbol{w}}_k\|^2$, which are independent of γ'. This implies that if we select γ' large enough, we will have $\sum_{k=2}^{K} (\gamma')^{\frac{2}{\alpha_k} - \frac{2}{\alpha_1}} \|\boldsymbol{w}_k\|^2 < \epsilon'$. This would contradict the assumption that $\boldsymbol{w}^*(\gamma') \in \mathcal{W}(\gamma')$ and therefore minimizes eq. 24. This implies that $\forall\epsilon$, $\exists\gamma_0$ such that $\forall\gamma > \gamma_0$, we have $\mathcal{W}(\gamma) \subset \mathcal{B}_\epsilon(\mathcal{W})$, which entails the Theorem.

□

C. Proof of Lemma 5

Proof. We assume by contradiction that yet $\boldsymbol{\theta}(\rho)$ does not have the form of eqs. 12 and 13. Without loss of generality we can write

$$\rho\boldsymbol{\theta}(\rho) = \mathbf{v}_k(\rho) \left[\gamma(\rho, \boldsymbol{\theta}(\rho))\right]^{\frac{1}{\alpha_k}} \,. \tag{25}$$

If $\mathbf{v}_k(\rho') = \boldsymbol{w}_k^* + o(1)$, for some $\boldsymbol{w}^* = [\boldsymbol{w}_1^*, \dots, \boldsymbol{w}_K^*] \in \mathcal{W}$. Then we could have written, from eqs. 25 and 15

$$\rho\boldsymbol{\theta}(\rho) = (\boldsymbol{w}_k^* + o(1)) \left[\gamma(\rho, \boldsymbol{\theta}(\rho))\right]^{\frac{1}{\alpha_k}} = (\boldsymbol{w}_k^* + o(1)) \left[\gamma^*(\rho)\right]^{\frac{1}{\alpha_k}}$$

which contradicts out assumption that $\rho\boldsymbol{\theta}(\rho)$ does not have the form of eq. 13 and 14.

Therefore $\exists\delta > 0$, such that $\forall\rho$, $\exists\rho' > \rho$: $\mathbf{v}(\rho') \notin \mathcal{B}_\delta(\mathcal{W})$. The norm of the solution in eq. 25

$$\rho'^2 = \sum_{k=1}^{K} \|\mathbf{v}_k(\rho')\|^2 \left[\gamma(\rho, \boldsymbol{\theta}(\rho))\right]^{\frac{2}{\alpha_k}} \,,$$

is equal to the norm of the solution with margin $\gamma^*(\rho')$

$$\rho'^2 = \sum_{k=1}^{K} \|\boldsymbol{w}_k^* + o(1)\|^2 \left[\gamma^*(\rho')\right]^{\frac{2}{\alpha_k}} \,.$$

Therefore, from eq. 15 we have

$$\sum_{k=1}^{K}\left\|\boldsymbol{w}_1^* + o\left(1\right)\right\|^2\left[\gamma\left(\rho,\boldsymbol{\theta}\left(\rho\right)\right)+C\right]^{\frac{2}{\alpha_k}} > \sum_{k=1}^{K}\left\|\mathbf{v}_k\left(\rho'\right)\right\|^2\left[\gamma\left(\rho,\boldsymbol{\theta}\left(\rho\right)\right)\right]^{\frac{2}{\alpha_k}}$$

and so, dividing by $\left[\gamma\left(\rho,\boldsymbol{\theta}\left(\rho\right)\right)\right]^{\frac{2}{\alpha_k}}$ we obtain

$$\left\|\boldsymbol{w}_1^*\right\|^2 + o\left(1\right) > \left\|\mathbf{v}_1\left(\rho'\right)\right\|^2 + o\left(1\right) \tag{26}$$

However, since $\mathbf{v}\left(\rho'\right) \notin \mathcal{B}_\delta\left(\mathcal{W}\right)$, $\exists\epsilon' > 0$ such that for all ρ': $\left\|\mathbf{v}_1\left(\rho'\right)\right\|^2 > \left\|\boldsymbol{w}_1^*\right\|^2 + \epsilon'$ plugging this into eq. 26 we obtain

$$o\left(1\right) > \epsilon' + o\left(1\right)$$

which is a contradiction. Therefore, $\mathbf{v}\left(\rho'\right)$ converges into $\mathcal{W}$, and eq. 25 can be written in the form of eqs. 12 and 13. □

D. Examples Section: Auxiliary Results

D.1. Showing that margin convergence implies convergence in the parameter space for homogeneous models

We need to show that $\max\limits_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}}\min\limits_{n} f_n\left(\boldsymbol{\theta}\right) - \min\limits_{n} f_n\left(\boldsymbol{\theta}_c\left(\rho\right)\right) \to 0$ implies that $\Theta_c\left(\rho\right)$ converges to $\Theta_m\left(\rho\right)$.

