# Convergence of Gradient Descent on Separable Data

**Mor Shpigel Nacson**[1] **Jason D. Lee**[2] **Suriya Gunasekar**[3] **Pedro H. P. Savarese**[3] **Nathan Srebro**[3] **Daniel Soudry**[1]
[1]Technion, Israel, [2]USC Los Angeles, USA, [3]TTI Chicago, USA

## Abstract

We provide a detailed study on the implicit bias of gradient descent when optimizing loss functions with strictly monotone tails, such as the logistic loss, over separable datasets. We look at two basic questions: (a) what are the conditions on the tail of the loss function under which gradient descent converges in the direction of the $L_2$ maximum-margin separator? (b) how does the rate of margin convergence depend on the tail of the loss function and the choice of the step size? We show that for a large family of super-polynomial tailed losses, gradient descent iterates on linear networks of any depth converge in the direction of $L_2$ maximum-margin solution, while this does not hold for losses with heavier tails. Within this family, for simple linear models we show that the optimal rates with fixed step size is indeed obtained for the commonly used exponentially tailed losses such as logistic loss. However, with a fixed step size the optimal convergence rate is extremely slow as $1/\log(t)$, as also proved in Soudry et al. (2018a). For linear models with exponential loss, we further prove that the convergence rate could be improved to $\log(t)/\sqrt{t}$ by using aggressive step sizes that compensates for the rapidly vanishing gradients. Numerical results suggest this method might be useful for deep networks.

## 1 INTRODUCTION

In learning over-parameterized models, where the training objective has multiple global optima, each optimization algorithm can have a distinct implicit bias. Hence, different algorithms learn different models with different generalization to the population loss. This effect of the implicit bias of the optimization algorithm on the learned model is particularly prominent in deep learning, where the generalization or the inductive bias is not sufficiently driven by explicit regularization or restrictions on the model capacity (Neyshabur et al., 2015; Zhang et al., 2017; Hoffer et al., 2017). Thus, in order to understand what is the true inductive bias in such high capacity models, it is important to rigorously understand how optimization affects the implicit bias.

Consider learning a homogeneous linear predictor $\mathbf{x} \to \mathbf{w}^\top \mathbf{x}$ using unregularized logistic regression over separable data. For this problem, Soudry et al. (2018a) showed that the gradient descent iterates converge in direction to the maximum-margin separator with unit $L_2$ norm, and this implicit bias holds independently of initialization and step size (given the step size is small enough). This is exactly the solution of the homogeneous hard margin support vector machine (SVM) where the $L_2$ norm constraint on the parameters $\mathbf{w}$ is explicitly added. More surprisingly, Soudry et al. (2018a) also showed that the rate of convergence to the maximum-margin solution is $O(1/\log(t))$. This is much slower compared to the rate of convergence of the loss function itself, which is shown to be $O(1/t)$. This implies that the classification boundary of logistic regression, and hence the generalization of the classifier, continues to change long after the 0-1 error on training examples has diminished to zero, or the logistic loss is very small. In a follow up work, Gunasekar et al. (2018a) showed that for exponential loss, gradient descent on fully connected deep linear networks also has the same bias asymptotically. However, the convergence rates were not analyzed in this work on deep linear networks.

Despite this recent line of interesting results, the implicit bias of gradient descent is not entirely understood even in simple linear classification tasks. For example, the analysis of Soudry et al. (2018a) and Gunasekar et al. (2018a) crucially relied on strict monotonicity of the loss function to get an initialization–independent characterization of the bias of gradient descent. However, in these work the results are derived specifically for tight exponential tailed losses and exponential loss, respectively. While exponential tailed losses such as logistic and cross entropy losses are indeed the most widely used losses in training deep neural net-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

works, we do not yet know: Do such losses with tight exponential tail have a special significance? Can a similar convergence to maximum-margin separator be achieved by other strictly monotonic losses? How is the rate of convergence to maximum-margin solution affected by the tail? Are there other ways to accelerate the convergence?

