Flavors of Margin: Implicit Bias of Steepest Descent in Homogeneous Neural Networks

Nikolaos Tsilivis
New York University
nt2231@nyu.edu

Gal Vardi
Weizmann Institute of Science
gal.vardi@weizmann.ac.il

Julia Kempe
New York University & Meta FAIR
kempe@meta.com

Abstract

We study the *implicit bias* of the general family of *steepest descent algorithms*, which includes gradient descent, sign descent and coordinate descent, in deep homogeneous neural networks. We prove that an algorithm-dependent geometric margin starts increasing once the networks reach perfect training accuracy and characterize the late-stage bias of the algorithms. In particular, we define a generalized notion of stationarity for optimization problems and show that the algorithms progressively reduce a (generalized) Bregman divergence, which quantifies proximity to such stationary points of a margin-maximization problem. We then experimentally zoom into the trajectories of neural networks optimized with various steepest descent algorithms, highlighting connections to the implicit bias of *Adam*.

1 Introduction

Overparameterized neural networks excel in many natural supervised learning applications. A theory that aims to explain their strong generalization performance places optimization at the forefront: in problems where many candidate models are available, the optimization algorithm implicitly selects well-generalizing ones (Neyshabur et al., 2015b). The term "implicitly" indicates that neither the unregularized loss function nor the architecture explicitly favors simple, well-generalizing solutions, but this occurs due to the choice of the optimization algorithm. Most existing theoretical results on this so-called *implicit bias of optimization* demonstrate, to some extent, that gradient descent in overparameterized problems biases the solution to be the *simplest*, in terms of the lowest possible ℓ_2 norm of the weights (Soudry et al., 2018; Ji & Telgarsky, 2019, 2020; Lyu & Li, 2020; Nacson et al., 2019).

Simplicity, however, is a term that depends on the setting. For instance, in logistic regression with many irrelevant features, an ℓ_1-regularized solution is simpler than an ℓ_2-regularized one (Ng, 2004). Moreover, in contemporary deep learning, Adam (Kingma & Ba, 2015), AdamW (Loshchilov & Hutter, 2019), and related optimization algorithms are preferred for language modeling (Zhang et al., 2020), and their implicit bias might be better suited for such applications than gradient descent. It is therefore important to understand the types of solutions favored by optimization algorithms beyond (stochastic) gradient descent, in order to address the current (and future) applications of deep learning.

In this work, we contribute to this line of research by studying the large family of *steepest descent* algorithms (Equation 1) with respect to an arbitrary norm $\|\cdot\|$ in deep, non-linear, homogeneous

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

neural networks. This class of methods generalizes gradient descent to optimization geometries other than the Euclidean, allowing the update rule to operate under a different norm. It includes coordinate descent (which has strong ties to boosting (Mason et al., 1999)) and sign gradient descent (which is closely related to Adam (Kunstner et al., 2023)) as special cases.

Our contributions. We provide a unifying, rigorous analysis of any steepest descent algorithm in classification settings with locally-Lipschitz, homogeneous neural networks trained using an exponentially-tailed loss. Specifically, we focus on the late stage of training (after the network has achieved perfect training accuracy) in the limit of an infinitesimal learning rate. Our first result characterizes the algorithm's tendency to increase an algorithm-dependent margin (Theorem 3.1): similar to prior work on gradient descent (Lyu & Li, 2020), we show that a *soft* version of the geometric margin starts increasing immediately after fitting the training data.

We then turn our attention to the asymptotic properties of the algorithm. In an attempt to find evidence of *margin maximization*, we define a notion of stationary points for optimization problems, which generalizes the usual Karush-Kuhn-Tucker one (Definition 3.4), along with approximate versions (Definition A.9). As we show, during training, the algorithms make implicit progress towards such stationary points of a margin maximization problem in a specific, geometric sense: they progressively reduce a *(generalized) Bregman divergence* (Definition 3.6), which quantifies how well the stationarity condition is satisfied. As this process concludes, the limit points of training are along the direction of a generalized KKT point of the algorithm-dependent geometric margin maximization problem (Theorem 3.8). For algorithms whose squared norm is a smooth function (for example, any ℓ_p norm for $p \geq 2$), Theorem 3.8 further implies directional convergence to KKT points of the same margin maximization problem (Corollary 3.8.1).

In total, these results provide evidence for (geometric) margin-maximization in any steepest descent algorithm and significantly generalize prior results that concerned gradient descent only (Lyu & Li, 2020; Nacson et al., 2019). Moreover, the generalized divergence can be interpreted as a measure of proximity to stationarity for optimization problems, similarly to what was proposed in prior definitions of approximate KKT points in the literature (Dutta et al., 2013), and could be of broader interest to the optimization community. We find it appealing and theoretically intriguing that, despite the non-convexity of the loss landscape of deep neural networks, such simple convex structures emerge once the data points separate.

Finally, in Section 4, we train neural networks with the three main steepest descent algorithms (gradient descent, sign gradient descent and coordinate descent). We perform experiments in: (a) teacher-student tasks, to elucidate the theoretical findings and assess the connection between implicit bias and generalization and (b) image classification tasks, to study the relationship between Adam and steepest descent algorithms.

1.1 Related Work

There have been numerous works studying the implicit bias of optimization in supervised learning and their relationship to geometric margin maximization - see (Vardi, 2023) for a survey.

Steepest descent algorithms with respect to non-Euclidean geometries have been explored before, both in supervised (e.g. Neyshabur et al. (2015a); Large et al. (2024)) and non-supervised (e.g. Carlson et al. (2015)) machine learning problems. The implicit bias of this family of optimization methods was first studied in generality in Gunasekar et al. (2018) in the context of linear models for separable data, where margin maximization was established. Their proof is based on a result on Adaboost due to Telgarsky (2013). Most related to our paper are the works of Nacson et al. (2019); Lyu & Li (2020) and Ji & Telgarsky (2020). Nacson et al. (2019) studied infinitesimal regularization and its connection to margin maximization in both homogeneous and non-homogeneous deep models, while also proving (directional) convergence of gradient descent to a first order point of an ℓ_2-margin maximization problem for homogeneous models under strong technical assumptions. Lyu & Li (2020); Ji & Telgarsky (2020), whose theoretical setup we mainly follow, significantly weakened the assumptions, under which such a result holds, and Lyu & Li (2020) further demonstrated the experimental benefits of margin maximization in terms of robustness. Kunin et al. (2023) generalized these results to a broader class of networks with varying degree of homogeneity, which includes networks with biases and skip connections. Vardi et al. (2022) identified cases where the KKT points of the ℓ_2 margin maximization problems are not (even locally) optimal.

The implicit bias of Adam (Kingma & Ba, 2015) has been previously studied in Wang et al. (2021, 2022) for homogeneous networks, where it is proven that it shares the same asymptotic properties as gradient descent (ℓ_2 margin maximization). Recently, Zhang et al. (2024) analyzed a version of the algorithm in linear models, without a numerical precision constant, which arguably better captures realistic training runs, and found bias towards ℓ_1 margin maximization - the same bias as in the case of sign gradient descent. This makes us optimistic that insights from our analysis, which covers sign gradient descent (steepest descent with respect to the ℓ_∞ norm), can shed light on the poorly understood implicit bias of Adam in deep neural networks. See also Xie & Li (2024) for a recently established connection between AdamW (Loshchilov & Hutter, 2019) and sign gradient descent. An additional motivation for studying steepest descent algorithms is in improving the robustness of deep neural networks: Tsilivis et al. (2024), recently, provided experimental evidence and theoretical arguments that deep networks adversarially trained with different steepest descent algorithms exhibit significant differences in their (robust) generalization error.

2 Background

Learning Setup We consider binary classification problems with deep, homogeneous, neural networks. Formally, let $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^m$ be a dataset of i.i.d. points sampled from an unknown distribution $\mathcal{D}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$ for all $i \in [m]$, and let $f(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ denote a neural network parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$. The vector $\boldsymbol{\theta}$ contains all the parameters of the neural network, concatenated into a single vector. We study training under an exponential loss $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i; \boldsymbol{\theta})}$. We focus on this setting for simplicity in the main text, but our results should readily generalize to more common losses, such as the logistic loss, as well as its multi-class generalization - the cross-entropy loss. See Section A.3 for details and extensions of our main result.

Algorithms The family of steepest descent algorithms generalizes gradient descent to different optimization geometries, allowing the update rule to operate under an arbitrary norm (instead of the usual Euclidean one) (Boyd & Vandenberghe, 2014). Formally, the update rule for *steepest descent* with respect to a norm $\|\cdot\|$ is:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \eta_t \Delta\boldsymbol{\theta}_t, \text{ where } \Delta\boldsymbol{\theta}_t \text{ satisfies} \\ \Delta\boldsymbol{\theta}_t &= \underset{\|\mathbf{u}\| \leq \|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_\star}{\arg\min} \langle \mathbf{u}, \nabla\mathcal{L}(\boldsymbol{\theta}_t) \rangle , \end{aligned} \tag{1}$$

where the *dual* norm $\|\cdot\|_\star$ of $\|\cdot\|$ is defined as $\|\mathbf{z}\|_\star = \max_{\mathbf{v}}\{|\langle \mathbf{z}, \mathbf{v}\rangle| : \|\mathbf{v}\| = 1\}$ for any $\mathbf{z}$, and η_t is a learning rate. Gradient descent can be derived from Equation 1 with $\|\cdot\| = \|\cdot\|_2$. See Appendix C for details on how steepest descent algorithms are closely related to popular adaptive methods, such as Adam (Kingma & Ba, 2015) and Shampoo (Gupta et al., 2018).

Assumptions & Technical Points In order to formally allow commonly used activation functions, such as the ReLU, we theoretically analyze loss landscapes that are not necessarily differentiable. That is, we consider Clarke's subdifferentials (Clarke, 1975) in our analysis:

$$\partial f := \operatorname{conv}\left\{ \lim_{k\to\infty} \nabla f(\mathbf{x}_k) : \mathbf{x}_k \to \mathbf{x}, f \text{ differentiable at } \mathbf{x}_k \right\}, \tag{2}$$

where $\operatorname{conv}(\cdot)$ stands for the convex hull of a set.
Furthermore, we analyze steepest descent in the limit of infinitesimal step size, i.e. *steepest flow*:

$$\frac{d\boldsymbol{\theta}}{dt} \in \left\{ \underset{\|\mathbf{u}\| \leq \|\mathbf{g}_t\|_\star}{\arg\min} \langle \mathbf{u}, \mathbf{g}_t \rangle : \mathbf{g}_t \in \partial\mathcal{L}(\boldsymbol{\theta}_t) \right\}. \tag{3}$$

This choice simplifies the analysis while still capturing the essence of the bias of the algorithms. Finally, we make the following assumptions:

(A1) Local Lipschitzness: For any $\mathbf{x}_i \in \mathbb{R}^d$, $f(\mathbf{x}_i; \cdot) : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz (and admits a chain rule - see Theorem A.2).

(A2) L-Homogeneity: We assume that f is L-homogeneous in the parameters, i.e. $f(\cdot; c\boldsymbol{\theta}) = c^L f(\cdot; \boldsymbol{\theta})$ for any $c > 0$.

(A3) Realizability. There is a $t_0 > 0$, such that $\mathcal{L}(\boldsymbol{\theta}_{t_0}) < 1$.

Assumption (A1) is a minimal assumption on the regularity of the network, while assumption (A2) includes many commonly used architectures. For instance, ReLU networks with an arbitrary number of layers, but without bias terms, satisfy (A1),(A2). Assumption (A3) ensures that the algorithm will succeed in classifying the training points and allows us to focus on what happens beyond that point of separation. Indeed, we are particularly interested in understanding the geometric properties of the model $f(\cdot;\boldsymbol{\theta}_t)$ as $t \to \infty$ (at convergence) – the so-called *implicit biases* of the learning algorithms.

3 Theory

We analyze the behavior of steepest descent algorithms in the late stage of training and study their geometric properties and how these relate to geometric, algorithm-specific, margins.

3.1 Algorithm-Dependent Margin Increases

In *linear* models, where $f(\mathbf{x};\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$, the concept of $\|\cdot\|_\star$*-geometric margin* [1], $\min_{i\in[m]} \frac{y_i \langle \boldsymbol{\theta}, \mathbf{x}_i \rangle}{\|\boldsymbol{\theta}\|}$, plays a central and fundamental role in the analysis of the convergence of training (Novikoff, 1963) as well as in the generalization of the final model (Vapnik, 1998). Ideally, we would like to track a similar quantity when training general, homogeneous, non-linear networks $f(\cdot;\boldsymbol{\theta})$ with steepest descent with respect to the $\|\cdot\|$ norm:

$$\gamma(\boldsymbol{\theta}) = \frac{\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|^L} = \min_{i\in[m]} y_i f\left(\mathbf{x}_i; \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right), \tag{4}$$

where recall that L is the level of homogeneity of the model. As it turns out, it is easier to follow the evolution of the following, *soft*, geometric margin:

$$\widetilde{\gamma}(\boldsymbol{\theta}) = -\frac{\log \mathcal{L}(\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|^L}. \tag{5}$$

The characterisation of "soft" comes from the definition of "softmax" (a.k.a. log-sum-exp), which is often used in machine learning. The same idea is used here to approximate the numerator of Equation 4. The soft margin $\widetilde{\gamma}(\boldsymbol{\theta})$ is at most an additive $\log m$ away from $\gamma(\boldsymbol{\theta})$ - see Lemma A.7.

We show next that, given the algorithm has reached a small value in the loss, the soft margin is non-decreasing. This theorem is similar to part of Lemma 5.1 in (Lyu & Li, 2020), which is the key lemma in their result. Ours is admittedly simpler, avoiding a beautiful polar decomposition which was crucial in their analysis, yet, unfortunately, pertinent to the ℓ_2 case only.