We denote $g\left(\boldsymbol{\theta}\right) = \min\limits_{n} f_n\left(\boldsymbol{\theta}\right)$. This is a continues function since $\forall n : f_n$ is continues. In addition, we define for some $\rho_0 > 0$

$$A_r = \left\{\boldsymbol{\theta} \in \left\{\Theta_c\left(\rho\right)\right\}_{\rho\geq\rho_o} : \left|g\left(\boldsymbol{\theta}\right) - g\left(\boldsymbol{\theta}_m\right)\right| \leq r\right\}$$

where $\boldsymbol{\theta}_m \in \Theta_m$. Using this definition we also define $d\left(\boldsymbol{\theta}\right)$ as the Euclidean distance between $\boldsymbol{\theta}$ and any point in the set Θ_m and $d\left(r\right)$ as the maximal distance for $\boldsymbol{\theta} \in A_r$:

$$d\left(\boldsymbol{\theta}\right) = \min_{\boldsymbol{y}\in\Theta_m}\left\|\boldsymbol{y} - \boldsymbol{\theta}\right\| ,$$
$$d\left(r\right) = \max_{\boldsymbol{\theta}\in A_r} d\left(\boldsymbol{\theta}\right) .$$

Note that the maximum in the last equation is obtained as the maximum of a continues function over a compact set.
We want to show that $\Theta_c\left(\rho\right)$ converges to Θ_m. From definition, this implies that $\forall\epsilon > 0\ \exists\rho_0$ such that $\forall\rho > \rho_0\ \Theta_c\left(\rho\right) \subset \mathcal{B}_\epsilon\left(\Theta_m\right)$, i.e., $\forall\boldsymbol{\theta}_c(\rho) \in \Theta_c\left(\rho\right)$: $\exists\boldsymbol{\theta}' \in \Theta_m : \left\|\boldsymbol{\theta}_c(\rho) - \boldsymbol{\theta}'\right\| < \epsilon$.
Assume in contradiction that this is not the case. This means that $\exists\epsilon > 0$ such that $\forall\rho_0 : \exists\rho > \rho_0$ and $\exists\boldsymbol{\theta}_c(\rho) \in \Theta_c\left(\rho\right)$ so that $\forall\boldsymbol{\theta}' \in \Theta_m : \left\|\boldsymbol{\theta}_c(\rho) - \boldsymbol{\theta}'\right\| > \epsilon$.
This implies that $\lim_{r\to 0} d\left(r\right) \neq 0$. Using the limit definition we get that $\exists\epsilon > 0$ so that $\forall\delta > 0$, $\exists\left|r\right| < \delta$ and $d\left(r\right) > \epsilon$. Using our notations this implies that

$$\exists\epsilon > 0 : \forall\delta : \exists\boldsymbol{\theta}' \in A_r \text{ s.t. } \left|g\left(\boldsymbol{\theta}'\right) - g\left(\boldsymbol{\theta}_m\right)\right| \leq r < \delta \text{ and } d\left(\boldsymbol{\theta}'\right) > \epsilon .$$

Next, we build a subsequence $\left\{\boldsymbol{\theta}_i\right\}_{i=1}^{\infty}$ by taking a decreasing series of $\left\{\delta_i\right\}_{i=1}^{\infty}$ and their associated $\boldsymbol{\theta}'$ from the last equation. Since $\boldsymbol{\theta}_i'$ are bounded, there exist a convergent subsequence $\left\{\widetilde{\boldsymbol{\theta}}_i\right\}_{i=1}^{\infty}$. For this subsequence, we obtain, using g continuity

$$\lim_{i\to\infty} g\left(\widetilde{\boldsymbol{\theta}}_i\right) = g\left(\lim_{i\to\infty}\widetilde{\boldsymbol{\theta}}_i\right) = g\left(\boldsymbol{\theta}_m\right)$$

which implies that $\exists\boldsymbol{\theta}_m^* \in \Theta_m$ so that $\lim_{i\to\infty}\widetilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_m^*$ which contradicts the fact that $d\left(\widetilde{\boldsymbol{\theta}}_i\right) > \epsilon > 0$. □

D.2. Auxiliary results for $f_n\left(\rho, \boldsymbol{\theta}\right) = \log^{1+\epsilon}\left(\rho \boldsymbol{\theta}^\top \mathbf{z}_n\right)$

First, we show the full derivation of $\gamma^*\left(\rho\right) - \gamma\left(\rho, \boldsymbol{\theta}_c\left(\rho\right)\right)$.

$$\begin{aligned}
&\gamma^*\left(\rho\right) - \gamma\left(\rho, \boldsymbol{\theta}_c\left(\rho\right)\right) \\
&= \log^{1+\epsilon}\left(\rho\widetilde{\gamma}^*\right) - \log^{1+\epsilon}\left(\rho\widetilde{\gamma}\left(\boldsymbol{\theta}_c\left(\rho\right)\right)\right) \\
&= \left(\log\left(\rho\right) + \log\left(\widetilde{\gamma}^*\right)\right)^{1+\epsilon} - \left(\log\left(\rho\right) + \log\left(\widetilde{\gamma}\left(\boldsymbol{\theta}_c\left(\rho\right)\right)\right)\right)^{1+\epsilon} \\
&= \log^{1+\epsilon}\left(\rho\right) + \left(1+\epsilon\right)\log^{\epsilon}\left(\rho\right)\log\left(\widetilde{\gamma}^*\right) \\
&- \log^{1+\epsilon}\left(\rho\right) - \left(1+\epsilon\right)\log^{\epsilon}\left(\rho\right)\log\left(\widetilde{\gamma}\left(\boldsymbol{\theta}_c\left(\rho\right)\right)\right) + o\left(\log^{\epsilon}\left(\rho\right)\right) \\
&= \left(1+\epsilon\right)\log^{\epsilon}\left(\rho\right)\left(\log\left(\widetilde{\gamma}^*\right) - \log\left(\widetilde{\gamma}\left(\boldsymbol{\theta}_c\left(\rho\right)\right)\right)\right) + o\left(\log^{\epsilon}\left(\rho\right)\right) \\
&\leq N\,.
\end{aligned}$$

Second, we need to show that $\left(\log\left(\widetilde{\gamma}^*\right) - \log\left(\widetilde{\gamma}\left(\boldsymbol{\theta}_c\left(\rho\right)\right)\right)\right) \to 0$ implies that $\Theta_c\left(\rho\right)$ converges to $\Theta_m\left(\rho\right)$.