Here we provide a detailed study of this problem, focusing on the rate of convergence of the margin:

1. *What are the conditions on the tail of the loss function under which gradient descent converges to the $L_2$ maximum-margin separator?* We show that convergence to the $L_2$ maximum-margin solution can be extended to losses with super polynomial tails, but not to losses with (sub) polynomial tails.
2. *Does a heavier or lighter tail gives a faster rate of convergence?* In our analysis, losses with exponential tails, which include the commonly used logistic loss, can indeed be shown to have the optimal rate of convergence of the margin.
3. *Extensions to deep linear networks.* We show that similar analysis and the same asymptotic rates hold more generally for linear networks with fully connected layers. Interestingly, the results suggest that increasing the number of layers (depth) decreases the convergence rate only marginally, even in the limit of infinite depth.
4. *For exponential loss, which obtains the optimal margin convergence rate, can we accelerate the convergence to the maximum-margin by using variable step sizes?* The answer is yes, and we show that using normalized gradient updates, i.e., step size proportional to the inverse gradient, we can get a much faster rate of $O(\log t/\sqrt{t})$ instead of $1/\log t$. Experimental results suggests this improvement in rate over standard gradient descent might also extend for non-linear neural networks.

## 2 SETUP AND REVIEW OF PREVIOUS RESULTS

Consider a dataset $\{\mathbf{x}_n, y_n\}_{n=1}^N$, with features $\mathbf{x}_n \in \mathbb{R}^d$ and binary labels $y_n \in \{-1, 1\}$. All the results in the paper are stated for data $\{\mathbf{x}_n, y_n\}_{n=1}^N$ which is *strictly linearly separable*, i.e., there exists a separator $\mathbf{w}_*$ such that $\forall n : y_n \mathbf{w}_*^\top \mathbf{x}_n > 0$.

We study learning homogenous linear predictors by minimizing unregularized empirical losses of the form

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} \ell\left(y_n \mathbf{w}^\top \mathbf{x}_n\right), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector or the linear predictor. To simplify notation, we assume that $\forall n : y_n = 1$ — this is without loss of generality, since we can always re-define $y_n \mathbf{x}_n$ as $\mathbf{x}_n$. We denote the data matrix by $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ and $\|\cdot\|$ denotes the $L_2$ norm.

The gradient descent (GD) updates for minimizing $\mathcal{L}(\mathbf{w})$ in eq. (1) with step size sequence $\{\eta_t\}_{t=0}^{\infty}$ is given by

$$\begin{aligned}\mathbf{w}(t+1) &= \mathbf{w}(t) - \eta_t \nabla \mathcal{L}(\mathbf{w}(t)) \\ &= \mathbf{w}(t) - \eta_t \sum_{n=1}^{N} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\mathbf{x}_n. \end{aligned} \tag{2}$$

We look at the iterates of GD on linearly separable datasets with monotonic loss functions.

**Definition 1.** *[Strict Monotone Loss] $\ell(u)$ is a differentiable strictly monotonically decreasing function bounded from below, i.e. $\forall u, \ell(u)' < 0$ and, without loss of generality, $\forall u, \ \ell(u) > 0$ and $\lim_{u\to\infty} \ell(u) = \lim_{u\to\infty} \ell'(u) = 0$. Also, $\limsup_{u\to-\infty} \ell'(u) \neq 0$.*

Examples of strict monotone losses include common classification losses such as logistic loss, exponential loss, and probit loss. A key property of interest with such losses is that the empirical risk in eq. (1) over separable data does not have any finite global minimizers. Thus, whenever the gradient descent updates in eq. (2) minimize the empirical loss $\mathcal{L}(\mathbf{w})$, the iterates $\mathbf{w}(t)$ will necessarily diverge to infinity. Nevertheless, in this case, even though the norm of the iterates $\|\mathbf{w}(t)\|$ diverge, the classification boundary is entirely specified by the direction of $\mathbf{w}(t)/\|\mathbf{w}(t)\|$. Can we say something interesting about which direction the iterates $\mathbf{w}(t)$ converge to?

For monotone losses with $-\ell'(u)$ satisfying the specific *tight exponential tail* property (defined below), Soudry et al. (2018a) characterized this direction to be the maximum-margin separator,

**Definition 2.** *[Tight Exponential Tail] A scalar function $h(u)$ has a tight exponential tail, if there exist positive constants $\mu_+, \mu_-$, and $\bar{u}$ such that $\forall u > \bar{u}$:*

$$(1 - \exp(-\mu_- u))e^{-u} \le h(u) \le (1 + \exp(-\mu_+ u))e^{-u}.$$

**Theorem 1** (Theorem 3 in Soudry et al. (2018a), rephrased). *For almost all linearly separable datasets $\{\mathbf{x}_n, y_n\}_{n=1}^N$, and any $\beta$-smooth $\mathcal{L}$ with a strictly monotone loss function $\ell$ (Definition 1), for which $-\ell'$ has a tight exponential tail (Definition 2), the gradient descent iterates $\mathbf{w}(t)$ in eq. (2) with any fixed step size satisfying*[1] *$\eta < 2\beta^{-1}$ and any initialization $\mathbf{w}(0)$, will behave as:*