Theorem 3.1 (Soft margin increases)**.** *For almost any $t > t_0$, it holds:*

$$\frac{d \log \widetilde{\gamma}}{dt} \geq L \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2 \left(\frac{1}{L\mathcal{L}(\boldsymbol{\theta}_t) \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}} - \frac{1}{\|\boldsymbol{\theta}_t\| \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|} \right) \geq 0.$$

Proof of a simplified version. We present a proof for a simplified version of this theorem here, covering differentiable networks f, while we defer the full proof to Appendix A.2. For differentiable losses, steepest flow corresponds to:

$$\frac{d\boldsymbol{\theta}}{dt} \in \underset{\|\mathbf{u}\| \leq \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_\star}{\arg\min} \langle \mathbf{u}, \nabla \mathcal{L}(\boldsymbol{\theta}_t) \rangle. \tag{6}$$

By the definition of the dual norm and chain rule, we have for any $t > 0$:

$$\left\| \frac{d\boldsymbol{\theta}}{dt} \right\| = \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_\star \quad \text{and} \quad \frac{d\mathcal{L}(\boldsymbol{\theta}_t)}{dt} = -\left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2. \tag{7}$$

[1] In this paper, we diverge from the established terminology when it comes to naming margins, by calling it $\|\cdot\|_\star$-geometric margin (instead of $\|\cdot\|$-geometric margin) when it is defined with respect to the $\|\cdot\|$ norm of the parameters. We believe this is proper, since the $\|\cdot\|_\star$-geometric margin in linear models maximizes the metric induced by the $\|\cdot\|_\star$ norm (and not its dual, $\|\cdot\|$).

Let $\mathbf{n}_t \in \partial\|\boldsymbol{\theta}_t\|$ (recall that a norm $\|\cdot\|$ might not be differentiable everywhere). For any $t > t_0$, we have:

$$\begin{aligned}
\frac{d\log\widetilde{\gamma}}{dt} &= \frac{d}{dt}\log\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)} - L\frac{d}{dt}\log\|\boldsymbol{\theta}_t\| \\
&= \frac{d}{dt}\log\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)} - L\left\langle\frac{\mathbf{n}_t}{\|\boldsymbol{\theta}_t\|}, \frac{d\boldsymbol{\theta}}{dt}\right\rangle && \text{(Chain rule)} \\
&\geq \frac{d}{dt}\log\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)} - L\frac{\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} && \text{(def. of dual norm and } \|\mathbf{n}_t\|_\star \leq 1\text{, Lemma A.4)} \\
&= -\frac{d\mathcal{L}(\boldsymbol{\theta}_t)}{dt}\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}} - L\frac{\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} && \text{(Chain rule)} \\
&= \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2\left(\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}} - \frac{L}{\|\boldsymbol{\theta}_t\|\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}\right). && \text{(Equation 7)}
\end{aligned} \tag{8}$$

The first term inside the parenthesis can be related to the second one via the following calculation:

$$\begin{aligned}
\langle\boldsymbol{\theta}_t, -\nabla\mathcal{L}(\boldsymbol{\theta}_t)\rangle &= \left\langle\boldsymbol{\theta}_t, \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} y_i \nabla f(\mathbf{x}_i;\boldsymbol{\theta}_t)\right\rangle \\
&= \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} y_i \langle\boldsymbol{\theta}_t, \nabla f(\mathbf{x}_i;\boldsymbol{\theta}_t)\rangle \\
&= L\sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t),
\end{aligned} \tag{9}$$

where the last equality follows from Euler's theorem for homogeneous functions. Now, observe that this last term can be lower bounded as:

$$\langle\boldsymbol{\theta}_t, -\nabla\mathcal{L}(\boldsymbol{\theta}_t)\rangle \geq L\sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} \min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t) \geq L\mathcal{L}(\boldsymbol{\theta}_t)\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}, \tag{10}$$

where we used the fact $e^{-\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} \leq \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} = \mathcal{L}(\boldsymbol{\theta}_t)$. We have made the first term of Equation 8 appear. By plugging Equation 10 into Equation 8, we get:

$$\begin{aligned}
\frac{d\log\widetilde{\gamma}}{dt} &\geq \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2\left(\frac{L}{\langle\boldsymbol{\theta}_t, -\nabla\mathcal{L}(\boldsymbol{\theta}_t)\rangle} - \frac{L}{\|\boldsymbol{\theta}_t\|\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}\right) \\
&\geq \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2\left(\frac{L}{\|\boldsymbol{\theta}_t\|\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_\star} - \frac{L}{\|\boldsymbol{\theta}_t\|\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}\right). \quad \text{(definition of dual norm)}
\end{aligned} \tag{11}$$

Noticing that $\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_\star = \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|$ (from Equation 7) concludes the proof. □

Remark 3.2. *Observe that it is the geometric margin induced by the dual norm of the algorithm that is non-decreasing, and not any geometric margin. The proof crucially relies on the fact that* $\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_\star = \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|$.

3.2 Convergence to Generalized Stationary Points of the Max-Margin Problem

The previous theorem provides evidence and is a first indication that steepest flow implicitly maximizes the $\|\cdot\|_\star$-geometric margin in deep neural networks. However, the monotonicity of the (soft) margin alone does not imply anything about its final value and its optimality. In this section, we provide a concrete characterization of the *asymptotic* behavior of steepest flow: we show that any limit point of the iterates produced by steepest flow is along the direction of a *generalized KKT* point of the following margin maximization (MM) optimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\
&\text{s.t. } y_i f(\mathbf{x}_i;\boldsymbol{\theta}) \geq 1,\ \forall i \in [m].
\end{aligned} \tag{MM}$$

Let us recall the definition of a Karush-Kuhn-Tucker point (Karush, 1939; Kuhn, H. W. and Tucker, A. W., 1951).

Definition 3.3 (KKT point)**.** *A feasible point $\boldsymbol{\theta} \in \mathbb{R}^p$ of (MM) is a Karush-Kuhn-Tucker (KKT) point, if there exist $\lambda_1, \ldots, \lambda_m \geq 0$ such that:*

1. $\partial \frac{1}{2}\|\boldsymbol{\theta}\|^2 + \sum_{i=1}^m \lambda_i \partial \left(1 - y_i f(\mathbf{x}_i; \boldsymbol{\theta})\right) \ni 0.$
2. $\lambda_i(1 - y_i f(\mathbf{x}_i; \boldsymbol{\theta})) = 0, \ \forall i \in [m].$

Notice that the first so-called *stationarity* condition is defined using set addition, since we are dealing with non-differentiable functions. See (Dutta et al., 2013) for more details on optimization problems with non-smooth objectives/constraints. Under some regularity assumptions, the KKT conditions become necessary conditions for global optimality and for non-convex problems like (MM) they might be the best characterization of optimality we can hope for. See Lemma A.11 for details.

In the following definition of generalized KKT points, we relax the stationarity condition and parameterize it by a non-negative function.

Definition 3.4 (d-generalized KKT point)**.** *Let $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$. A feasible point $\boldsymbol{\theta} \in \mathbb{R}^p$ of (MM) is called a d-**generalized KKT** point if there exist $\lambda_1, \ldots, \lambda_m \geq 0$, $\mathbf{h}_i \in \partial f(\mathbf{x}_i; \boldsymbol{\theta})$ and $\mathbf{k} \in \partial \frac{1}{2}\|\boldsymbol{\theta}\|^2$ such that:*

1. $d\left(\sum_{i=1}^m \lambda_i y_i \mathbf{h}_i, \mathbf{k}\right) = 0.$
2. $\lambda_i(1 - y_i f(\mathbf{x}_i; \boldsymbol{\theta})) = 0, \ \forall i \in [m].$

Remark 3.5. *When d in the definition of a d-generalized KKT point is any metric, we readily recover the original definition of KKT point.*

In Appendix B, we demonstrate how to construct a progress measure for optimization problems leveraging the above notion of stationarity (as well as its approximate version - see Definition A.9).

As we will show, the function d, which in our case measures proximity of steepest flow to stationarity, is a generalized *Bregman divergence* induced by the dual norm of the algorithm (squared).

Definition 3.6 (Generalized Bregman divergence)**.** *Let $\psi : \mathbb{R}^p \to \mathbb{R}$ with $\psi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_\star^2$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$. We define the (generalized) Bregman divergence $D^{\mathbf{m}}_{\frac{1}{2}\|\cdot\|_\star^2}(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ induced by ψ as follows:*

$$D^{\mathbf{m}}_{\frac{1}{2}\|\cdot\|_\star^2}(\mathbf{y}, \mathbf{z}) = \frac{1}{2}\|\mathbf{y}\|_\star^2 - \frac{1}{2}\|\mathbf{z}\|_\star^2 - \langle \mathbf{m}, \mathbf{y} - \mathbf{z} \rangle, \tag{12}$$

where $\mathbf{m} \in \partial \frac{1}{2}\|\mathbf{z}\|_\star^2$.

Remark 3.7. *Notice that if the function $\psi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_\star^2$ is differentiable, then the subdifferental defined at any point collapses to a single element: the gradient of ψ. If, further, ψ is strictly convex, then Equation 12 coincides with the usual Bregman divergence induced by ψ, defined as $D_\psi(\mathbf{y}, \mathbf{z}) = \psi(\mathbf{y}) - \psi(\mathbf{z}) - \langle \nabla\psi(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle$. Bregman divergences (Bregman, 1967) generalize the Euclidean squared distance in different geometries and have found numerous applications in machine learning (A. Nemirovskii and D. Yudin, 1983; Banerjee et al., 2005).*

We are now ready to state our main result.

Theorem 3.8. *Under assumptions (A1), (A2), (A3), consider steepest flow with respect to a norm $\|\cdot\|$ (Equation 3) on the exponential loss $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i; \boldsymbol{\theta})}$. Then, any limit point $\bar{\boldsymbol{\theta}}$ of $\left\{\frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|}\right\}_{t \geq 0}$ is along the direction of a $D^{\widetilde{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}$-generalized KKT point, of the following optimization problem:*

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ &\text{s.t. } y_i f(\mathbf{x}_i; \boldsymbol{\theta}) \geq 1, \ \forall i \in [m], \end{aligned} \tag{13}$$

where $D^{\widetilde{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}$ is a (generalized) Bregman divergence induced by $\frac{1}{2}\|\cdot\|_\star^2$ and $\widetilde{\boldsymbol{\theta}} = \left(\min_{i \in [m]} y_i f(\mathbf{x}_i; \bar{\boldsymbol{\theta}})\right)^{-\frac{1}{L}} \bar{\boldsymbol{\theta}}$.

Theorem 3.8 states that the iterates induced by steepest flow have very specific, geometric properties: not only do they asymptotically approach (in direction) a generalized KKT point of a margin

maximization problem, but also, as the proof of this theorem and, in particular, Proposition A.15 tells us, they implicitly make progress towards stationarity by decreasing a Bregman divergence between the gradient of the objective function and the gradient of the constraints of (MM). The full proof can be found in Appendix A.

While the previous result is not strong enough to guarantee convergence to a KKT point for any instantiation of the algorithm norm $\|\cdot\|$, it immediately implies it in the case of a norm whose square is a *smooth* function. We can prove the following corollary for this special class of steepest flows.

Corollary 3.8.1. *Under assumptions (A1), (A2), (A3), any limit point $\bar{\boldsymbol{\theta}}$ of $\left\{\frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|}\right\}_{t\geq 0}$ produced by steepest flow (Equation 3) with respect to a norm $\|\cdot\|$, whose square is a smooth function, on the exponential loss, is along the direction of a KKT point of the optimization problem (MM).*

The proof relies on a fundamental relationship between smoothness of a function and strong convexity of its convex conjugate (Proposition A.19), and can be found in Appendix A. The main contribution of Lyu & Li (2020), which characterizes the implicit bias of gradient flow in homogeneous deep networks, can be recovered by the above result when $\|\cdot\| = \|\cdot\|_2$. Additionally, Corollary 3.8.1 generalizes this result in at least the following cases of algorithm norms:

- Any ℓ_p norm with $p \in [2, \infty)$ – see, for example, Lemma 17 in Shalev-Shwartz (2007) for a proof of its smoothness.
- Any norm induced by a positive definite symmetric matrix – i.e. $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle, \mathbf{x} \in \mathbb{R}^d, \mathbf{A} \in \mathbb{R}^{d\times d}$.
- Any $(2, D)$-smooth norm (i.e. a norm such that $\frac{d^2}{dt^2}\|\mathbf{x} + t\mathbf{z}\|^2 \leq 2D^2\|\mathbf{z}\|^2$ for all $\mathbf{x}, \mathbf{z}$ and for a $D > 0$) – see Kakade et al. (2008) for details.

To the best of our knowledge, this is a first result about the implicit bias of an algorithm in the parameter space of deep homogeneous neural networks which is not about ℓ_2-geometric margin maximization.

4 Experiments

In this section, we train neural networks with various steepest descent algorithms (gradient descent-`GD`, coordinate descent-`CD`, sign descent-`SD`) to confirm the validity and measure the robustness of the theoretical claims, and to discuss the connection between `Adam` and steepest descent algorithms. Amongst other quantities, we measure the three relevant geometric margins during training, which, in the context of one-hidden layer neural networks with homogeneous activations and without biases, become:

$$\gamma_1 = \min_{i\in[m]} \frac{y_i f(\mathbf{x}_i; \boldsymbol{\theta})}{\|\boldsymbol{\theta}\|_\infty^2}, \quad \gamma_2 = \min_{i\in[m]} \frac{y_i f(\mathbf{x}_i; \boldsymbol{\theta})}{\|\boldsymbol{\theta}\|_2^2}, \quad \gamma_\infty = \min_{i\in[m]} \frac{y_i f(\mathbf{x}_i; \boldsymbol{\theta})}{\|\boldsymbol{\theta}\|_1^2}. \tag{14}$$

4.1 Teacher – Student Experiments

We first perform experiments in a controlled environment, where the generative process consists of Gaussian data passed through a one-hidden layer ("teacher") neural network, which is sparse. Specifically:

$$\mathbf{x} \sim \mathcal{N}(0, I_d), \quad y = \operatorname{sgn}\left(f_{\text{teacher}}(\mathbf{x}; \boldsymbol{\theta}^\star)\right) = \operatorname{sgn}\left(\sum_{j=1}^{k} u_j^\star \sigma\left(\langle \mathbf{w}_j^\star, \mathbf{x}\rangle\right)\right), \tag{15}$$

where $\sigma(u) = \max(u, 0)$ is the ReLU activation, $\operatorname{sgn}(\cdot)$ returns the sign of a number, and $\|\boldsymbol{\theta}^\star\|_0$ is assumed to be small. We train ("student") neural networks of the same architecture, but of larger width and with randomly initialized weights: $f_{\text{student}}(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^{k'} u_j \sigma\left(\langle \mathbf{w}_j, \mathbf{x}\rangle\right)$, with width $k' > k$ and $w_{jl} \sim \mathcal{U}\left[-\frac{\alpha}{d}, \frac{\alpha}{d}\right], j \in [k'], l \in [d], u_j \in \mathcal{U}\left[-\frac{\alpha}{k'}, \frac{\alpha}{k'}\right]$ (for `CD` we use: $w_{jl} \sim \mathcal{U}\left[-\frac{\alpha}{k'}, \frac{\alpha}{k'}\right], j \in [k'], l \in [d]$ in order to keep all the individual parameters to the same scale at initialization). The magnitude of initialization α can control how fast the implicit bias of