We denote $g\left(\boldsymbol{\theta}\right) = \log\left(\min_n \boldsymbol{\theta}^\top \mathbf{x}_n\right)$. The rest of the proof is identical to the proof for the homogeneous case in Appendix D.1.

E. Proofs in Section 4

E.1. Proof of Theorem 2

Define $S = \{n : f_n(\bar{\boldsymbol{\theta}}) = \gamma^*(1)\}$, where $\gamma^*(1)$ is the optimal margin attainable by a unit norm $\boldsymbol{\theta}$.

Lemma 6. *Under the setting of Theorem 2,*

$$\nabla f_n(\boldsymbol{\theta}(t)) = \nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + O(Bg(t)^{\alpha-1}\|\boldsymbol{\delta}(t)\|). \tag{27}$$

For $n \in S$, the second term is asymptotically negligible as a function of t,

$$\nabla f_n(\boldsymbol{\theta}(t)) = \nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + o(\nabla f_n(g(t)\bar{\boldsymbol{\theta}}))$$

Proof. By Taylor's theorem,

$$\nabla f_n(\boldsymbol{\theta}(t)) = \nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + \int_{s=0}^{s=1} \nabla^2 f_n(g(t)\bar{\boldsymbol{\theta}} + sg(t)\boldsymbol{\delta}(t))g(t)\boldsymbol{\delta}(t)\,\mathrm{d}s\,.$$

Let $\bar{\boldsymbol{\theta}}_s(t) := g(t)\bar{\boldsymbol{\theta}} + sg(t)\boldsymbol{\delta}(t)$. We bound the integrand in the second term.

$$\|\nabla^2 f_n(\bar{\boldsymbol{\theta}}_s(t))\| = \nabla^2 f_n\Big(\frac{\bar{\boldsymbol{\theta}}_s(t)}{\|\bar{\boldsymbol{\theta}}_s\|}\Big)\|\bar{\boldsymbol{\theta}}_s(t)\|^{\alpha-2} \leq B\|\bar{\boldsymbol{\theta}}_s(t)\|^{\alpha-2},$$

where $B = \max_{\|\boldsymbol{\theta}\|\leq 1}\|\nabla^2 f_n(\boldsymbol{\theta})\| < \infty$ since $\nabla^2 f_n$ is a continuous function maximized over a compact set.

Thus

$$\begin{aligned}
\nabla f_n(\boldsymbol{\theta}(t)) &= \nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + O(B\|\bar{\theta}_s(t)\|^{\alpha-1}\|\boldsymbol{\delta}(t)\|) \\
&= \nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + O(Bg(t)^{\alpha}(1+o(1))^{\alpha-1}\|\boldsymbol{\delta}(t)\|) \\
&= \nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + O(Bg(t)^{\alpha-1}\|\boldsymbol{\delta}(t)\|) && (\alpha \text{ is a constant independent of } t.)
\end{aligned}$$

$\nabla f_n(g(t)\bar{\boldsymbol{\theta}}) = g(t)^{\alpha-1}\nabla f_n(\bar{\boldsymbol{\theta}})$, and for $n \in S$, $\|\nabla f_n(\bar{\boldsymbol{\theta}})\| > 0$ via constraint qualification (Assumption 5). Thus for $n \in S$ and using $\|\boldsymbol{\delta}(t)\| = o(1)$,

$$\nabla f_n(\boldsymbol{\theta}(t)) = \nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + o(\nabla f_n(g(t)\bar{\boldsymbol{\theta}})).$$

□

Lemma 7. *Let $S = \{n : f_n(\bar{\boldsymbol{\theta}}) = \gamma^*(1)\}$. Under the conditions of Theorem 2, $a_n = 0$ for $n \notin S$.*

Proof.

$$\ell_n(t) = \exp(-f_n(\boldsymbol{\theta}(t))) = \exp(-g(t)^\alpha f_n(\bar{\boldsymbol{\theta}})) \exp(-g(t)^\alpha \nabla f_n(\bar{\boldsymbol{\theta}})^T \delta(t)) \exp(-g(t)^\alpha o(\delta(t))).$$

On the other hand, $\ell_n(t) = h(t)a_n + h(t)\epsilon_n(t)$, so $\frac{\ell_n(t)}{\|\ell(t)\|_1} \to a_n$.

Consider $n \notin S$ so $f_n(\bar{\boldsymbol{\theta}}) = \gamma_n > \gamma^*(1)$.

$$\begin{aligned} \frac{\ell_n(t)}{\|\ell(t)\|_1} &\leq \frac{\exp(-g(t)^\alpha(\gamma_n - \epsilon))}{\exp(-g(t)^\alpha(\gamma^*(1) + \epsilon)} \\ &\to 0 && \text{(since } \gamma^*(1) + \epsilon < \gamma_n - \epsilon \text{ for } \epsilon \text{ appropriately small)} \end{aligned}$$

Thus $a_n > 0$ only if $n \in S$.