$$\mathbf{w}(t) = \hat{\mathbf{w}} \log t + \boldsymbol{\rho}(t), \tag{3}$$

[1]Note that for exponential loss $\ell(u) = \exp(-u)$, $\mathcal{L}(\mathbf{w})$ does not have a global smoothness parameter $\beta$. However, with $\eta < 1/\mathcal{L}(\mathbf{w}(0))$ it is straightforward to show the gradient descent iterates maintain bounded local smoothness $\beta(t) \le \mathcal{L}(\mathbf{w}(t)) \le \mathcal{L}(\mathbf{w}(0))$, so we will have $\eta < \beta(t)^{-1}$ for all iterates, which suffices for the result to extend to exponential loss.

*where the residual $\boldsymbol{\rho}(t)$ is bounded and $\hat{\mathbf{w}}$ is the following $L_2$ max margin separator:*

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \text{ s.t. } y_n \mathbf{w}^\top \mathbf{x}_n \geq 1. \tag{4}$$

In Theorem 1 and in the remainder of the paper, *almost all* datasets refers to all datasets except a measure zero set of $\{\mathbf{x}_n\}_n$, e.g., with probability 1, any dataset sampled from an absolutely continuous distribution.

Interestingly, and somewhat surprisingly, Theorem 1 implies logarithmically slow convergence in direction to the $L_2$ max-margin separator. This slow convergence rate also applies to the margin. This, in contrast, is much slower compared to the rate of convergence of the loss $\mathcal{L}(\mathbf{w}(t))$ itself, which can be shown to decay as $O(1/t)$ (see Lemma 1 in Soudry et al. (2018b)).

**Multilayer linear networks.** In a recent follow up work, Gunasekar et al. (2018a) extend such results to fully connected deep linear networks, where the objective is non-convex. A multi-layer linear network consists of nodes arranged in $L$ layers. We use the convention that for an $L$ layer network, the inputs features $\mathbf{x}$ form the source nodes in the zeroth layer $l = 0$ and the output is sink node in the final layer $l = L$. Let $d_l$ for $l = 0, 1, \ldots, L$ denote the number of nodes in layer $l$. The network is parameterized by weight matrices $\mathcal{W} = \{\mathbf{W}_l \in \mathbb{R}^{d_{l-1}\times d_l} : l = 1, 2, \ldots, L\}$. Every such network represents a linear mapping given as follows:

$$\mathbf{w} = \mathcal{P}(\mathcal{W}) := \mathbf{W}_1 \cdot \mathbf{W}_2 \cdot \ldots \cdot \mathbf{W}_L \in \mathbb{R}^d.$$

Unlike logistic regression, where the parameters of the linear model $\mathbf{w} \in \mathbb{R}^d$ are learned directly by minimizing the training loss, in training linear networks, the objective is instead minimized over the parameters of the network $\mathcal{W} = \{\mathbf{W}_l \in \mathbb{R}^{d_{l-1}\times d_l} : l = 1, 2, \ldots, L\}$. The empirical loss is given by:

$$\mathcal{L}_{\mathcal{P}}(\mathcal{W}) = \mathcal{L}(\mathcal{P}(\mathcal{W})) = \sum_{n=1}^{N} \ell\left(y_n \langle \mathcal{P}(\mathcal{W}), \mathbf{x}_n\rangle\right). \tag{5}$$

Gradient descent iterates $\mathcal{W}(t) = \{\mathbf{W}_l(t)\}_{l=1}^L$ for the above objective are given by:

$$\forall l, \mathbf{W}_l(t+1) = \mathbf{W}_l(t) - \eta_t \nabla_{\mathbf{W}_l} \mathcal{L}_{\mathcal{P}}(\mathcal{W}(t)), \tag{6}$$

and the corresponding sequence of linear predictors along the gradient descent path is given by,

$$\mathbf{w}(t) = \mathcal{P}(\mathcal{W}(t)) = \mathbf{W}_1(t) \cdot \ldots \cdot \mathbf{W}_L(t) \in \mathbb{R}^d. \tag{7}$$

For the special case of exponential loss, Gunasekar et al. (2018a) showed that the linear separator $\mathbf{w}(t)$ in eq. (6) learned by gradient descent on fully connected network (under additional conditions on convergence of the net parameters and gradients, and convergence of the loss) again converges in the direction of the $L_2$ maximum-margin separator (Theorem 1 in Gunasekar et al. (2018a)). This result, however, only applies to exponential loss and does not specify how quickly the margin of $\mathbf{w}(t)$ converges to the maximum-margin (in case of convergence).