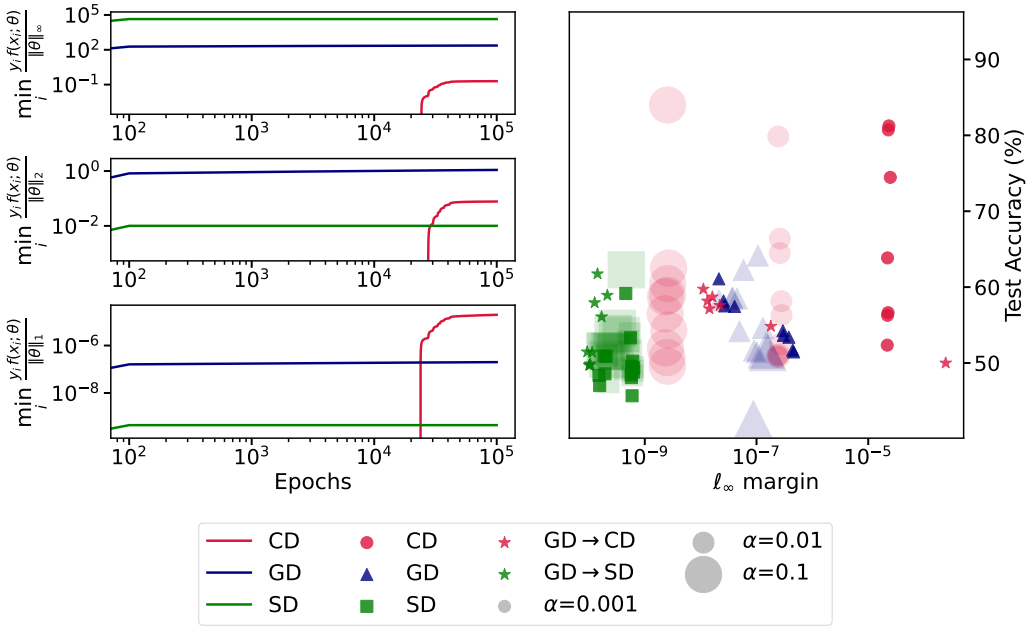

Figure 1: **Evaluation of steepest descent algorithms in a teacher-student setup.** *Left:* Geometric margins ($\gamma_1, \gamma_2, \gamma_\infty$ in Equation 14) over the course of training (average over 20 different seeds). *Right:* Final test accuracy vs final ℓ_∞ margin (γ_∞). Each point in the 2d space corresponds to a different run (only showing runs that did not diverge). Larger points correspond to larger initialization scales α. The star points are produced by switching from GD to CD (red) or SD (green), right after the point of perfect train accuracy.

the algorithm kicks in, with smaller values entering this so-called "rich" regime faster (Woodworth et al., 2020). We compare the performance of (full batch) GD, CD and SD in minimizing the empirical exponential loss, consisting of m independent points sampled from the generative process of Equation 15. Section D contains full experimental details. According to Theorems 3.1, 3.8, we expect GD to favor solutions with small ℓ_2 norm. This is equivalent to a small sum of the product of the magnitude of incoming and outcoming weights across all neurons (Theorem 1 in (Neyshabur et al., 2015b)). On the other hand, CD will seek to minimize the ℓ_1 norm of the parameters, which translates to a narrow network with sparse 1st-layer weights. Finally, SD's bias towards small $\|\boldsymbol{\theta}\|_\infty$ solutions does not appear to be useful for generalizing from few samples in this task. Therefore, we expect CD $>$ GD $>$ SD in terms of generalization.

Figure 1 displays our main findings. We summarize our key findings below:

(i) **Margins increase past t_0.** As expected from Lemma 3.1, we observe that, right after the point of separation, each algorithm implicitly increases its corresponding geometric margin (Figure 1 left). Furthermore, we observe that the ordering of the algorithms is as expected for each margin (SD attains larger ℓ_1 margin than GD and CD, etc.), despite the fact that Theorem 3.8 only guarantees convergence to a KKT point (at best) of the margin maximization problem - note the log-log plot.

(ii) **Smaller initialization produces larger geometric margin**. A smaller magnitude of initialization α causes a larger eventual value of the geometric margin (see Figure 1 right for CD and γ_∞, where this effect is stronger, and Figure 4 in Appendix D for γ_1, γ_2).

(iii) **Importance of early-stage dynamics for generalization**. Figure 1, right, shows the final test accuracy of the networks (20 different runs) vs the value of their final ℓ_∞ margin (γ_∞). We observe that, while there exist more CD runs with good generalization (red circles), these do not always coincide with larger γ_∞. Furthermore, intervening in the algorithms to encourage or discourage γ_∞-maximization does not result in significant generalization changes: after running GD until the point of perfect train accuracy, we switch to either SD (green stars) or CD (red stars) to directly control the late stage geometric properties of

the model. Switching to CD seems to bear marginal benefits in terms of generalization, even though all the switched runs reach smaller values of ℓ_∞ margin compared to the full, no-switching, GD runs. These benefits, however, pale in comparison to the full CD runs. Switching to SD, on the other hand, results in smaller γ_∞ and similar or marginally worse test accuracy. See also Figure 4 in Appendix D for test accuracy vs the other two geometric margins. We conclude that it is unlikely that large generalization benefits can solely and causally be linked to larger geometric margins in this setup, and it appears that the early stage dynamics play an important role for generalization.

4.2 Connection between Adam and Sign-GD

Adaptive optimization methods like Adam (Kingma & Ba, 2015) have been popular in deep learning applications, yet theoretically their value has been questioned (Wilson et al., 2017) and their properties remain poorly understood. Wang et al. (2021, 2022) studied the implicit bias of Adam in homogeneous networks and concluded that Adam shares the same asymptotic properties as GD. More recently, this conclusion has been challenged, in the sense that this asymptotic property crucially depends on a precision parameter of the algorithm and does not capture realistic runs of the algorithm (see Section C.1 for details). In particular, it was shown that in linear models, Adam, without this precision parameter, implicitly maximizes the ℓ_1-geometric margin (Zhang et al., 2024), a property shared with SD and not GD. Indeed, it is easy to see that Adam without momentum, and ignoring the precision parameters, is equivalent to SD (see Section C.1). Setting the precision parameter to 0, on the other hand, is not useful in applications, as small initial values of the gradient result in divergence of the loss. A question arises: what, then, are the relevant geometric properties of Adam *in practice*?

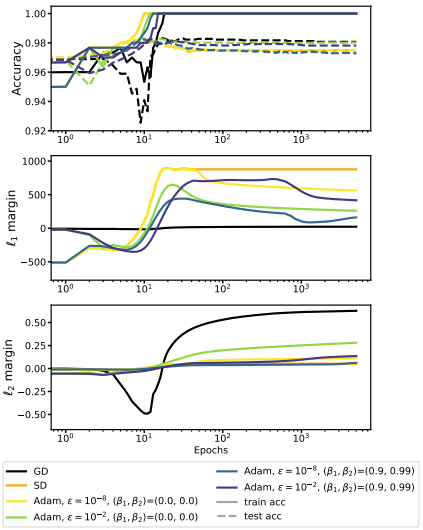

Figure 2: **Accuracy, and ℓ_1, ℓ_2 margins during training for GD, SD and Adam on MNIST (3 random seeds)**. Adam is parameterized by a numerical precision constant ϵ and two momentum parameters (β_1, β_2) (defaulting to 10^{-8} and $(0.9, 0.99)$). We observe that Adam behaves similar to SD for the period right after the point of perfect train accuracy.

Figure 2 provides some experimental answers to this question, in light of Theorems 3.1, 3.8. On a pair of digits extracted from MNIST we train two-layer neural networks with GD, SD and Adam, with small initialization. See Section D for experimental details. We observe that, as soon as the algorithms reach 100% train accuracy, the margins start to increase (as Theorem 3.1 suggests); SD reaches a larger value of γ_1, while GD reaches a larger value of γ_2. Interestingly, Adam with the default hyperparameters (precision $\epsilon = 10^{-8}$ and non-zero momentum), initially, behaves similar to SD, increasing γ_1, before it starts decreasing it, in order to slowly start increasing γ_2! Curiously, larger values of ϵ increase γ_1 even further and start the second phase slower, but more aggressively. Notice, however, that train and test accuracies have long converged, so it is unlikely that a typical run would have lasted long enough to see the second phase of ℓ_2-margin maximization (in particular, the loss value needs to be smaller than 10^{-7} in order to observe such behavior). Similar observations hold for Adam without momentum (recall that without momentum and for $\epsilon \to 0$, we recover SD). Therefore, it appears that the ℓ_1 bias of SD (Theorems 3.1, 3.8 for $\|\cdot\| = \|\cdot\|_\infty$) more faithfully describes a typical run of Adam in neural networks.

5 Conclusion

In our work, we considered the large family of steepest descent algorithms with respect to an arbitrary norm $\|\cdot\|$ and provided a unifying theoretical analysis of their late-stage implicit bias when training homogeneous neural networks. Furthermore, we introduced a notion of stationarity for optimization problems (defined with respect to a Bregman divergence induced by the algorithm norm),

which, as we showed, captures the implicit progress of the algorithms and might be of independent interest. Theorem 3.8 does not preclude the possibility that any steepest descent algorithm will converge to a KKT point; and our positive result (Corollary 3.8.1) shows that this is the case for smooth squared algorithm norm. It would be interesting and theoretically challenging to generalize this result to any norm or show a counterexample.

Our results can reinforce several recent efforts that attempt to understand deep learning through the lens of implicit bias. In particular, questions about generalization, robustness, and privacy can now be asked more broadly: (a) is it possible to extract training samples from neural networks optimized with Adam, leveraging its connection to sign gradient descent, similarly to what has been shown to be possible for gradient descent (Haim et al., 2022)? (b) can we leverage our implicit bias results to design more sample-efficient algorithms for robust training, as argued by Tsilivis et al. (2024)? (c) is benign overfitting a general property of first-order methods, or are current results (e.g. Frei et al. (2022); Shamir (2023)) specifically tailored to gradient descent?

Acknowledgments. NT and JK acknowledge support through the NSF under award 1922658. GV is supported by research grants from the Center for New Scientists at the Weizmann Institute of Science, and the Shimon and Golde Picker – Weizmann Annual Grant. We would like to thank Zhiyuan Li and Kaifeng Lyu for a useful discussion. Part of this work was done while NT was visiting the Toyota Technological Institute of Chicago (TTIC) during the winter of 2024, and NT would like to thank everyone at TTIC for their hospitality, which enabled this work. Part of this work was done while JK and NT were hosted by the Centre Sciences de Donnees at the École Normale Supérieure (ENS) in 2023/24, whose hospitality we gratefully acknowledge. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

A. Nemirovskii and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman Divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology. abs/2409.20325, 2024.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014. ISBN 978-0-521-83378-3.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553.

David E. Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned Spectral Descent for Deep Learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2971–2979, 2015.

Frank H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

Damek Davis, Dmitriy Drusvyatskiy, Sham M. Kakade, and Jason D. Lee. Stochastic Subgradient Method Converges on Tame Functions. *Found. Comput. Math.*, 20(1):119–154, 2020.

Joydeep Dutta, Kalyanmoy Deb, Rupesh Tulshyan, and Ramnik Arora. Approximate KKT points and a proximity measure for termination. *J. Glob. Optim.*, 56(4), 2013.

W. Fenchel. On Conjugate Convex Functions. *Canadian Journal of Mathematics*, 1(1):73–77, 1949.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2668–2703. PMLR, 2022.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing Implicit Bias in Terms of Optimization Geometry. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1827–1836. PMLR, 2018.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned Stochastic Tensor Optimization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1837–1845. PMLR, 2018.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing Training Data From Trained Neural Networks. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pp. 793–800. Curran Associates, Inc., 2008.

William Karush. Minima of Functions of Several Variables with Inequalities as Side Conditions. Master's thesis, Department of Mathematics, University of Chicago, 1939.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kuhn, H. W. and Tucker, A. W. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pp. 481–492. University of California Press, 1951.

Daniel Kunin, Atsushi Yamamura, Chao Ma, and Surya Ganguli. The Asymmetric Maximum Margin Bias of Quasi-Homogeneous Neural Networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable Optimization in the Modular Norm. abs/2405.14813, 2024. URL `https://doi.org/10.48550/arXiv.2405.14813`.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Kaifeng Lyu and Jian Li. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting Algorithms as Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Mor Shpigel Nacson, Suriya Gunasekar, Jason D. Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4683–4692. PMLR, 2019.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-Normalized Optimization in Deep Neural Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2422–2430, 2015a.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015b.

Andrew Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 78, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385.

Albert B. J. Novikoff. On Convergence Proofs For Perceptrons. 1963.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS 2017 Workshop on Autodiff*, 2017.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.

Shai Shalev-Shwartz. *Online learning: theory, algorithms and applications*. PhD thesis, Hebrew University of Jerusalem, Israel, 2007.

Ohad Shamir. The Implicit Bias of Benign Overfitting. *J. Mach. Learn. Res.*, 24:113:1–113:40, 2023. URL `https://jmlr.org/papers/v24/22-0784.html`.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. *J. Mach. Learn. Res.*, 19:70:1–70:57, 2018.

Matus Telgarsky. Margins, Shrinkage, and Boosting. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 307–315, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Nikolaos Tsilivis, Natalie Frank, Nathan Srebro, and Julia Kempe. The Price of Implicit Bias in Adversarially Robust Generalization. abs/2406.04981, 2024. URL `https://arxiv.org/abs/2406.04981`.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

Gal Vardi. On the Implicit Bias in Deep-Learning Algorithms. *Commun. ACM*, 66(6):86–93, 2023.

Gal Vardi, Ohad Shamir, and Nati Srebro. On Margin Maximization in Linear and ReLU Networks. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The Implicit Bias for Adaptive Optimization Algorithms on Homogeneous Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10849–10858. PMLR, 2021.

Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does Momentum Change the Implicit Regularization on Separable Data? In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4148–4158, 2017.