□

Theorem 5 (Theorem 2)**.** *$\bar{\boldsymbol{\theta}}$ satifies the first-order optimality of margin problem.*

Proof. From the gradient dynamics,

$$\begin{aligned} \dot{\boldsymbol{\theta}}(t) &= \sum_n \exp(-f_n(\boldsymbol{\theta}(t)))\nabla f_n(\boldsymbol{\theta}(t)) \\ &= \sum_n (h(t)a_n + h(t)\epsilon_n(t))(\nabla f_n(g(t)\bar{\boldsymbol{\theta}}) + \Delta_n(t)), \end{aligned}$$

where $\Delta_n(t) = \int_{s=0}^{s=1} \nabla^2 f_n(g(t)\bar{\boldsymbol{\theta}} + sg(t)\delta(t))g(t)\delta(t)ds$. By multiplying out and using $a_n = 0$ for $n \notin S$ (Lemma 7),

$$\begin{aligned} \dot{\boldsymbol{\theta}}(t) = &\underbrace{\sum_{n \in S} h(t)a_n \nabla f_n(g(t)\boldsymbol{\theta})}_{I} \\ &+ \underbrace{h(t)\sum_{n \in S} a_n \Delta_n(t)}_{II} + \underbrace{h(t)\sum_n \epsilon_n(t)\nabla f_n(g(t)\boldsymbol{\theta})}_{III} + \underbrace{\sum_n h(t)\epsilon_n(t)\Delta_n(t)}_{IV} \end{aligned}$$

Via constraint qualification (Assumption 5), $I = \Omega(g(t)^{\alpha-1}h(t))$ and the second part of Lemma 6, $II = o(I)$.

Since $\epsilon_{tn} = o(1)$, then $III = o(I)$. By the first part of Lemma 6, $IV = O(Bg(t)^{\alpha-1}\|\delta(t)\|) = o(I)$ since $\|\delta(t)\| \to 0$.

Since I is the largest term then after normalization,

$$\frac{\dot{\boldsymbol{\theta}}(t)}{\|\dot{\boldsymbol{\theta}}(t)\|} = \sum_{n \in S} a_n \nabla f_n(g(t)\bar{\theta}) + o(1). \tag{28}$$

Since $\lim\limits_{t\to\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} = \lim\limits_{t\to\infty} \frac{\dot{\boldsymbol{\theta}}(t)}{\|\dot{\boldsymbol{\theta}}(t)\|}$ (Gunasekar et al., 2018b), then

$$\lim_{t\to\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} = \sum_{n \in S} a_n \nabla f_n(g(t)\bar{\boldsymbol{\theta}}). \tag{29}$$

Thus $\bar{\boldsymbol{\theta}}$ satisfies the first-order optimality conditions of 18. □

E.2. Proof of Theorem 3

Proof. The proof is similar to the proof of Theorem 2. From Equations 29 and 28 in the proof of Theorem 2, we see that

$$\lim_{t\to\infty} -\frac{\nabla_\theta \mathcal{L}(\rho\widetilde{\boldsymbol{\theta}}(t))}{\|\nabla_\theta \mathcal{L}(\rho\widetilde{\boldsymbol{\theta}}(t)\|} = \lim_{t\to\infty} \frac{\widetilde{\boldsymbol{\theta}}(t)}{\|\widetilde{\boldsymbol{\theta}}(t)\|}. \qquad \square$$

E.3. Proof of Theorem 4

Proof. The proof is adapted from the ideas outlined in Theorem 7 in (Rosset et al., 2004a).

For any $\boldsymbol{\theta}_\infty \in \Theta_c^\infty$, from definition, let $\{\rho_i, \boldsymbol{\theta}_{\rho_i}\}_{i=1}^\infty$ denote a sequence such that $\rho_i \to \infty$, $\boldsymbol{\theta}_{\rho_i} \in \Theta_c(\rho_i)$ and $\boldsymbol{\theta}_{\rho_i} \to \boldsymbol{\theta}_\infty$. Thus, for any $\epsilon > 0$, $\exists i_0$ such that $\forall i > i_0$, $\|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_{\rho_i}\| \leq \epsilon$.

We need to show that $\boldsymbol{\theta}_\infty \in \Theta^*_{m,N}$. We will prove this theorem by induction, where we show that for all $k = 0, 1, 2, \ldots, N$, $\boldsymbol{\theta}_\infty \in \Theta^*_{m,k}$.

Recall that $\Theta^*_{m,0} = \mathbb{S}^{d-1}$. From the definition of constrained path $\forall i : \boldsymbol{\theta}_{\rho_i} \in \Theta_c(\rho_i) \subseteq \mathbb{S}^{d-1}$. Thus, $\lim_{i\to\infty} \boldsymbol{\theta}_{\rho_i} = \boldsymbol{\theta}_\infty \in \mathbb{S}^{d-1} = \Theta^*_{m,0}$, which proves the base case of induction.

Assume that for some k, $\boldsymbol{\theta}_\infty \in \Theta^*_{m,k}$. We need to show the inductive argument that $\boldsymbol{\theta}_\infty \in \Theta^*_{m,k+1}$.