# 3 MAIN RESULTS

In this section, we provide a detailed analysis of the implicit bias in linear models focusing on convergence and rate of convergence of margin under general tails and with variable step sizes. We use the following standard notation on asymptotic behaviour: (a) $f(u) = \omega(g(u)) \Leftrightarrow \lim_{u\to\infty} \left|\frac{f(u)}{g(u)}\right| = \infty$, (b) $f(u) = o(g(u)) \Leftrightarrow \lim_{u\to\infty} \frac{f(u)}{g(u)} = 0$, (c) $f(u) = O(g(u)) \Leftrightarrow \limsup_{u\to\infty} \frac{|f(u)|}{g(u)} < \infty$, (d) $f(u) = \Omega(g(u)) \Leftrightarrow \lim\inf_{u\to\infty} \frac{f(u)}{g(u)} > 0$, and (e) $f(u) = \Theta(g(u)) \Leftrightarrow \Omega(g(u)) = f(u) = O(g(u))$.

Previous results, summarized in Section 2, show that when minimizing exponentially tailed losses on separable datasets, gradient descent converges to the $L_2$ max-margin separator with a very slow rate of $1/\log(t)$. While commonly used classification losses such as logistic loss, cross entropy loss, and exponential loss indeed have tight exponential tail, the significance of the exponential tail is not fully understood. What are the general conditions on the tail under which gradient descent converges to the maximum-margin solution? Can the rate of convergence be accelerated by choosing a heavier or lighter tail?

## 3.1 Linear networks with general tails

We first show that for a large family of strictly monotone losses with super–polynomial tails specified (Assumption 1 below), gradient descent iterates converge to the maximum-margin solution. We will later also analyze the rate of convergence for this family of loss functions.

**Assumption 1.** *$\ell(u)$ is analytic and satisfies the following:*

1. ***Strict monotonicity:*** *$\ell$ satisfies Definition 1. Since, $\forall u$, $\ell'(u) < 0$, let $\ell'(u) = -\exp(-f(u))$.*
2. ***Super-polynomial tail:*** *$\ell(u)$ has a "super-polynomial tail" if $\forall M > 0$, $\exists u_0$ such that $\forall u \geq u_0$, $-\ell'(u) \leq u^{-M}$. This is equivalent to $f(u) = \omega(\log(u))$.*
3. ***Asymptotically convex:*** *$\exists u_0$ such that $\forall u > u_0$, $\ell''(u) > 0$. For strictly monotone decreasing losses, this is equivalent to $\forall u > u_0$, $f'(u) = \frac{\ell''(u)}{-\ell'(u)} > 0$.*
4. ***Non-oscillatory tail:*** *$\lim_{u\to\infty} uf'(u)$ exists. For losses with super-polynomial tails where $f(u) = \omega(\log(u))$, this condition implies $f'(u) = \omega\left(u^{-1}\right)$.*

**Remark 1.** *Assumption 1 captures a large family strictly monotone losses with super-polynomial tails that are relevant for binary classification tasks, and the last condition is rather technical to avoid undesirable oscillatory behaviour* *like $f(u) = u + \sin(u)$. In particular, the assumption includes the following special cases:*

- *Logistic loss $\ell(u) = \log(1 + e^{-u})$, for which $f(u) = \log(1 + e^u) = \omega(\log(u))$ and $f'(u) = \frac{e^u}{1+e^u} = \omega\left(u^{-1}\right)$.*
- *Other losses with tight exponential tail (Definition 2), like the exponential loss $\ell(u) = \exp(-u)$.*
- *"Poly-exponential" tailed losses given by $\ell'(u) = -\exp(-u^\nu)$ for degree $\nu > 0$, e.g., the probit loss.*
- *Sub-exponential super-polynomial tails like $\ell'(u) = -u^{-\log^\mu(u)}$ for $\mu > 0$.*

For depth–$L$ linear networks, we first show that the implicit bias of gradient descent for exponential loss from Gunasekar et al. (2018a) can be extended more broadly to super-polynomial tailed losses specified in Assumption 1.