Blake E. Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 2020.

Shuo Xie and Zhiyuan Li. Implicit Bias of AdamW: ℓ_∞-Norm Constrained Optimization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Chenyang Zhang, Difan Zou, and Yuan Cao. The Implicit Bias of Adam on Separable Data. abs/2406.10650, 2024.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are Adaptive Methods Good for Attention Models? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

A Missing Proofs

In this section, we provide proofs for the results stated in the main text.

A.1 Steepest Flow

We first present a series of technical results, which are about steepest flow in the case of non-differentiable loss functions. In what follows, we will denote with $\mathbf{g}_t^\star$ any loss subderivative with minimum $\|\cdot\|_\star$ norm, i.e. $\mathbf{g}_t^\star \in \arg\min_{\mathbf{u}\in\partial\mathcal{L}(\boldsymbol{\theta}_t)} \|\mathbf{u}\|_\star$. In the case of subdifferentials, chain rule holds as an inclusion:

Theorem A.1 (Theorem 2.3.9 and 2.3.10 in Clarke (1990))**.** *Let $z_1, \ldots, z_n : \mathbb{R}^d \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz functions and define $\mathbf{z} = (z_1, \ldots, z_n)$. Let $(f \circ \mathbf{z})(\mathbf{x}) = f(z_1(\mathbf{x}), \ldots, z_n(\mathbf{x}))$ be the composition of $\mathbf{z}$ with f. Then, it holds:*

$$\partial(f \circ \mathbf{z})(\mathbf{x}) \subseteq \operatorname{conv}\left\{\sum_{i=1}^{n} \alpha_i \mathbf{h}_i : \boldsymbol{\alpha} \in \partial f(z_1(\mathbf{x}), \ldots, z_n(\mathbf{x})), \mathbf{h}_i \in \partial z_i(\mathbf{x})\right\}. \tag{16}$$

To further analyze steepest flows and to guarantee loss monotonicity, we need a stronger chain rule result. This can be achieved for a large class of locally Lipschitz functions, as per the following theorem which is due to Davis et al. (2020).

Theorem A.2. *(Theorem 5.8 in Davis et al. (2020)) If $F : \mathbb{R}^k \to \mathbb{R}$ is locally Lipschitz and Whitney C^1-stratifiable, then it admits a chain rule: for all arcs (functions which are absolutely continuous on every compact subinterval) $\mathbf{u} : [0, \infty) \to \mathbb{R}^k$, almost all $t \geq 0$, and all $\mathbf{g} \in \partial F(\mathbf{u}(t))$, it holds:*

$$\frac{dF(\mathbf{u}(t))}{dt} = \left\langle \mathbf{g}, \frac{d\mathbf{u}(t)}{dt} \right\rangle. \tag{17}$$

Whitney C^1-stratifiability includes a large family of functions, including functions defined in an o-minimal structure which has been a standard assumption in the literature Ji & Telgarsky (2020). It excludes some pathological functions - see, for instance, Appendix J in Lyu & Li (2020). This version of chain rule allows us to derive the following central properties of steepest flows.

Proposition A.3. *Let $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$ and assume that $\mathcal{L}$ admits a chain rule. Then, for the steepest flow iterates of Equation 1, it holds for almost any $t \geq 0$:*

$$\frac{d\mathcal{L}}{dt} = -\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2 \leq 0, \tag{18}$$

and

$$\left\langle \frac{d\boldsymbol{\theta}}{dt}, -\mathbf{g}_t^\star \right\rangle = \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2 = \|\mathbf{g}_t^\star\|_\star^2, \tag{19}$$

where $\mathbf{g}_t^\star \in \arg\min_{\mathbf{u}\in\partial\mathcal{L}(\boldsymbol{\theta}_t)} \|\mathbf{u}\|_\star$.

Proof. From Theorem A.2, for almost any $t \geq 0$, it holds $\forall\ \mathbf{g}_t \in \partial\mathcal{L}(\boldsymbol{\theta}_t)$:

$$\frac{d\mathcal{L}}{dt} = \left\langle \mathbf{g}_t, \frac{d\boldsymbol{\theta}}{dt} \right\rangle. \tag{20}$$

Applying this for the element of $\partial\mathcal{L}(\boldsymbol{\theta}_t), \mathbf{g}_t'$, that corresponds to $\frac{d\boldsymbol{\theta}}{dt}$ from the definition of steepest flow Equation 3, we get:

$$\frac{d\mathcal{L}}{dt} = \left\langle \mathbf{g}_t', \frac{d\boldsymbol{\theta}}{dt} \right\rangle = -\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2, \tag{21}$$

where the last equality follows from the definition of the dual norm. But, Equation 20 for $\mathbf{g}_t^\star \in \arg\min_{\mathbf{u}\in\partial\mathcal{L}(\boldsymbol{\theta}_t)} \|\mathbf{u}\|_\star$, yields:

$$\left|\frac{d\mathcal{L}}{dt}\right| = \left|\left\langle \mathbf{g}_t^\star, \frac{d\boldsymbol{\theta}}{dt} \right\rangle\right| \leq \|\mathbf{g}_t^\star\|_\star \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|. \tag{22}$$

Thus, combining Equation 21, Equation 22, we obtain:

$$\left\| \frac{d\boldsymbol{\theta}}{dt} \right\| \leq \|\mathbf{g}_t^\star\|_\star, \tag{23}$$

which implies that the update rule Equation 3 iq equivalent to:

$$\frac{d\boldsymbol{\theta}}{dt} \in \left\{ \underset{\|\mathbf{u}\| \leq \|\mathbf{g}_t\|_\star}{\arg\min} \langle \mathbf{u}, \mathbf{g}_t \rangle : \mathbf{g}_t^\star \in \underset{\mathbf{u} \in \partial \mathcal{L}(\boldsymbol{\theta}_t)}{\arg\min} \|\mathbf{u}\|_\star \right\}. \tag{24}$$

Therefore, from the definition of the dual norm, we have:

$$\left\langle \frac{d\boldsymbol{\theta}}{dt}, -\mathbf{g}_t^\star \right\rangle = \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2 = \|\mathbf{g}_t^\star\|_\star^2. \tag{25}$$

□

Hence, under the mild assumptions of Theorem A.2, the loss is non-increasing during training.

A.2 Late Phase Implicit Bias

A useful standard characterization of the subdifferential of a norm is the following:

Lemma A.4.

$$\partial \|\mathbf{x}\| = \{\mathbf{v} : \langle \mathbf{v}, \mathbf{x} \rangle = \|\mathbf{x}\|, \|\mathbf{v}\|_\star \leq 1\}$$

We present the proofs for our results about the late stage of training in steepest flow algorithms. The next lemma quantifies the behavior of the smooth margin past the point t_0 (where, recall, zero classification error is achieved).

Theorem A.5 (Soft margin increases - full version)**.** *For almost any $t > t_0$, it holds:*

$$\frac{d \log \widetilde{\gamma}}{dt} \geq L \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2 \left(\frac{1}{L\mathcal{L}(\boldsymbol{\theta}_t) \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}} - \frac{1}{\|\boldsymbol{\theta}_t\| \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|} \right) \geq 0.$$

Proof. Let $\mathbf{n}_t \in \partial \|\boldsymbol{\theta}_t\|$. We have:

$$\begin{aligned}
\frac{d \log \widetilde{\gamma}}{dt} &= \frac{d}{dt} \log \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)} - L \frac{d}{dt} \log \|\boldsymbol{\theta}_t\| \\
&= \frac{d}{dt} \log \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)} - L \left\langle \frac{\mathbf{n}_t}{\|\boldsymbol{\theta}_t\|}, \frac{d\boldsymbol{\theta}}{dt} \right\rangle \qquad \text{(Chain rule)} \\
&\geq \frac{d}{dt} \log \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)} - L \frac{\left\| \frac{d\boldsymbol{\theta}}{dt} \right\|}{\|\boldsymbol{\theta}_t\|} \qquad \text{(definition of dual norm and } \|\mathbf{n}_t\|_\star \leq 1) \\
&= -\frac{d\mathcal{L}(\boldsymbol{\theta}_t)}{dt} \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t) \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}} - L \frac{\left\| \frac{d\boldsymbol{\theta}}{dt} \right\|}{\|\boldsymbol{\theta}_t\|} \qquad \text{(Chain rule)} \\
&= \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2 \left(\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t) \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}} - \frac{L}{\|\boldsymbol{\theta}_t\| \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|} \right) \qquad \text{(Equation 18).}
\end{aligned} \tag{26}$$

But, the first term inside the parenthesis can be related to the second one via the following calculation. Recall that, by Theorem A.2, for any $\mathbf{g}_t \in \partial \mathcal{L}(\boldsymbol{\theta}_t)$ there exist $\mathbf{h}_1 \in \partial y_1 f(\mathbf{x}_1; \boldsymbol{\theta}_t), \ldots, \mathbf{h}_m \in \partial y_m f(\mathbf{x}_m; \boldsymbol{\theta}_t)$ such that $\mathbf{g}_t = \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i; \boldsymbol{\theta}_t)} \mathbf{h}_i$. Thus, for a minimum norm subderivative $\mathbf{g}_t^\star$, we have:

$$\begin{aligned}
\langle \boldsymbol{\theta}_t, -\mathbf{g}_t^\star \rangle &= \left\langle \boldsymbol{\theta}_t, \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i; \boldsymbol{\theta}_t)} \mathbf{h}_i^\star \right\rangle \\
&= \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i; \boldsymbol{\theta}_t)} \langle \boldsymbol{\theta}_t, \mathbf{h}_i^\star \rangle \\
&= L \sum_{i=1}^m e^{-y_i f(\mathbf{x}_i; \boldsymbol{\theta}_t)} y_i f(\mathbf{x}_i; \boldsymbol{\theta}_t),
\end{aligned} \tag{27}$$

where the last equality follows from Euler's theorem for homogeneous functions (whose generalization for subderivatives can be found in Theorem B.2 in Lyu & Li (2020)). Now, observe that this last term can be lower bounded as:

$$\langle \boldsymbol{\theta}_t, -\mathbf{g}_t^\star \rangle \geq L \sum_{i=1}^{m} e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} \min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t) \geq L\mathcal{L}(\boldsymbol{\theta}_t) \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}, \tag{28}$$

where we used the fact $e^{-\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} \leq \sum_{i=1}^{m} e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)}$. We have made the first term of Equation 26 appear. By plugging Equation 28 into Equation 26, we get:

$$\begin{aligned} \frac{d\log\widetilde{\gamma}}{dt} &\geq \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2 \left(\frac{L}{\langle \boldsymbol{\theta}_t, -\mathbf{g}_t^\star \rangle} - \frac{L}{\|\boldsymbol{\theta}_t\| \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|} \right) \\ &\geq \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2 \left(\frac{L}{\|\boldsymbol{\theta}_t\| \|\mathbf{g}_t^\star\|_\star} - \frac{L}{\|\boldsymbol{\theta}_t\| \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|} \right) \qquad \text{(definition of dual norm).} \end{aligned} \tag{29}$$

Noticing that $\|\mathbf{g}_t^\star\|_\star = \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|$ (from Proposition A.13) concludes the proof. □

By extending Lemma B.6 of Lyu & Li (2020), we can further prove that the loss converges to 0 and, thus, the norm of the iterates diverges to infinity.

Lemma A.6. *As $t \to \infty$, $\mathcal{L}(\boldsymbol{\theta}_t) \to 0$ and $\|\boldsymbol{\theta}_t\| \to \infty$.*

Proof. We suppress the dependence of the loss and the iterates from time t, when it is obvious from the context.

From the definition of the steepest flow update and chain rule (eq. Equation 18), we have

$$-\frac{d\mathcal{L}}{dt} = \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|^2 = \|\mathbf{g}_t^\star\|_\star^2 \geq \frac{1}{\|\boldsymbol{\theta}\|^2} \langle \boldsymbol{\theta}, -\mathbf{g}_t^\star \rangle^2, \tag{30}$$

where we applied Equation 19, Equation 18 and the definition of the dual norm. But, as we showed in Equation 28, the above inner product can be upper bounded by a function of the loss, so, by plugging in, we get:

$$-\frac{d\mathcal{L}}{dt} \geq \frac{L^2}{\|\boldsymbol{\theta}\|^2} \left(\mathcal{L} \log \frac{1}{\mathcal{L}} \right)^2 = \frac{L^2}{\left(\log \frac{1}{\mathcal{L}}\right)^{2/L}} \widetilde{\gamma}^{2/L}(t) \left(\mathcal{L} \log \frac{1}{\mathcal{L}} \right)^2 \geq \frac{L^2}{\left(\log \frac{1}{\mathcal{L}}\right)^{2/L}} \widetilde{\gamma}^{2/L}(t_0) \left(\mathcal{L} \log \frac{1}{\mathcal{L}} \right)^2, \tag{31}$$

which follows from the definition of the margin (Equation 5) and its monotonicity (Lemma A.5). By rearranging:

$$-\frac{d\mathcal{L}}{dt} \frac{1}{\mathcal{L}^2} \left(\log \frac{1}{\mathcal{L}} \right)^{2/L-2} \geq L^2 \widetilde{\gamma}(t_0)^{2/L}, \tag{32}$$

and integrating over time from t_0 to $t > t_0$, we further have:

$$\int_{t_0}^{t} \left(\log \frac{1}{\mathcal{L}} \right)^{2/L-2} \frac{d}{dt} \frac{1}{\mathcal{L}} dt \geq L^2 \widetilde{\gamma}(t_0)^{2/L} (t - t_0), \tag{33}$$

or, by a change of variables,

$$\int_{1/\mathcal{L}(t_0)}^{1/\mathcal{L}(t)} (\log u)^{2/L-2} \, du \geq L^2 \widetilde{\gamma}(t_0)^{2/L} (t - t_0). \tag{34}$$

The RHS diverges to infinity as $t \to \infty$, hence so does the LHS, which can only happen if $\mathcal{L} \to 0$. In order for $\mathcal{L}(\boldsymbol{\theta}_t) = \sum_{i=1}^{m} e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} = \sum_{i=1}^{m} e^{-y_i \|\boldsymbol{\theta}_t\|^L f\left(\mathbf{x}_i; \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|}\right)}$ to go to zero, it must be $\|\boldsymbol{\theta}_t\| \to \infty$. □

The following Lemma quantifies the connection between soft and hard margin.