Recall that for all $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$, we introduced the notation $n^*_\ell(\boldsymbol{\theta}) \in [N]$ reiterated below to denote the index corresponding to the ℓ^{th} smallest margin of $\boldsymbol{\theta}$

$$\begin{aligned} n^*_1(\boldsymbol{\theta}) &= \arg\min_n f_n(\boldsymbol{\theta}) \\ n^*_k(\boldsymbol{\theta}) &= \arg\min_{n \notin \{n^*_\ell(\boldsymbol{\theta})\}_{l=1}^{k-1}} f_n(\boldsymbol{\theta}) \quad \text{for } k \geq 2, \end{aligned} \tag{30}$$

where in the minimization on the right, ties are broken arbitrarily.

Using the above notation, $\Theta^*_{m,k+1}$ is given by

$$\Theta^*_{m,k+1} = \arg\max_{\boldsymbol{\theta}\in\Theta^*_{m,k}} f_{n^*_{k+1}(\boldsymbol{\theta})}(\boldsymbol{\theta}).$$

If possible, let $\boldsymbol{\theta}_\infty \notin \Theta^*_{m,k+1}$ and let $\boldsymbol{\theta}' \in \Theta^*_{m,k+1}$. Using the inductive assumption and the definition of $\Theta^*_{m,k+1}$, we have we have $\boldsymbol{\theta}_\infty, \boldsymbol{\theta}' \in \Theta^*_{m,k}$. From the definition of $\Theta^*_{m,k+1}$, we can deduce the following,

$$\begin{aligned} \forall \ell \leq k,\ f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) &= f_{n^*_\ell(\boldsymbol{\theta}')}(\boldsymbol{\theta}') \\ \gamma := f_{n^*_{k+1}(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) &> f_{n^*_{k+1}(\boldsymbol{\theta}')}(\boldsymbol{\theta}') := \gamma' \end{aligned} \tag{31}$$

Recall that $\mathcal{L}(\boldsymbol{\theta}) = \sum_n \exp(-f_n(\boldsymbol{\theta}))$, where f_n are α-positive homogeneous

Step 1. *Upper bound on $\mathcal{L}(\rho\boldsymbol{\theta}')$.*

$$\begin{aligned} \mathcal{L}(\rho\boldsymbol{\theta}') &= \sum_n \exp(-\rho^\alpha f_n(\boldsymbol{\theta}')) = \sum_{\ell=1}^k \exp(-\rho^\alpha f_{n^*_\ell(\boldsymbol{\theta}')}(\boldsymbol{\theta}')) + \sum_{\ell=k+1}^N \exp(-\rho^\alpha f_{n^*_\ell(\boldsymbol{\theta}')}(\boldsymbol{\theta}')) \\ &\overset{(a)}{\leq} \sum_{\ell=1}^k \exp(-\rho^\alpha f_{n^*_\ell(\boldsymbol{\theta}')}(\boldsymbol{\theta}')) + N\exp(-\rho^\alpha\gamma') \overset{(b)}{=} \sum_{\ell=1}^k \exp(-\rho^\alpha f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + N\exp(-\rho^\alpha\gamma'), \end{aligned} \tag{32}$$

where (a) follows since for all $\ell > k+1$, we have $f_{n^*_\ell(\boldsymbol{\theta}')}(\boldsymbol{\theta}') \geq f_{n^*_{k+1}(\boldsymbol{\theta}')}(\boldsymbol{\theta}') = \gamma'$, and (b) follows since $\forall \ell \leq k$, $f_{n^*_l(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) = f_{n^*_l(\boldsymbol{\theta}')}(\boldsymbol{\theta}')$.

Step 2. *Lower bound on $\mathcal{L}(\rho\boldsymbol{\theta}_\infty)$.*

$$\begin{aligned}\mathcal{L}(\rho\boldsymbol{\theta}_\infty) &= \sum_{\ell=1}^{k}\exp(-\rho^\alpha f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + \sum_{\ell=k+1}^{N}\exp(-\rho^\alpha f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty))\\ &\overset{a}{\geq} \sum_{\ell=1}^{k}\exp(-\rho^\alpha f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + \exp(-\rho^\alpha\gamma),\end{aligned} \tag{33}$$

where (a) follows from using $f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) \geq 0$ for all ℓ and using $f_{n^*_{k+1}(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) = \gamma$.

Step 3. *Lower bound on $\mathcal{L}(\rho_i\boldsymbol{\theta}_{\rho_i})$ for large enough i.* Recall that the sequence of $\boldsymbol{\theta}_{\rho_i} \to \boldsymbol{\theta}_\infty$ satisfies $\boldsymbol{\theta}_{\rho_i} \in \Theta_c(\rho_i)$. From the definition of constrained path, we have for all i, $\mathcal{L}(\rho_i\boldsymbol{\theta}_{\rho_i}) \leq \mathcal{L}(\rho_i\boldsymbol{\theta}_\infty)$.

We first show the following lemma:

Lemma 8. *For any $\epsilon > 0$, $\exists i_0 > 0$ such that $\forall i \geq i_0$, $\mathcal{L}(\rho_i\boldsymbol{\theta}_{\rho_i}) \geq \sum_{\ell=1}^{k}\exp(-\rho_i^\alpha f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + \exp(-\rho_i^\alpha(\gamma-\epsilon))$.*

Proof. Recall that $\forall\ell \leq k$, $\boldsymbol{\theta}_\infty \in \Theta^*_{m,\ell}$. Note from the definition of $n^*_\ell(\boldsymbol{\theta})$ that

$$f_{n^*_\ell(\boldsymbol{\theta})}(\boldsymbol{\theta}) = \min_{\{n_\ell\}_{\ell'=1}^{\ell}}\max_{\ell'\in[\ell]} f_{n_{\ell'}}.$$