**Theorem 2.** *For any depth L, almost all linearly separable datasets, almost all initialization and any bounded sequence of step sizes $\{\eta_t\}$, consider the sequence $\mathcal{W}(t) = \{\mathbf{W}_l(t)\}_{l=1}^L$ of gradient descent updates in eq.* (6) *for minimizing the empirical loss $\mathcal{L}_\mathcal{P}(\mathcal{W})$ (eq.* (5)*) with a strictly monotone loss function $\ell$ satisfying Assumption 1, i.e.: $\ell'(u) = -\exp(-f(u)) < 0$, where asymptotically $f'(u) > 0$ and $f'(u) = \omega\left(u^{-1}\right)$.*

*If (a) $\mathcal{W}(t)$ minimizes the empirical loss, i.e. $\mathcal{L}_\mathcal{P}(\mathcal{W}(t)) \to 0$, (b) $\mathcal{W}(t)$, and consequently $\mathbf{w}(t) = \mathcal{P}(\mathbf{w}(t))$, converge in direction to yield a separator with positive margin, and (c) the gradients with respect to the linear predictors $\nabla_\mathbf{w}\mathcal{L}(\mathbf{w}(t))$ converge in direction, then the limit direction is given by,*

$$\bar{\mathbf{w}}_\infty = \lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|},$$

*where*

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \text{ s.t. } \mathbf{w}^\top \mathbf{x}_n \geq 1. \tag{8}$$

This theorem is proved in Appendix D, while the basic ideas are sketched in Appendix C for $L = 1$.

**Remark 2.** *Theorem 2 covers a large family of super-polynomial tails specified under Assumption 1. Conversely, for (sub) polynomial tails, we may not converge to the maximum-margin separator. In Appendix H, we show that we do not converge to the max margin if $\ell(u)$ has polynomial tail. Additionally, with the hinge loss (which it is neither differentiable or strictly monotonic) we generally do not converge to the maximum-margin without regularization, as then GD typically converges to a finite minimizer that depends on the initialization.*

**Remark 3.** *Rosset et al. (2004) also investigated the connection between the loss function choice and the maximum-margin solution. In this work, Rosset et al. (2004) considered linear models with monotone loss functions and explicit norm regularization. We discuss the connections between Rosset et al. (2004) results and ours in appendix A.*

**Remark 4.** *Gunasekar et al. (2018a) characterized the implicit bias of gradient descent for fully connected linear networks for the special case of exponential loss $\ell(u) = \exp(-u)$. Theorem 2 generalizes this characterization to a larger family of losses, which in particular includes the commonly used logistic loss. Logistic loss, despite having the same exponential tail as the exponential loss, was not explicitly analyzed in Gunasekar et al. (2018a).*

We now continue to characterizing the convergence rates.

### 3.2 Rates of convergence

To calculate the convergence rates we will make an additional assumption.

**Assumption 2.** *$f(u)$ is real analytic on $\mathbb{R}_{++}$ and satisfies $\forall k \in \mathbb{N} : \left|\frac{f^{(k+1)}(u)}{f'(u)}\right| = O\left(u^{-k}\right)$.*

While the above assumption is not required to show asymptotic convergence of gradient descent to the maximum-margin separator (Theorem 2), we do require the additional assumption to calculate the rates. This assumption implies that the loss tail does not decay too fast. In particular, Assumption 2 is *not* satisfied by super-poly-exponential tails like $\ell'(u) = \exp(-\exp(u^\nu))$ for $\nu > 0$ or $\ell'(u) = \exp(-\exp(\log^\mu(u)))$ for $\mu > 1$, and additionally avoids oscillatory functions like $\sin(u)$.

Nevertheless, a large class of interesting monotone functions satisfy this assumption, including cases where $f(u)$ is polynomial and poly-logarithmic functions. Within this family, we look at the margin rate of convergence of the gradient descent iterates, for $L = 1$ in two regimes:

1. $f'(u) = \omega(1)$, which implies $-\ell'(u) = \omega(\exp(-u))$. This case includes loss functions with tails *lighter* than the exponential tail, for example poly-exponential tail $\ell(u) = \exp(-u^\nu)$ with s strictly greater than one exponent, $\nu > 1$.
2. $f'(u) = \omega(u^{-1})$ and $f'(u) = o(1)$: or $-\ell'(u) = o(\exp(-u))$. This case includes loss functions with tails *heavier* than the exponential tail, such as $\ell(u) = \exp(-u^\nu)$ for $\nu < 1$ or $\ell(u) = \exp(-\log^\mu(u))$ for $\mu > 0$.