Lemma A.7. *For any $\boldsymbol{\theta}$, it holds:*

$$\frac{\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|^L} - \log m \leq \widetilde{\gamma} \leq \frac{\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|^L}. \tag{35}$$

Proof. Follows from:

$$e^{-\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta})} \leq \mathcal{L}(\boldsymbol{\theta}) \leq m e^{-\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta})}. \tag{36}$$

□

The last part of the proof consists of characterizing the (directional) convergence of the iterates in relation to stationary points of the following optimization problem (re-introduced here for convenience):

$$\begin{aligned} &\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ &\text{s.t. } y_i f(\mathbf{x}_i;\boldsymbol{\theta}) \geq 1,\ \forall i \in [m]. \end{aligned} \tag{37}$$

Under some regularity assumptions, the KKT conditions (Definition 3.3) become necessary for global optimality (yet, not sufficient):

Definition A.8. *We say that a feasible point of Equation 37 satisfies the Mangasarian-Fromovitz Constraint Qualifications if there exists* $\mathbf{v} \in \mathbb{R}^p$ *such that for all* $i \in [m]$ *with* $1 - y_i f(\mathbf{x}_i;\boldsymbol{\theta}) = 0$ *and for all* $\mathbf{h} \in \partial\left(1 - y_i f(\mathbf{x}_i;\boldsymbol{\theta})\right)$*, it holds:*

$$\langle \mathbf{v}, \mathbf{h} \rangle > 0. \tag{38}$$

Our proof uses the following relaxed notion of stationarity.

Definition A.9 ((d,ϵ,δ)- approximate KKT point)**.** *Let* $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$*. A feasible point* $\boldsymbol{\theta}$ *of equation 37 is called an* $(d,\epsilon,\delta)-$*approximate KKT point if there exist* $\lambda_1,\ldots,\lambda_m \geq 0$*,* $\mathbf{h}_i \in \partial f(\mathbf{x}_i;\boldsymbol{\theta})$ *and* $\mathbf{k} \in \partial\frac{1}{2}\|\boldsymbol{\theta}\|^2$ *such that:*

1. $d\left(\sum_{i=1}^m \lambda_i y_i \mathbf{h}_i, \mathbf{k}\right) \leq \epsilon$
2. $\sum_{i=1}^m \lambda_i (y_i f(\mathbf{x}_i;\boldsymbol{\theta}) - 1) \leq \delta$.

We first show that we can always construct a feasible point of Equation 37 from a scaled version of $\boldsymbol{\theta}_t$.

Lemma A.10. *For any* $t > 0$*,* $\widetilde{\boldsymbol{\theta}}_t = \frac{\boldsymbol{\theta}_t}{\left(\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)\right)^{\frac{1}{L}}}$ *is a feasible point of Equation 37.*

Proof. From the homogeneity of f, we have:

$$y_i f(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_t) = y_i f\left(\mathbf{x}_i; \frac{\boldsymbol{\theta}_t}{\left(\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)\right)^{\frac{1}{L}}}\right) = \frac{y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)}{\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)} \geq 1 \tag{39}$$

for all $i \in [m]$. So $\widetilde{\boldsymbol{\theta}}_t$ is a feasible point of Equation 37. □

The next Lemma shows that Problem 37 satisfies the Mangasarian-Fromovitz Constraint Qualifications:

Lemma A.11. *Problem 37 satisfies the Mangasarian-Fromovitz Constraint Qualifications at every feasible point* $\boldsymbol{\theta}$*.*

Proof. Let $\mathbf{h}_i \in \partial(1 - y_i f(\mathbf{x}_i;\boldsymbol{\theta}))$ and $\mathbf{v} = -\boldsymbol{\theta}$, then for all $i \in [m]$ satisfying $y_i f(\mathbf{x}_i;\boldsymbol{\theta}) = 1$, we have from Euler's theorem for homogeneous functions:

$$\langle \mathbf{v}, \mathbf{h}_i \rangle = L y_i f(\mathbf{x}_i;\boldsymbol{\theta}) = L > 0. \tag{40}$$

□

Our proof uses core ideas from the theory of conjugate functions and Fenchel's duality.

Definition A.12 (Convex conjugate)**.** *Let* $\psi : \mathbb{R}^p \to \mathbb{R}$*. We denote by* $\psi^\star(\cdot)$ *the convex conjugate of* $\psi(\cdot)$*:*

$$\psi^\star(\boldsymbol{\omega}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^p} \{\langle \boldsymbol{\omega}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})\}. \tag{41}$$

We will make use of the following properties of conjugate functions.

Proposition A.13. *(Conjugate subgradient theorem - Theorem 23.5 in Rockafellar (1970), Theorem 4.20 in Beck (2017)) Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be convex and closed. For any $\boldsymbol{\theta}^\star \in \partial\psi^\star(\boldsymbol{\theta})$, it holds $\partial\psi(\boldsymbol{\theta}^\star) \ni \boldsymbol{\theta}$.*

Lemma A.14. *(Fenchel-Young inequality) (Fenchel, 1949) For any $\psi : \mathbb{R}^p \to \mathbb{R}$ and $\boldsymbol{\omega}, \boldsymbol{\theta} \in \mathbb{R}^p$, it holds:*

$$\langle \boldsymbol{\theta}, \boldsymbol{\omega} \rangle \leq \psi(\boldsymbol{\theta}) + \psi^\star(\boldsymbol{\omega}). \tag{42}$$

Next, we show that the scaled version of the iterates from Lemma A.10, $\widetilde{\boldsymbol{\theta}}_t$, is an $\left(D^{\widetilde{\boldsymbol{\theta}}_t}_{\frac{1}{2}\|\cdot\|_\star^2}, \epsilon(t), \delta(t)\right)$-approximate KKT point for $\epsilon(t)$ and $\delta(t)$ that vanish as t increases.

Proposition A.15. *For any $t > t_0$, $\widetilde{\boldsymbol{\theta}}_t = \frac{\boldsymbol{\theta}_t}{\left(\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)\right)^{\frac{1}{L}}}$ is an $\left(D^{\widetilde{\boldsymbol{\theta}}_t}_{\frac{1}{2}\|\cdot\|_\star^2}, \epsilon(t), \delta(t)\right)$-approximate KKT point of Equation 37, with:*

$$\begin{aligned} \epsilon(t) &= \frac{1}{\widetilde{\gamma}(t_0)^{\frac{2}{L}}}\left(1 - \left\langle \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star}\right\rangle\right), \\ \delta(t) &= \frac{m}{eL\widetilde{\gamma}(t_0)^{\frac{2}{L}}\log\frac{1}{\mathcal{L}}}, \end{aligned} \tag{43}$$

with $\mathbf{g}_t^\star \in \arg\min_{\mathbf{u}\in\partial\mathcal{L}(\boldsymbol{\theta}_t)} \|\mathbf{u}\|_\star$.

Proof. We suppress the dependence of the loss and the iterates from the time index t, when it is obvious from the context. From Lemma A.10, we know that $\widetilde{\boldsymbol{\theta}}$ is a feasible point. To simplify the notation, let $q_{\min} = \min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta})$. We will denote by $\widetilde{\mathbf{k}} \in \partial\frac{1}{2}\|\widetilde{\boldsymbol{\theta}}\|^2$ any subgradient of $\frac{1}{2}\|\cdot\|^2$ at $\widetilde{\boldsymbol{\theta}}$. Let, as previously stated, $\mathbf{g}_t^\star \in \arg\min_{\mathbf{u}\in\partial\mathcal{L}(\boldsymbol{\theta}_t)} \|\mathbf{u}\|_\star$ and $\mathbf{h}_i^\star \in \partial f(\mathbf{x}_i;\boldsymbol{\theta}), i \in [m]$, such that $\mathbf{g}_t^\star = -\sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta})} y_i \mathbf{h}_i^\star$ (whose existence is guaranteed from chain rule - Theorem A.1). Finally, we define $\widetilde{\mathbf{h}}_i^\star = q_{\min}^{\frac{1}{L}-1}\mathbf{h}_i^\star$ for all $i \in [m]$, for which it holds: $\widetilde{\mathbf{h}}_i^\star \in \partial f(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}})$ from Theorem B.2(a) in Lyu & Li (2020).

Given all these definitions, we set $\lambda_i = \frac{\|\boldsymbol{\theta}\|}{\|\mathbf{g}_t^\star\|_\star} q_{\min}^{1-\frac{2}{L}} e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta})} \geq 0$. The dual vector from the (d, ϵ, δ)-stationarity definition can be simplified to:

$$\begin{aligned} \sum_{i=1}^m \lambda_i y_i \widetilde{\mathbf{h}}_i^\star &= \sum_{i=1}^m \lambda_i q_{\min}^{\frac{1}{L}-1} y_i \mathbf{h}_i^\star \qquad \text{(Thm B.2(a) in Lyu \& Li (2020))} \\ &= \frac{\|\boldsymbol{\theta}\|}{q_{\min}^{\frac{1}{L}}\|\mathbf{g}_t^\star\|_\star}\sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta})} y_i \mathbf{h}_i \\ &= -\frac{\|\boldsymbol{\theta}\|\mathbf{g}_t^\star}{q_{\min}^{\frac{1}{L}}\|\mathbf{g}_t^\star\|_\star}, \end{aligned} \tag{44}$$

which is a scaled version of the (minimum norm) subderivative of the loss.

Let $\psi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_\star^2$ be the potential function that we shall use in order to define our divergence. For this specific ψ, it holds: $\psi^\star(\boldsymbol{\omega}) = \frac{1}{2}\|\boldsymbol{\omega}\|^2$ (see for instance Example 3.27 in Boyd & Vandenberghe (2014) for a derivation). Recall that in the definition of $D^{\mathbf{m}}_{\frac{1}{2}\|\cdot\|_\star^2}$ (Equation 12) there is an extra choice that we have to make; the one of the subderivative $\mathbf{m}$. In what follows, we will specifically measure "distance" between $\sum_{i=1}^m \lambda_i y_i \widetilde{\mathbf{h}}_i$ and $\widetilde{\mathbf{k}}$ using $D^{\widetilde{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}(\cdot,\cdot)$, i.e. by picking $\mathbf{m} = \widetilde{\boldsymbol{\theta}}$. This is possible, since from Proposition A.13 it holds that $\widetilde{\boldsymbol{\theta}} \in \partial\frac{1}{2}\|\widetilde{\mathbf{k}}\|_\star^2$. Finally, let $\mathbf{r} \in \partial\|\widetilde{\boldsymbol{\theta}}\|$ be the subgradient of $\|\cdot\|$ that stems from the chain rule of $\frac{1}{2}\|\cdot\|^2$ evaluated at $\widetilde{\boldsymbol{\theta}}$. We calculate the divergence between the

two vectors:

$$
\begin{aligned}
&D^{\widetilde{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}\left(\sum_{i=1}^m \lambda_i y_i \widetilde{\mathbf{h}}_i, \widetilde{\mathbf{k}}\right) \\
&\quad= \frac{1}{2}\left\|-\frac{\|\boldsymbol{\theta}\|\mathbf{g}_t^\star}{q_{\min}^{\frac{1}{L}}\|\mathbf{g}_t^\star\|_\star}\right\|_\star^2 - \frac{1}{2}\left\|\widetilde{\mathbf{k}}\right\|_\star^2 - \left\langle \frac{\boldsymbol{\theta}}{q_{\min}^{\frac{1}{L}}}, -\frac{\|\boldsymbol{\theta}\|\mathbf{g}_t^\star}{q_{\min}^{\frac{1}{L}}\|\mathbf{g}_t^\star\|_\star} - \widetilde{\mathbf{k}}\right\rangle \\
&\quad= \frac{1}{2}\frac{\|\boldsymbol{\theta}\|^2}{q_{\min}^{\frac{2}{L}}} - \frac{1}{2}\left\|\|\widetilde{\boldsymbol{\theta}}\|\mathbf{r}\right\|_\star^2 - \left\langle \frac{\boldsymbol{\theta}}{q_{\min}^{\frac{1}{L}}}, \frac{-\|\boldsymbol{\theta}\|\mathbf{g}_t^\star}{q_{\min}^{\frac{1}{L}}\|\mathbf{g}_t^\star\|_\star} - \|\widetilde{\boldsymbol{\theta}}\|\mathbf{r}\right\rangle \qquad \text{(Chain rule)} \\
&\quad= \frac{\|\boldsymbol{\theta}\|^2}{q_{\min}^{\frac{2}{L}}}\left(\frac{1}{2} - \frac{1}{2}\|\mathbf{r}\|_\star^2 - \left\langle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star}\right\rangle + \left\langle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \mathbf{r}\right\rangle\right) \\
&\quad\le \frac{\|\boldsymbol{\theta}\|^2}{q_{\min}^{\frac{2}{L}}}\left(\frac{1}{2} - \frac{1}{2}\|\mathbf{r}\|_\star^2 - \left\langle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star}\right\rangle + \frac{1}{2}\left\|\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\right\|^2 + \frac{1}{2}\|\mathbf{r}\|_\star^2\right) \qquad \text{(Equation 42)} \\
&\quad= \frac{\|\boldsymbol{\theta}\|^2}{q_{\min}^{\frac{2}{L}}}\left(1 - \left\langle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star}\right\rangle\right) \\
&\quad\le \frac{1}{\widetilde{\gamma}^{\frac{2}{L}}}\left(1 - \left\langle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star}\right\rangle\right) \le \frac{1}{\widetilde{\gamma}(t_0)^{\frac{2}{L}}}\left(1 - \left\langle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star}\right\rangle\right),
\end{aligned}
\tag{45}
$$

where the last 2 inequalities follow from the relation between smooth and hard margin (Lemma A.7), and the monotonicity of the former. For the second condition of an approximate KKT point, we have:

$$
\begin{aligned}
\sum_{i=1}^m \lambda_i\left(y_i f(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}) - 1\right) &= \frac{\|\boldsymbol{\theta}\|}{\|\mathbf{g}_t^\star\|_\star}\sum_{i=1}^m q_{\min}^{1-\frac{2}{L}} e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta})}\left(\frac{y_i f(\mathbf{x}_i;\boldsymbol{\theta})}{q_{\min}} - 1\right) \\
&= \frac{\|\boldsymbol{\theta}\|}{q_{\min}^{\frac{2}{L}}\|\mathbf{g}_t^\star\|_\star}\sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta})}\left(y_i f(\mathbf{x}_i;\boldsymbol{\theta}) - q_{\min}\right).
\end{aligned}
\tag{46}
$$