Since $\boldsymbol{\theta}_{\rho_i} \to \boldsymbol{\theta}_\infty$, using continuity of min and max of finite number of continuous functions, we have

$$\forall\bar{\epsilon} > 0, \exists i_0(\bar{\epsilon}), \text{such that } \forall i \geq i_0(\bar{\epsilon}), \forall\ell \leq N, \qquad f_{n^*_\ell(\boldsymbol{\theta}_{\rho_i})}(\boldsymbol{\theta}_{\rho_i}) \leq f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) - \bar{\epsilon}. \tag{34}$$

Now consider the two cases for $i \geq i_0(\epsilon)$:

(a) If $\boldsymbol{\theta}_{\rho_i} \in \Theta^*_{m,k}$, then by definition, $\forall\ell \leq k$, $f_{n^*_\ell(\boldsymbol{\theta}_{\rho_i})}(\boldsymbol{\theta}_{\rho_i}) = f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)$. Thus,

$$\begin{aligned}\mathcal{L}(\rho_i\boldsymbol{\theta}_{\rho_i}) &\geq \sum_{\ell=1}^{k}\exp(-\rho_i^\alpha f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + \exp(-\rho_i^\alpha f_{n^*_{k+1}(\boldsymbol{\theta}_{\rho_i})}(\boldsymbol{\theta}_{\rho_i}))\\ &\overset{(a)}{\geq} \sum_{\ell=1}^{k}\exp(-\rho_i^\alpha f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + \exp(-\rho_i^\alpha(\gamma-\epsilon)),\end{aligned}$$

where (a) follows from using eq. 34 to get $f_{n^*_{k+1}(\boldsymbol{\theta}_{\rho_i})}(\boldsymbol{\theta}_{\rho_i}) \leq f_{n^*_{k+1}(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) - \epsilon = \gamma - \epsilon$.

(b) If $\boldsymbol{\theta}_{\rho_i} \notin \Theta^*_{m,k}$, let $\ell \leq k$ be the smallest number such that $\boldsymbol{\theta}_{\rho_i} \notin \Theta^*_{m,l}$. So for all $\ell' < \ell$, $f_{n^*_{\ell'}(\boldsymbol{\theta}_{\rho_i})}(\boldsymbol{\theta}_{\rho_i}) = f_{n^*_{\ell'}(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)$, but $f_{n^*_\ell(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty) - f_{n^*_\ell(\boldsymbol{\theta}_{\rho_i})}(\boldsymbol{\theta}_{\rho_i}) := \epsilon_i > 0$.

$$\begin{aligned}\mathcal{L}(\rho_i\boldsymbol{\theta}_{\rho_i}) &\geq \sum_{\ell'=1}^{\ell-1}\exp(-\rho_i^\alpha f_{n^*_{\ell'}(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + \exp(-\rho_i^\alpha f_{n^*_\ell(\boldsymbol{\theta}_{\rho_i})}(\boldsymbol{\theta}_{\rho_i}))\\ &= \sum_{\ell'=1}^{\ell-1}\exp(-\rho_i^\alpha f_{n^*_{\ell'}(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + \exp(\rho_i^\alpha\epsilon_i)\exp(-\rho_i^\alpha f_{n^*_\ell(\boldsymbol{\theta}_{\rho_\infty})}(\boldsymbol{\theta}_{\rho_\infty}))\end{aligned}$$

On the other hand,

$$\mathcal{L}(\rho_i\boldsymbol{\theta}_\infty) \leq \sum_{\ell'=1}^{\ell-1}\exp(-\rho_i^\alpha f_{n^*_{\ell'}(\boldsymbol{\theta}_\infty)}(\boldsymbol{\theta}_\infty)) + N\exp(-\rho_i^\alpha f_{n^*_\ell(\boldsymbol{\theta}_{\rho_\infty})}(\boldsymbol{\theta}_{\rho_\infty}))$$

Since, $\rho_i \to \infty$ and $\epsilon_i > 0$, for large enough i, we have $\exp(\rho_i^\alpha\epsilon_i) - N > 0$ Thus, for large enough i, from the above two equations, we will have $\mathcal{L}(\rho_i\boldsymbol{\theta}_{\rho_i}) - \mathcal{L}(\rho_i\boldsymbol{\theta}_\infty) > 0$, which is a contradiction, since $\boldsymbol{\theta}_{\rho_i} \in \Theta_c(\rho_i)$. Thus, this case cannot happen for large enough i

This completes the proof of the claim □

Step 4. *Remaining steps in the proof.* For any $\epsilon > 0$, from eqs. 32, 33, and Lemma 8 in Step 3, we have the following for large enough i's

$$\mathcal{L}(\rho_i \boldsymbol{\theta}_{\rho_i}) - \mathcal{L}(\rho_i \boldsymbol{\theta'}) \geq \exp(-\rho_i^{\alpha}(m_1 - \epsilon)) - N \exp(\rho_i^{\alpha} m_2) \overset{(a)}{>} 0. \tag{35}$$

where (a) follows since $m_2 < m_1$ and above equation holds for arbitrarily small ϵ.

In eq. 35, we have obtained a contradiction since for $\boldsymbol{\theta}' \in \mathbb{S}^{d-1}$, $\boldsymbol{\theta}_{\rho_i} \in \Theta_c(\rho_i) \implies \mathcal{L}(\rho_i \boldsymbol{\theta}_{\rho_i}) \leq \mathcal{L}(\rho_i \boldsymbol{\theta}')$.

This completes the proof of the theorem. □

F. The Regularization Path

F.1. The Regularization Path

The regularization path is given by the following set, $\forall c > 0$:

$$\Theta_r(c) = \left\{ \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} : \ \arg\min_{\boldsymbol{\theta}} \ \mathcal{L}\left(\boldsymbol{\theta}\right) + \frac{1}{c} \left\|\boldsymbol{\theta}\right\|^2 \right\}. \tag{36}$$

Lemma 9. $\forall c > 0 : \ \Theta_r\left(c\right)$ *is not empty, i.e.,* $\forall c : \ \min_{\boldsymbol{\theta}} \ \mathcal{L}\left(\boldsymbol{\theta}\right) + \frac{1}{c}\left\|\boldsymbol{\theta}\right\|^2$ *exists.*

Proof. Note that $\forall c > 0 : \ \mathcal{L}\left(\boldsymbol{\theta}\right) + \frac{1}{c}\left\|\boldsymbol{\theta}\right\|^2$ is coercive since $\mathcal{L}\left(\boldsymbol{\theta}\right)$ is lower bounded. Thus, the minimum of $\mathcal{L}\left(\boldsymbol{\theta}\right) + \frac{1}{c}\left\|\boldsymbol{\theta}\right\|^2$ is attained as the minimum of a continuous coercive function over a nonempty closed set. □

If Assumption 1 is satisfied, then as $c \to \infty$ we have that $\left\|\boldsymbol{\theta}_r\left(c\right)\right\| \to \infty$ where $\boldsymbol{\theta}_r\left(c\right) \in \Theta_r\left(c\right)$. We state this result in the following lemma.

Lemma 10. *If* $\exists \rho_0$ *such that* $\mathcal{L}^*\left(\rho\right)$ *is strictly monotonically decreasing for any* $\rho \geq \rho_0$*, and* $\boldsymbol{\theta}_r\left(c\right) \in \arg\min_{\boldsymbol{\theta}} \ \mathcal{L}\left(\boldsymbol{\theta}\right) + \frac{1}{c}\left\|\boldsymbol{\theta}\right\|^2$ *then as* $c \to \infty$*, we have* $\left\|\boldsymbol{\theta}_r\left(c\right)\right\| \to \infty$.

Proof. From lemma 9 we have that $\forall c > 0$, $\arg\min_{\boldsymbol{\theta}} \ \mathcal{L}\left(\boldsymbol{\theta}\right) + \frac{1}{c}\left\|\boldsymbol{\theta}\right\|^2$ has an optimal solution (at least one). We assume, in contradiction, that $\exists M > \rho_0$ so that $\forall c_0 : \exists c > c_0$ with $\left\|\boldsymbol{\theta}_r\left(c\right)\right\| \leq M$. For some $\epsilon > 0$ we denote $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \mathcal{L}\left(\left(M+\epsilon\right)\boldsymbol{\theta}\right)$, i.e., $\mathcal{L}\left(\left(M+\epsilon\right)\boldsymbol{\theta}^*\right) = \mathcal{L}^*\left(M+\epsilon\right)$. We have that

$$\begin{aligned}
&\mathcal{L}\left(\left(M+\epsilon\right)\boldsymbol{\theta}^*\right) + \frac{1}{c}\|\left(M+\epsilon\right)\boldsymbol{\theta}^*\| \\
&= \mathcal{L}\left(\boldsymbol{\theta}_r\left(c\right)\right) + \frac{1}{c}\|\boldsymbol{\theta}_r\left(c\right)\| + \mathcal{L}\left(\left(M+\epsilon\right)\boldsymbol{\theta}^*\right) - \mathcal{L}\left(\boldsymbol{\theta}_r\left(c\right)\right) + \frac{1}{c}\|\left(M+\epsilon\right)\boldsymbol{\theta}^*\| - \frac{1}{c}\|\boldsymbol{\theta}_r\left(c\right)\| \\
&\leq \mathcal{L}\left(\boldsymbol{\theta}_r\left(c\right)\right) + \frac{1}{c}\|\boldsymbol{\theta}_r\left(c\right)\| + \mathcal{L}^*\left(M+\epsilon\right) - \mathcal{L}^*\left(M\right) + \frac{1}{c}\|\left(M+\epsilon\right)\boldsymbol{\theta}^*\|
\end{aligned}$$

Note that $\mathcal{L}^*\left(M+\epsilon\right) - \mathcal{L}^*\left(M\right) < 0$ since we assume that $\mathcal{L}^*\left(\rho\right)$ is strictly monotonically decreasing for any $\rho \geq \rho_0$. For sufficiently large c we get that

$$\mathcal{L}\left(\left(M+\epsilon\right)\boldsymbol{\theta}^*\right) + \frac{1}{c}\|\left(M+\epsilon\right)\boldsymbol{\theta}^*\| < \mathcal{L}\left(\boldsymbol{\theta}_r\left(c\right)\right) + \frac{1}{c}\|\boldsymbol{\theta}_r\left(c\right)\|$$

which contradicts our assumption that $\boldsymbol{\theta}_r\left(c\right)$ is an optimal solution. □

F.2. Connections between regularization and constrained paths

For convex loss function, the regularization and constrained paths are known to be equivalent. For general loss function, we state the following basic result.