From eq. Equation 30 and Equation 28, we can lower bound the dual norm of the subderivate:

$$
\|\mathbf{g}_t^\star\|_\star \ge \frac{L}{\|\boldsymbol{\theta}\|}\mathcal{L}\log\frac{1}{\mathcal{L}} \ge \frac{L}{\|\boldsymbol{\theta}\|}e^{-q_{\min}}\log\frac{1}{\mathcal{L}}. \tag{47}
$$

By plugging in back to eq. Equation 46, we obtain

$$
\begin{aligned}
\sum_{i=1}^m \lambda_i\left(y_i f(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}) - 1\right) &\le \frac{\|\boldsymbol{\theta}\|^2}{q_{\min}^{\frac{2}{L}} L e^{-q_{\min}}\log\frac{1}{\mathcal{L}}}\sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta})}\left(y_i f(\mathbf{x}_i;\boldsymbol{\theta}) - q_{\min}\right) \\
&= \frac{\|\boldsymbol{\theta}\|^2}{q_{\min}^{\frac{2}{L}} L\log\frac{1}{\mathcal{L}}}\sum_{i=1}^m e^{-(y_i f(\mathbf{x}_i;\boldsymbol{\theta}) - q_{\min})}\left(y_i f(\mathbf{x}_i;\boldsymbol{\theta}) - q_{\min}\right) \\
&\le \frac{1}{\widetilde{\gamma}(t_0)^{\frac{2}{L}} L\log\frac{1}{\mathcal{L}}}\sum_{i=1}^m e^{-(y_i f(\mathbf{x}_i;\boldsymbol{\theta}) - q_{\min})}\left(y_i f(\mathbf{x}_i;\boldsymbol{\theta}) - q_{\min}\right) \qquad \text{(Lemmata A.7, A.5)} \\
&\le \frac{m}{e\widetilde{\gamma}(t_0)^{\frac{2}{L}} L\log\frac{1}{\mathcal{L}}},
\end{aligned}
\tag{48}
$$

since the function $u \mapsto e^{-u}u, u > 0$ has a maximum value of e^{-1}. □

Before we proceed with the main result, we state and prove two useful Lemmata. The first one lower bounds the alignment between normalized iterates and normalized loss gradients. This Lemma is key for showing that the alignment goes to 1 as $t \to \infty$.

Lemma A.16. *For all $t_2 > t_1 \geq t_0$, there exists $t_\star \in (t_1, t_2)$ such that:*

$$\left(\frac{1}{\left\langle \frac{\boldsymbol{\theta}_{t_\star}}{\|\boldsymbol{\theta}_{t_\star}\|}, \frac{-\mathbf{g}^\star_{t_\star}}{\|\mathbf{g}^\star_{t_\star}\|_\star} \right\rangle} - 1 \right) \leq \frac{1}{L} \frac{\log \frac{\widetilde{\gamma}(t_2)}{\widetilde{\gamma}(t_1)}}{\int_{t_1}^{t_2} \frac{\left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|}{\|\boldsymbol{\theta}_t\|} dt}, \tag{49}$$

for all $\mathbf{g}^\star_{t_\star} \in \arg\min_{\mathbf{u} \in \partial \mathcal{L}(\boldsymbol{\theta}_{t_\star})} \|\mathbf{u}\|_\star$

Proof. From Lemma A.5, we have for all $\mathbf{g}^\star_t \in \arg\min_{\mathbf{u} \in \partial \mathcal{L}(\boldsymbol{\theta}_t)} \|\mathbf{u}\|_\star$:

$$\begin{aligned} \frac{d \log \widetilde{\gamma}}{dt} &\geq L \left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|^2 \left(\frac{1}{\langle \boldsymbol{\theta}_t, -\mathbf{g}^\star_t \rangle} - \frac{1}{\|\boldsymbol{\theta}_t\| \left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|} \right) \\ &= L \frac{\left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|}{\|\boldsymbol{\theta}_t\|} \left(\frac{1}{\left\langle \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|}, \frac{-\mathbf{g}^\star_t(\boldsymbol{\theta}_t)}{\|\mathbf{g}^\star_t\|_\star} \right\rangle} - 1 \right). \end{aligned} \tag{50}$$

We then integrate the two sides from t_1 to $t_2 > t_1 > t_0$:

$$\int_{t_1}^{t_2} \left(\frac{1}{\left\langle \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|}, \frac{-\mathbf{g}^\star_t}{\|\mathbf{g}^\star_t\|_\star} \right\rangle} - 1 \right) \frac{\left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|}{\|\boldsymbol{\theta}_t\|} dt \leq \frac{1}{L} \log \frac{\widetilde{\gamma}(t_2)}{\widetilde{\gamma}(t_1)}. \tag{51}$$

The desired existential statement follows from a proof by contradiction. □

Next, we bound the rate of change of the normalized iterates.

Lemma A.17. *For any $t > 0$, it holds:*

$$\left\| d \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|} \right\| \leq 2 \frac{\left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|}{\|\boldsymbol{\theta}_t\|}. \tag{52}$$

Proof. The rate of change of the normalized iterates can be written as follows:

$$\begin{aligned} d \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|} &= \frac{1}{\|\boldsymbol{\theta}_t\|} \frac{d\boldsymbol{\theta}_t}{dt} + \boldsymbol{\theta}_t \left(-\frac{1}{\|\boldsymbol{\theta}_t\|^2} \frac{d\|\boldsymbol{\theta}_t\|}{dt} \right) \\ &= \frac{1}{\|\boldsymbol{\theta}_t\|} \frac{d\boldsymbol{\theta}_t}{dt} + \boldsymbol{\theta}_t \left(-\frac{1}{\|\boldsymbol{\theta}_t\|^2} \left\langle \mathbf{n}_t, \frac{d\boldsymbol{\theta}_t}{dt} \right\rangle \right), \qquad \text{(Chain rule)} \end{aligned} \tag{53}$$

where $\mathbf{n}_t \in \partial \|\boldsymbol{\theta}_t\|$. So, by the triangle inequality, its norm is bounded by:

$$\begin{aligned} \left\| d \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|} \right\| &\leq \frac{\left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|}{\|\boldsymbol{\theta}_t\|} + \frac{1}{\|\boldsymbol{\theta}_t\|} \left| \left\langle \mathbf{n}_t, \frac{d\boldsymbol{\theta}_t}{dt} \right\rangle \right| \\ &\leq 2 \frac{\left\| \frac{d\boldsymbol{\theta}_t}{dt} \right\|}{\|\boldsymbol{\theta}_t\|}. \qquad \text{(definition of dual norm and } \|\mathbf{n}_t\|_\star \leq 1) \end{aligned} \tag{54}$$

□

We are, now, ready to state and prove our main result.

Theorem A.18. *For steepest flow (eq. Equation 3) on the exponential loss, under assumptions A1, A2, A3, any limit point $\bar{\boldsymbol{\theta}}$ of $\left\{ \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|} \right\}_{t \geq 0}$ is along the direction of a $D^{\widetilde{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|^2_\star}$-generalized KKT point, $\widetilde{\boldsymbol{\theta}} := \frac{\bar{\boldsymbol{\theta}}}{\left(\min_{i \in [m]} y_i f(\mathbf{x}_i; \bar{\boldsymbol{\theta}})\right)^{\frac{1}{L}}}$, of the following optimization problem:*

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2} \|\boldsymbol{\theta}\|^2 \\ &\text{s.t. } y_i f(\mathbf{x}_i; \boldsymbol{\theta}) \geq 1, \ \forall i \in [m]. \end{aligned} \tag{55}$$

Proof. Our strategy will be to consider any limit point $\bar{\boldsymbol{\theta}}$ and construct $\left(D^{\widetilde{\boldsymbol{\theta}}_t}_{\frac{1}{2}\|\cdot\|_\star^2}, \epsilon(t), \delta(t)\right)$-approximate KKT points that converge to it, with vanishing $\epsilon(t), \delta(t)$.

Let $\epsilon_m = \frac{1}{m}$ for any $m > 0$. We construct a sequence $\{t_m\}_{m\geq 0}$, by induction, in the following sense. Suppose $t_1 < \ldots < t_{m-1}$ have been constructed already. Since $\bar{\boldsymbol{\theta}}$ is a limit point of the normalized iterates and $\log \widetilde{\gamma}_t \to \log \widetilde{\gamma}_\infty < \infty$ (as $\widetilde{\gamma}_t$ is non-decreasing and bounded from above), there exists $s_m > t_{m-1}$ such that:

$$\left\| \frac{\boldsymbol{\theta}_{s_m}}{\|\boldsymbol{\theta}_{s_m}\|} - \bar{\boldsymbol{\theta}} \right\| \leq \epsilon_m = \frac{1}{m} \qquad \text{and} \qquad \frac{1}{L} \log \frac{\widetilde{\gamma}_\infty}{\widetilde{\gamma}_{s_m}} \leq \epsilon_m^2 = \frac{1}{m^2}. \tag{56}$$

Since $\frac{d \log \|\boldsymbol{\theta}_t\|}{dt} \leq \frac{\left\|\frac{d\boldsymbol{\theta}_t}{dt}\right\|}{\|\boldsymbol{\theta}_t\|}$, we have that $\lim_{t\to\infty} \int_{t_A}^{t} \frac{\left\|\frac{d\boldsymbol{\theta}_{t'}}{dt'}\right\|}{\|\boldsymbol{\theta}_{t'}\|} dt' = \infty$ for all $t_A > 0$. Thus, there exists $s'_m > s_m$ such that $\int_{s_m}^{s'_m} \frac{\left\|\frac{d\boldsymbol{\theta}_t}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} dt = \frac{1}{m}$. Now, from Lemma A.16, we know there exists $t_\star \in (s_m, s'_m)$ with:

$$\left(\frac{1}{\left\langle \frac{\boldsymbol{\theta}_{t_\star}}{\|\boldsymbol{\theta}_{t_\star}\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star} \right\rangle} - 1 \right) \leq \frac{1}{L} \frac{\log \frac{\widetilde{\gamma}_{s'_m}}{\widetilde{\gamma}_{s_m}}}{\int_{s_m}^{s'_m} \frac{\left\|\frac{d\boldsymbol{\theta}_t}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} dt} \leq \frac{\frac{1}{m^2}}{\frac{1}{m}} = \frac{1}{m}, \tag{57}$$

which implies $\left\langle \frac{\boldsymbol{\theta}_{t_\star}}{\|\boldsymbol{\theta}_{t_\star}\|}, \frac{-\mathbf{g}_t^\star}{\|\mathbf{g}_t^\star\|_\star} \right\rangle \geq \frac{1}{1+\frac{1}{m}} \to 1$ as $m \to \infty$. On the other hand, for the normalized iterates we have:

$$\left\| \frac{\boldsymbol{\theta}_{t_\star}}{\|\boldsymbol{\theta}_{t_\star}\|} - \bar{\boldsymbol{\theta}} \right\| \leq \left\| \frac{\boldsymbol{\theta}_{t_\star}}{\|\boldsymbol{\theta}_{t_\star}\|} - \frac{\boldsymbol{\theta}_{s_m}}{\|\boldsymbol{\theta}_{s_m}\|} \right\| + \left\| \frac{\boldsymbol{\theta}_{s_m}}{\|\boldsymbol{\theta}_{s_m}\|} - \bar{\boldsymbol{\theta}} \right\| \overset{\text{eq. Equation 56}}{\leq} \left\| \frac{\boldsymbol{\theta}_{t_\star}}{\|\boldsymbol{\theta}_{t_\star}\|} - \frac{\boldsymbol{\theta}_{s_m}}{\|\boldsymbol{\theta}_{s_m}\|} \right\| + \frac{1}{m}. \tag{58}$$

To deal with the first term, we can leverage Lemma A.17 which bounds the rate of change of the normalized iterates:

$$\left\| \frac{\boldsymbol{\theta}_{t_\star}}{\|\boldsymbol{\theta}_{t_\star}\|} - \bar{\boldsymbol{\theta}} \right\| \leq 2 \int_{s_m}^{t_\star} \frac{\left\|\frac{d\boldsymbol{\theta}_t}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} dt + \frac{1}{m} \leq 2 \int_{s_m}^{s'_m} \frac{\left\|\frac{d\boldsymbol{\theta}_t}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} dt + \frac{1}{m} = \frac{3}{m} \to 0. \tag{59}$$

Hence, by picking t_m as $t_\star$, we constructed a time sequence such that, for any limit point $\bar{\boldsymbol{\theta}}$, $\frac{\boldsymbol{\theta}_{t_m}}{\|\boldsymbol{\theta}_{t_m}\|} \to \bar{\boldsymbol{\theta}}$ and also $\left\langle \frac{\boldsymbol{\theta}_{t_m}}{\|\boldsymbol{\theta}_{t_m}\|}, \frac{-\mathbf{g}_t^\star(\boldsymbol{\theta}_{t_m})}{\|\mathbf{g}_t^\star(\boldsymbol{\theta}_{t_m})\|_\star} \right\rangle \to 1$.

Then, from Proposition A.15, we know that $\widetilde{\boldsymbol{\theta}}_{t_m} = \frac{\boldsymbol{\theta}_{t_m}}{\left(\min_{i\in[m]} y_i f(\mathbf{x}_i; \boldsymbol{\theta}_{t_m})\right)^{\frac{1}{L}}}$ is an $\left(D^{\widetilde{\boldsymbol{\theta}}_t}_{\frac{1}{2}\|\cdot\|_\star^2}, \epsilon(t_m), \delta(t_m)\right)$-approximate KKT point of Equation 37. But $\epsilon(t_m) \to 0$ (since the alignment goes to 1) and $\delta(t_m) \to 0$ as the loss goes to zero (Lemma A.6), thus the sequence satisfies the conditions of Proposition A.20, which shows that the limit point of the sequence is a generalized KKT point. This concludes the proof of our claim. □

While the previous result is not strong enough to guarantee convergence to an approximate KKT point for a general algorithm norm $\|\cdot\|$, it immediately implies it in the case of a smooth norm. The proof relies on a fundamental relationship between smoothness of a function and strong convexity of its convex conjugate.

Proposition A.19. *(Conjugate Correspondence Theorem - Thm. 5.26 in Beck (2017)) Let $\sigma > 0$. If ψ is a $\frac{1}{\sigma}$-smooth convex function, then its conjugate $\psi^\star$ is σ-strongly convex.*

We can prove the following corollary for a special class of steepest flows.