Lemma 11. *$\forall c > 0, \forall \theta_r \in \Theta_r(c) : \exists \rho$ so that $\theta_r \in \Theta_c(\rho)$, and, If $\exists \rho_0$ such that $\mathcal{L}^*(\rho)$ is strictly monotonically decreasing for any $\rho \geq \rho_0$ then $\forall \theta_r \in \Theta_r(c)$: $\Theta_c(\|\theta_r\|) \subset \Theta_r(c)$.*

Proof. To prove Lemma 11 we combine the results from the following two lemmas. □

Lemma 12. *$\forall c > 0, \forall \theta_r \in \Theta_r(c) : \exists \rho$ so that $\theta_r \in \Theta_c(\rho)$.*

Proof. For some $c > 0$, let $\boldsymbol{\theta}_r^*(c) \in \Theta_r(c)$. From $\Theta_r(c)$ definition (eq. 36) $\|\boldsymbol{\theta}_r^*(c)\| = 1$ and thus $\boldsymbol{\theta}_r^*(c)$ is a feasible solution of eq. 5. Additionally, $\exists \alpha > 0$ so that $\forall \boldsymbol{\theta}^{(1)}$:

$$\mathcal{L}(\alpha\boldsymbol{\theta}_r^*(c)) + \frac{1}{c}\|\alpha\boldsymbol{\theta}_r^*(c)\|^2 \leq \mathcal{L}\left(\alpha\boldsymbol{\theta}^{(1)}(c)\right) + \frac{1}{c}\left\|\alpha\boldsymbol{\theta}^{(1)}(c)\right\|^2 .$$

For $\rho = \alpha$ we have that $\boldsymbol{\theta}_r^*(c) \in \Theta_c(\rho)$ since $\forall \boldsymbol{\theta}^{(1)}$:

$$\mathcal{L}(\rho\boldsymbol{\theta}_r^*(c)) = \mathcal{L}(\rho\boldsymbol{\theta}_r^*(c)) + \frac{\rho^2}{c}\|\boldsymbol{\theta}_r^*(c)\|^2 - \frac{\rho^2}{c} \leq \mathcal{L}\left(\rho\boldsymbol{\theta}^{(1)}(c)\right) + \frac{\rho^2}{c}\left\|\boldsymbol{\theta}^{(1)}(c)\right\|^2 - \frac{\rho^2}{c}$$

and particularly, $\forall \boldsymbol{\theta}^{(2)}$ such that $\left\|\boldsymbol{\theta}^{(2)}\right\| \leq 1$:

$$\mathcal{L}(\rho\boldsymbol{\theta}_r^*(c)) \leq \mathcal{L}\left(\rho\boldsymbol{\theta}^{(2)}(c)\right) + \frac{\rho^2}{c}\left\|\boldsymbol{\theta}^{(2)}(c)\right\|^2 - \frac{\rho^2}{c} \leq \mathcal{L}\left(\rho\boldsymbol{\theta}^{(2)}(c)\right) .$$

□

Lemma 13. *$\forall c > 0, \forall \boldsymbol{\theta}_r \in \Theta_r(c)$: $\Theta_c(\|\boldsymbol{\theta}_r\|) \subset \Theta_r(c)$.*

Proof. Note that from Lemma 2

$$\Theta_c(\rho) = \arg\min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|\leq 1}\mathcal{L}(\rho\boldsymbol{\theta}) = \left\{\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} : \arg\min_{\boldsymbol{\theta}}\mathcal{L}\left(\rho\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right)\right\} .$$

For some $c > 0$, let $\boldsymbol{\theta}_r^*(c) \in \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \frac{1}{c}\|\boldsymbol{\theta}\|^2$. For $\rho = \|\boldsymbol{\theta}_r^*(c)\|$ and $\boldsymbol{\theta}_c^* \in \arg\min_{\boldsymbol{\theta}}\mathcal{L}\left(\rho\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right)$ we have that $\forall \boldsymbol{\theta}$

$$\mathcal{L}\left(\rho\frac{\boldsymbol{\theta}_c^*}{\|\boldsymbol{\theta}_c^*\|}\right) \leq \mathcal{L}\left(\rho\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right) .$$

Thus, we have that

$$\mathcal{L}\left(\rho\frac{\boldsymbol{\theta}_c^*}{\|\boldsymbol{\theta}_c^*\|}\right) + \frac{1}{c}\|\boldsymbol{\theta}_c^*\|^2 \leq \mathcal{L}\left(\rho\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right) + \frac{1}{c}\|\boldsymbol{\theta}_c^*\|^2$$

Thus, $\forall \boldsymbol{\theta}^{(2)}$ so that $\left\|\boldsymbol{\theta}^{(2)}\right\| = \rho = \|\boldsymbol{\theta}_r^*(c)\|$ we have that

$$\mathcal{L}\left(\rho\frac{\boldsymbol{\theta}_c^*}{\|\boldsymbol{\theta}_c^*\|}\right) + \frac{\rho^2}{c} \leq \mathcal{L}\left(\boldsymbol{\theta}^{(2)}\right) + \frac{1}{c}\left\|\boldsymbol{\theta}^{(2)}\right\|^2 .$$

In particular, this implies that

$$\mathcal{L}\left(\rho\frac{\boldsymbol{\theta}_c^*}{\|\boldsymbol{\theta}_c^*\|}\right) + \frac{\rho^2}{c} \leq \mathcal{L}(\boldsymbol{\theta}_r^*(c)) + \frac{1}{c}\|\boldsymbol{\theta}_r^*(c)\|^2 .$$

□