Corollary A.19.1. *For steepest flow (eq. Equation 3) with respect to a norm $\|\cdot\|$, whose square is a smooth function, on the exponential loss, under assumptions A1, A2, A3, any limit point $\bar{\boldsymbol{\theta}}$ of $\left\{\frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|}\right\}_{t\geq 0}$ is along the direction of a KKT point of optimization problem Equation 37.*

Proof. From Proposition A.19, if $\frac{1}{2}\|\cdot\|^2$ is $\frac{1}{\sigma}$-smooth w.r.t. $\|\cdot\|$, then the function $\frac{1}{2}\|\cdot\|_\star^2$ is σ-strongly convex w.r.t. $\|\cdot\|_\star$. Thus, the function $D^{\boldsymbol{\theta}}_{\frac{1}{2}\|\cdot\|_\star^2}$ is defined with respect to a strongly convex function and it becomes a proper Bregman divergence. Hence, from Theorem 5.24 in Beck (2017), for $\widetilde{\mathbf{h}}_i^\star = q_{\min}^{\frac{1}{L}-1}\mathbf{h}_i^\star$, where $\widetilde{\mathbf{h}}_i^\star \in \partial f(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}), i \in [m]$ such that $\mathbf{g}_t^\star = -\sum_{i=1}^m e^{-y_i f(\mathbf{x}_i;\boldsymbol{\theta})} y_i \mathbf{h}_i^\star$ and $\widetilde{\mathbf{k}} \in \partial\frac{1}{2}\|\widetilde{\boldsymbol{\theta}}\|^2$, it holds:

$$D^{\widetilde{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}\left(\sum_{i=1}^m \lambda_i y_i \widetilde{\mathbf{h}}_i^\star, \widetilde{\mathbf{k}}\right) \geq \sigma \left\|\sum_{i=1}^m \lambda_i y_i \widetilde{\mathbf{h}}_i^\star - \widetilde{\mathbf{k}}\right\|_\star. \tag{60}$$

In other words, if $D^{\widetilde{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is 0, so is the difference $\boldsymbol{\alpha}-\boldsymbol{b}$ for any $\boldsymbol{\alpha},\boldsymbol{b}$. As a result, and from the equivalence of the norms, the sequence $\widetilde{\boldsymbol{\theta}}_{t_m} = \frac{\boldsymbol{\theta}_{t_m}}{\left(\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_{t_m})\right)^{\frac{1}{L}}}$ from the proof of Theorem A.18, induces a sequence of $(\epsilon(t_m),\delta(t_m))$-approximate KKT points, which converges to a KKT point of Equation 37. By Theorem C.4 in Lyu & Li (2020) (which is itself based on a result due to Dutta et al. (2013)), we get that $\frac{\bar{\boldsymbol{\theta}}}{\left(\min_{i\in[m]} y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)\right)^{\frac{1}{L}}}$ is a KKT point of Equation 37. □

The following technical result was used in the proof of Theorem A.18.

Proposition A.20. *Let (MM) be the following optimization problem:*

$$\begin{aligned} &\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ &\text{s.t. } y_i f(\mathbf{x}_i;\boldsymbol{\theta}) \geq 1,\ \forall i \in [m]. \end{aligned} \tag{MM}$$

Let $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ be a sequence of feasible, (d,ϵ_t,δ_t)-approximate KKT points, with $d := D^{\boldsymbol{\theta}_t}_{\frac{1}{2}\|\cdot\|_\star^2}$ with $\epsilon_t \downarrow 0, \delta_t \downarrow 0$ and $\boldsymbol{\theta}_t \to \bar{\boldsymbol{\theta}}$. Assume that the Mangasarian-Fromovitz- Constraint Qualifications hold at $\bar{\boldsymbol{\theta}}$. Then, $\bar{\boldsymbol{\theta}}$ is a $D^{\bar{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}$-generalized KKT point of (MM).

Proof. Our proof closely follows the proof of Theorem 3.6 in (Dutta et al., 2013), which is the direct analog for (ϵ,δ)-approximate KKT points.

By the definition of the (d,ϵ_t,δ_t) stationarity, for each $t > 0$, there exist $\mathbf{h}_i^t \in \partial f(\mathbf{x}_i;\boldsymbol{\theta}_t), i \in [m], \mathbf{k}^t \in \partial\frac{1}{2}\|\boldsymbol{\theta}_t\|^2$ and $\lambda_i^t \geq 0, i \in [m]$ such that:

(i) $D^{\boldsymbol{\theta}_t}_{\frac{1}{2}\|\cdot\|_\star^2}\left(\sum_{i=1}^m \lambda_i^t y_i \mathbf{h}_i^t, \mathbf{k}^t\right) \leq \epsilon_t$.

(ii) $\sum_{i=1}^m \lambda_i^t \left(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t) - 1\right) \leq \delta_t$.

We will show that the sequence $\{\boldsymbol{\lambda}^t\}_{t\geq 0}$ is bounded. Assume on the contrary that is not. Consider $\hat{\boldsymbol{\lambda}} = \frac{\boldsymbol{\lambda}^t}{\|\boldsymbol{\lambda}^t\|}$, which is bounded and wlog, it is: $\hat{\boldsymbol{\lambda}} \to \widetilde{\boldsymbol{\lambda}}$ with $\|\hat{\boldsymbol{\lambda}}\| = 1$. Note that the sequences $\{\mathbf{h}_i^t\}, \{\mathbf{k}^t\}$ are bounded, as elements of a subdifferential, by the Lipschitz constant of the corresponding function. Hence, they also converge to, say, $\bar{\mathbf{h}}_i$ for all $i \in [m]$ and $\bar{\mathbf{k}}$, respectively. From condition (i), we have:

$$\frac{1}{\|\boldsymbol{\lambda}^t\|^2} D^{\boldsymbol{\theta}_t}_{\frac{1}{2}\|\cdot\|_\star^2}\left(\sum_{i=1}^m \lambda_i^t y_i \mathbf{h}_i^t, \mathbf{k}^t\right) \leq \frac{\epsilon_t}{\|\boldsymbol{\lambda}_t\|^2}, \tag{61}$$

or

$$\frac{1}{2}\left\|\sum_{i=1}^m \frac{\lambda_i^t}{\|\boldsymbol{\lambda}^t\|} y_i \mathbf{h}_i^t\right\|_\star^2 - \frac{1}{2}\left\|\frac{\mathbf{k}^t}{\|\boldsymbol{\lambda}^t\|}\right\|_\star^2 - \left\langle \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\lambda}^t\|}, \sum_{i=1}^m \frac{\lambda_i^t}{\|\boldsymbol{\lambda}\|} y_i \mathbf{h}_i^t - \frac{\mathbf{k}^t}{\|\boldsymbol{\lambda}\|}\right\rangle \leq \frac{\epsilon_t}{\|\boldsymbol{\lambda}_t\|^2}, \tag{62}$$

hence, by taking $t \to \infty$, we get:

$$\frac{1}{2}\left\|\sum_{i=1}^m \widetilde{\lambda}_i y_i \bar{\mathbf{h}}_i\right\|_\star^2 \leq 0, \tag{63}$$

which implies that there exists $\widetilde{\boldsymbol{\lambda}} \neq \mathbf{0}$ such that $\sum_{i=1}^m \widetilde{\lambda}_i y_i \bar{\mathbf{h}}_i = \mathbf{0}$, where recall $\bar{\mathbf{h}}_i \in \partial f(\mathbf{x}_i; \bar{\boldsymbol{\theta}})$. An existence of such a vector is prohibited by the Mangasarian-Fromovitz-Constraint Qualifications which hold at $\bar{\boldsymbol{\theta}}$. Thus, $\{\boldsymbol{\lambda}^t\}_{t\geq 0}$ is bounded and $\boldsymbol{\lambda}^t \to \bar{\boldsymbol{\lambda}}$ for some $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^m$. Hence, by taking the limit $t \to \infty$, we have:

$$D^{\bar{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}\left(\sum_{i=1}^m \bar{\lambda}_i y_i \bar{\mathbf{h}}_i, \bar{\mathbf{k}}\right) \leq 0, \tag{64}$$

where $\bar{\mathbf{h}}_i \in \partial f(\mathbf{x}_i; \bar{\boldsymbol{\theta}}), i \in [m]$ and $\bar{\mathbf{k}} \in \partial \frac{1}{2}\|\bar{\boldsymbol{\theta}}\|$. From (ii), we obtain $\sum_{i=1}^m \bar{\lambda}_i \left(y_i f(\mathbf{x}_i; \bar{\boldsymbol{\theta}}) - 1\right) \leq 0$. However, $y_i f(\mathbf{x}_i; \bar{\boldsymbol{\theta}}) - 1 \geq 0$ ($\bar{\boldsymbol{\theta}}$ is a feasible point of (P)) and $\bar{\lambda}_i \geq 0$ for all $i \in [m]$, thus it holds:

$$\sum_{i=1}^m \bar{\lambda}_i \left(y_i f(\mathbf{x}_i; \bar{\boldsymbol{\theta}}) - 1\right) = 0, \tag{65}$$

which concludes the proof that $\bar{\boldsymbol{\theta}}$ is a $D^{\bar{\boldsymbol{\theta}}}_{\frac{1}{2}\|\cdot\|_\star^2}$-generalized KKT point.

□

A.3 Generalization to other losses

The previous results can be generalized to any loss with exponential tails. In particular, let us proceed to the following definition:

Definition A.21. *Let $\Phi : \mathbb{R} \to \mathbb{R}$. Assume that $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^m e^{-\Phi(y_i f(\mathbf{x}_i;\boldsymbol{\theta}))}, \boldsymbol{\theta} \in \mathbb{R}^p$, where $f : \mathbb{R}^d \to \mathbb{R}, y_i \in \{\pm 1\}$. We call the function $l : \mathbb{R} \to \mathbb{R}, l(u) := e^{-\Phi(u)}$, exponentially tailed, if the following conditions hold:*

- *(i) Φ is continuously differentiable.*
- *(ii) $\Phi'(u) > 0$ for all $u \in \mathbb{R}$.*
- *(iii) The function $g(u) = \Phi'(u)u$ is non-decreasing in $[0, \infty)$.*

Notice that the definition above covers the exponential loss for $\Phi(u) = u$ and the logistic loss for $\Phi(u) = -\log\log(1 + e^{-u})$. To accommodate different loss functions, Assumption A3 needs to be adjusted as follows:

- There is a $t_0 > 0$, such that $\mathcal{L}(\boldsymbol{\theta}_{t_0}) < e^{-\Phi(0)}$.

See Section A in (Lyu & Li, 2020) for a more general, albeit technical, definition that allows the extension of the full analysis to general exponentially-tailed losses.

Under these conditions, we can define the soft margin as follows:

$$\widetilde{\gamma} = \frac{l^{-1}(\mathcal{L})}{\|\boldsymbol{\theta}\|^L} = \frac{\Phi^{-1}\left(\log \frac{1}{\mathcal{L}}\right)}{\|\boldsymbol{\theta}\|^L},$$

and prove a strict generalization of Theorem 3.1.

Theorem A.22 (Soft margin increases - general loss function)**.** *For almost any $t > t_0$, it holds:*

$$\frac{d\log\widetilde{\gamma}}{dt} \geq L\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2 \left(\frac{\left(\Phi^{-1}\right)'\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)}{L\mathcal{L}(\boldsymbol{\theta}_t)\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)} - \frac{1}{\|\boldsymbol{\theta}_t\|\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}\right) \geq 0.$$

Proof. Let $\mathbf{n}_t \in \partial\|\boldsymbol{\theta}_t\|$. We have:

$$\begin{aligned}
\frac{d\log\widetilde{\gamma}}{dt} &= \frac{d}{dt}\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right) - L\frac{d}{dt}\log\|\boldsymbol{\theta}_t\| \\
&= \frac{d}{dt}\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right) - L\left\langle\frac{\mathbf{n}_t}{\|\boldsymbol{\theta}_t\|}, \frac{d\boldsymbol{\theta}}{dt}\right\rangle \qquad \text{(Chain rule)} \\
&\geq \frac{d}{dt}\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right) - L\frac{\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} \qquad \text{(definition of dual norm and } \|\mathbf{n}_t\|_\star \leq 1) \\
&= -\frac{d\mathcal{L}(\boldsymbol{\theta}_t)}{dt}\frac{\left(\Phi^{-1}\right)'\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)}{\mathcal{L}(\boldsymbol{\theta}_t)\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)} - L\frac{\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}{\|\boldsymbol{\theta}_t\|} \qquad \text{(Chain rule)} \\
&= \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2\left(\frac{\left(\Phi^{-1}\right)'\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)}{\mathcal{L}(\boldsymbol{\theta}_t)\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)} - \frac{L}{\|\boldsymbol{\theta}_t\|\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}\right) \qquad \text{(Equation 18).}
\end{aligned} \tag{66}$$

But, the first term inside the parenthesis can be related to the second one via the following calculation. Recall that, by the chain rule for locally Lipschitz functions (Theorem A.2), for any $\mathbf{g}_t \in \partial\mathcal{L}(\boldsymbol{\theta}_t)$ there exist $\mathbf{h}_1 \in \partial y_1 f(\mathbf{x}_1;\boldsymbol{\theta}_t), \ldots, \mathbf{h}_m \in \partial y_m f(\mathbf{x}_m;\boldsymbol{\theta}_t)$ such that $\mathbf{g}_t = \sum_{i=1}^m e^{-\Phi(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))}\Phi'(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t)))\mathbf{h}_i$. Thus, for a minimum norm subderivative $\mathbf{g}_t^\star$, we have:

$$\begin{aligned}
\langle\boldsymbol{\theta}_t, -\mathbf{g}_t^\star\rangle &= \left\langle\boldsymbol{\theta}_t, \sum_{i=1}^m e^{-\Phi(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))}\Phi'(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))\mathbf{h}_i^\star\right\rangle \\
&= \sum_{i=1}^m e^{-\Phi(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))}\Phi'(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))\,\langle\boldsymbol{\theta}_t, \mathbf{h}_i^\star\rangle \\
&= L\sum_{i=1}^m e^{-\Phi(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))}\Phi'(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t),
\end{aligned} \tag{67}$$

where the last equality follows from Euler's theorem for homogeneous functions (whose generalization for subderivatives can be found in Theorem B.2 in Lyu & Li (2020)). But, now observe that as per assumption, $u \to \Phi'(u)u$ is non-decreasing and this last term can be lower bounded as:

$$\langle\boldsymbol{\theta}_t, -\mathbf{g}_t^\star\rangle \geq L\sum_{i=1}^m e^{-\Phi(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))}\Phi'\left(\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)\right)\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right), \tag{68}$$

where we used the fact $y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t) \leq \Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)$ for all $i \in [m]$ (by the monotonicity of Φ the definition of $\mathcal{L}$). Leveraging the fundamental property between the derivative of a function and its inverse's, we further get:

$$\langle\boldsymbol{\theta}_t, -\mathbf{g}_t^\star\rangle \geq L\sum_{i=1}^m e^{-\Phi(y_i f(\mathbf{x}_i;\boldsymbol{\theta}_t))}\frac{\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)}{\left(\Phi^{-1}\right)'\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t}\right)} = L\mathcal{L}(\boldsymbol{\theta}_t)\frac{\Phi^{-1}\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t)}\right)}{\left(\Phi^{-1}\right)'\left(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}_t}\right)}. \tag{69}$$

We have made the first term of Equation 66 appear. By plugging Equation 69 into Equation 66, we get:

$$\begin{aligned}
\frac{d\log\widetilde{\gamma}}{dt} &\geq \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2\left(\frac{L}{\langle\boldsymbol{\theta}_t, -\mathbf{g}_t^\star\rangle} - \frac{L}{\|\boldsymbol{\theta}_t\|\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}\right) \\
&\geq \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|^2\left(\frac{L}{\|\boldsymbol{\theta}_t\|\|\mathbf{g}_t^\star\|_\star} - \frac{L}{\|\boldsymbol{\theta}_t\|\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|}\right) \qquad \text{(definition of dual norm).}
\end{aligned} \tag{70}$$

Noticing that $\|\mathbf{g}_t^\star\|_\star = \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|$ (from Proposition A.13) concludes the proof. □

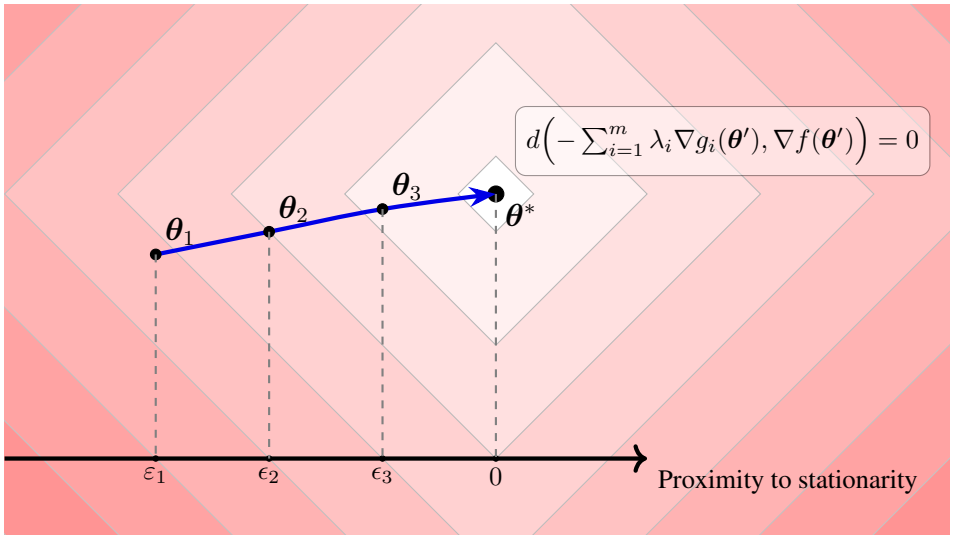

Figure 3: **An illustration of the reduction of the Bregman proximity measure.** Once it separates the training points, steepest descent in homogeneous networks implicitly makes progress towards generalized stationary points $\boldsymbol{\theta}^\star$ of a margin maximization problem (Theorem A.18).

B Bregman proximity measure

In the proof of our main result (Theorem A.18), we constructed a sequence of approximate generalized KKT points (Definition A.9). However, in many cases, while solving (non-convex) optimization problems, we only have feasible points without any evidence of optimality or stationarity. In such cases, it is useful to come up with a *progress measure* of approximate stationarity that can also serve as a stopping criterion for the optimization algorithm. Consider an optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta}) \\ &\text{s.t. } g_i(\boldsymbol{\theta}) \leq 0 \ \forall i \in [m], \end{aligned} \tag{P}$$

where we assume that $f, \{g_i\}_{i=1}^m$ are differentiable for the sake of brevity. For a feasible point $\boldsymbol{\theta}'$ of (P) and a non-negative function $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$, we define the *$d$-Bregman proximity measure* as the solution of the following optimization problem:

$$\begin{aligned} &\min_{\epsilon,\lambda_1,\ldots,\lambda_m} \epsilon \\ &\text{s.t. } d\left(-\sum_{i=1}^m \lambda_i \nabla g_i(\boldsymbol{\theta}'), \nabla f(\boldsymbol{\theta}')\right) \leq \epsilon, \\ &\sum_{i=1}^m \lambda_i g_i(\boldsymbol{\theta}') \geq -\epsilon, \\ &\lambda_i \geq 0 \ \forall i \in [m]. \end{aligned} \tag{71}$$

This definition mirrors and generalizes the definitions of (Dutta et al., 2013), which were inspired by approximate KKT points (whose proximity is measured using the Euclidean distance as d). However, as we saw in our analysis, there are many cases of problems where a proximity measure would be better defined using alternatives functions. Figure 3 conceptually illustrates the reduction of the Bregman divergence in a possible converging path. The relaxation of the slackness constraints in the form of $\sum_{i=1}^m \lambda_i g_i(\boldsymbol{\theta}') \geq -\epsilon$ is essential for ensuring that the proximity measure captures proximity to stationarity - see Section 3.2 in (Dutta et al., 2013) for a discussion.

C Relationship to Adam and Shampoo

The family of steepest descent algorithms includes simplified versions (momentum turned-off) of two adaptive methods, Adam and Shampoo, which have been very popular for training deep neural networks.

C.1 Adam

Adam (Kingma & Ba, 2015) is a popular adaptive optimization method, which is frequently used in deep learning. Following our previous notation, the update rule of Adam amounts to:

$$\begin{aligned} \mathbf{m}_t &= \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\nabla\mathcal{L}(\boldsymbol{\theta}_t) \\ \mathbf{v}_t &= \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\nabla\mathcal{L}(\boldsymbol{\theta}_t)^2 \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta_t \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}+\epsilon}, \end{aligned} \tag{72}$$

where the $\sqrt{\cdot}, {}^2, \div$ operations are overloaded to operate elementwise in vectors. Parameters β_1, β_2 control the memory of the update rule, while ϵ is a numerical precision parameter. Notice that for $\beta_1 = \beta_2 = \epsilon = 0$, we recover sign-gradient descent.

Wang et al. (2022)studied the implicit bias of (72) in linear networks establishing bias towards ℓ_2 margin maximization, while Zhang et al. (2024) analyzed the case of $\epsilon = 0$ and generic $\beta_1, \beta_2 \in [0, 1)$ also in linear networks and showed bias towards ℓ_∞ margin maximization.

C.2 Shampoo

Shampoo (Gupta et al., 2018) is an adaptive optimization algorithm, which has recently gained popularity in deep learning applications. For each weight matrix $\mathbf{W}_t$ and its corresponding gradient matrix $\mathbf{G}_t$, the update rule of Shampoo without momentum amounts to (Bernstein & Newhouse, 2024):

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \mathbf{U}_t \mathbf{V}_t^T, \tag{73}$$

where $\mathbf{U}_t, \mathbf{V}_t$ contain the left and right singular vectors of $\mathbf{G}_t$, i.e., $\mathbf{G}_t = \mathbf{U}_t \boldsymbol{\Sigma}_t \mathbf{V}_t$. Bernstein & Newhouse (2024) recently noticed that this update corresponds to steepest descent (in matrix space) with respect to the *spectral norm* $\sigma_{\max}(\cdot)$. This is equivalent to an architecture-dependent norm in parameter space. For instance, if $\boldsymbol{\theta} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$, then Shampoo without momentum corresponds to steepest descent with respect to the norm $\|\boldsymbol{\theta}\|_S := \max\left(\sigma_{\max}(\mathbf{W}_1), \ldots, \sigma_{\max}(\mathbf{W}_L)\right)$.

D Experimental Details

All experiments are implemented in PyTorch (Paszke et al., 2017).

Teacher-student experiments We use the following hyperparameters: $d = 2^{32}, k = 64, k' = 1024, m = 250$, learning rate $\eta = 6 \times 10^{-3}$ and density $\frac{\|\boldsymbol{\theta}^\star\|_0}{k'(d+1)} = 0.0001$ (3 coordinates active per neuron). We vary the scale of initialization in $\{0.1, 0.01, 0.001\}$ and we train for 10^5 epochs. Each random seed affects the draw of the datasets and the initialization of the parameters of the network. Test accuracy is estimated using 20,000 unseen data drawn from the same generative process.

MNIST We use a constant learning rate of 3×10^{-3} and 1-hidden layer neural networks of width 128, optimizing the logistic loss. The digits that we extract are '3' and '6' (100 training points). Each random seed corresponds to a different draw of the training dataset and different initialization. Sign gradient descent runs were very effective in minimizing the training loss, and we stopped the training early after the loss reached value smaller than 10^{-7} in order to avoid numerical issues. We depict the final value, repeated for as many epochs as shown in the figures (as if the model has indeed converged).

Figure 5 shows accuracy and margins for a different pair of digits ("2" vs "7").

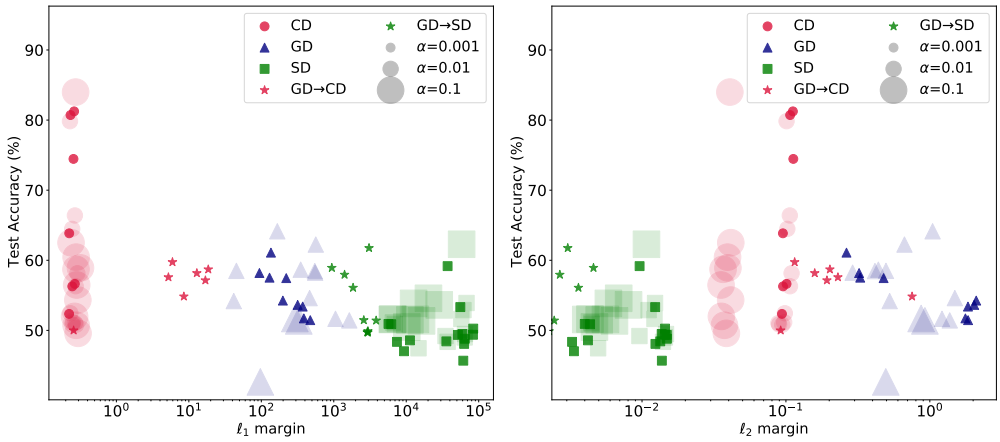

Figure 4: **Geometric margins vs test accuracy in a teacher-student setup.** *Left:* ℓ_1 margin. *Right:* ℓ_2 margin. Each point corresponds to a different run (different random seed).

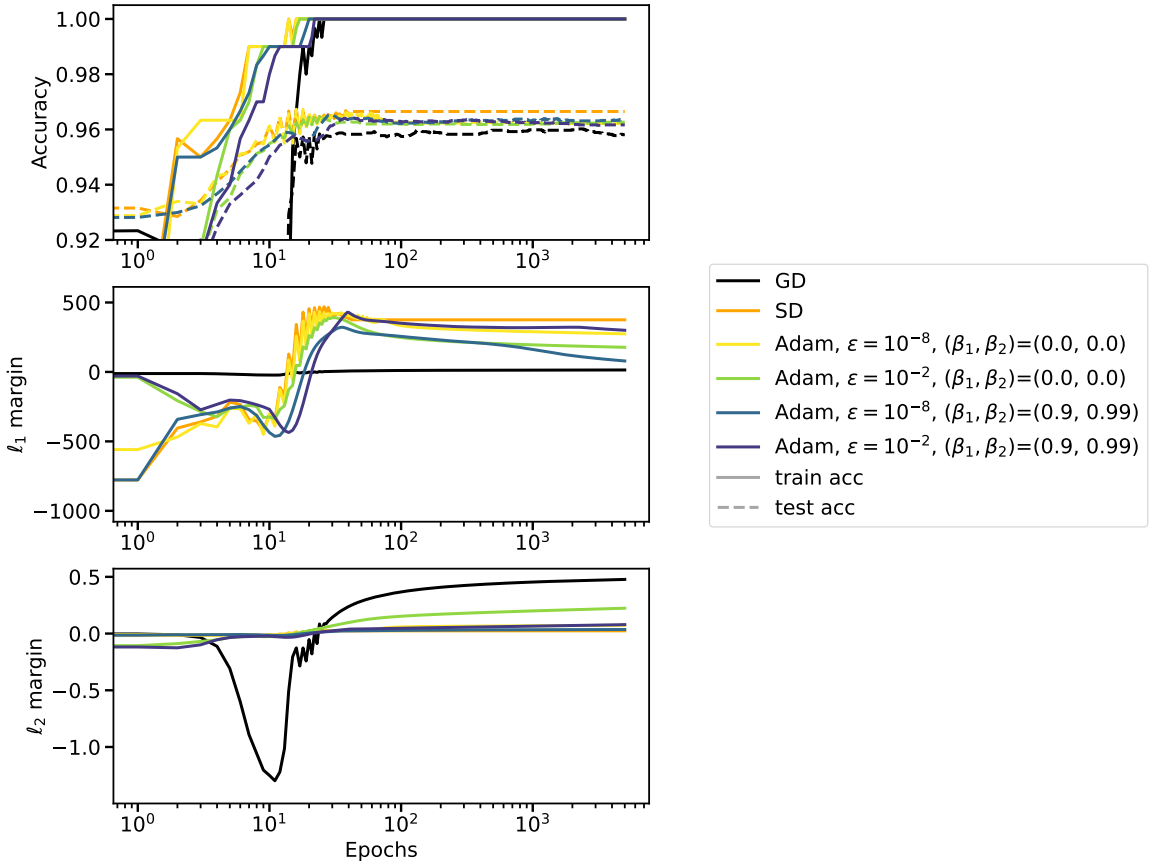

Figure 5: **Relationship between Adam and steepest descent algorithms.** Digits '2' and '7'.