We first look at the rates for the special case of $L = 1$ where the parameters $\mathbf{w}$ of the linear models are directly learned using gradient descent. This is the setting analyzed in Soudry et al. (2018a) with tight exponential tailed losses. The following theorem is proved in Appendix F.

**Theorem 3.** *For almost all linearly separable datasets, almost all initialization, any bounded sequence of step sizes $\{\eta_t\} < 2\beta^{-1}$, and a single layer $L = 1$, consider the sequence of gradient descent updates in eq.* (2) *for minimizing the empirical loss $\mathcal{L}(\mathbf{w})$ (eq.* (1)*) with a strictly monotone $\beta$-smooth loss function $\ell$ satisfying*

*$\ell'(u) = -\exp(-f(u)) < 0$, where asymptotically $f'(u) = \Omega\left(\frac{1}{u}\log^{1+\epsilon}(u)\right)$ for some $\epsilon > 0$ and satisfies Assumption 2.*

*If (a) $\mathbf{w}(t)$ converges in direction to yield a separator with positive margin, and (b) the gradients with respect to the linear predictors $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}(t))$ converge in direction, then the margin convergence of $\mathbf{w}(t)$ to the max margin $\gamma = \max_{\mathbf{w}}\min_n \frac{\mathbf{w}^\top \mathbf{x}_n}{\|\mathbf{w}\|}$ satisfies:*

1. *If $f'(u) = \omega(1)$ (which implies $f(t) = \omega(t)$), then*

$$\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = O\left(\frac{1}{f^{-1}(\log(t))}\right).$$

2. *If $f'(u) = o(1)$ and $f$ is strictly concave, then*

$$\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \Omega\left(\frac{1}{\log(t)}\right)$$

*and the optimal rate is obtained for exponential loss.*

From the proof of Theorem 3, we can also calculate the rates of convergence for the normalized direction $\mathbf{w}(t)/\|\mathbf{w}(t)\|$ to the maximum-margin separator $\hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|$, as well as the convergence of the angle between them.

**Corollary 1.** *We examine super-polynomial tailed losses satisfying the assumptions of the previous Theorem, when the loss tail does not decay too fast, i.e. $\left|\frac{f'(u)}{f(u)}\right| = O(u^{-1})$. The optimal rate of convergence to the max margin of GD with fixed step size is $1/\log(t)$. This optimal rate is attained by exponentially tailed losses, where $f(u) = \Theta(u)$ (or $f'(u) = \Theta(1)$). This includes the popular losses of logistic loss and exponential loss.*

*Proof.* For the case of $f'(u) = \omega(1)$, $f(t) = \omega(t) \Rightarrow f^{-1}(t) = o(t)$ and thus, the rate for this case $O\left(\frac{1}{f^{-1}(\log(t))}\right)$ is sub-optimal compared with the rate for exponential loss which is $1/\log(t)$ (from Theorem 1). In appendix sections H.3, H.4 we give a positive example that demonstrates that this upper bound is tight, i.e., it is obtained for some datasets, and a negative example which shows a case in which the upper bound is not obtained. In general as long as the loss tail does not decay too fast, i.e. $\left|\frac{f'(u)}{f(u)}\right| = O(u^{-1})$, the rate in this case is $\Omega\left(\frac{1}{\log(t)}\right)$ (see appendix E.5). Secondly, for the case of $f'(u) = o(1)$ the asymptotic rate is $\Omega(1/\log(t))$, so the optimal rate we can hope for with any tail is $O(1/\log(t))$. In appendix F we show that the exponential tail obtains this optimal rate. Additionally, in Appendix J, we show that for the special case of poly-exponential losses $\ell'(u) = -\exp(-u^\nu)$ with $0.25 < \nu \leq 1$, the rate is indeed $O(1/\log(t))$ and the constants in the rates for $\nu < 1$ are strictly worse than that of exponential tail with $\nu = 1$. □

**Remark 5.** *Note that for $L = 1$ the optimization objective (eq. (5) ) is convex in the optimization variables and hence, by Lemma 1 in Soudry et al. (2018a), the assumption in Theorem 2 that $\mathcal{L}_P(\mathcal{W}(t)) \to 0$ is satisfied for appropriate choices of step size. Moreover for the special case of poly-exponential tails with $\ell'(u) = -\exp(-u^\nu)$ for $\nu > 0.25$, the convergence to the maximum-margin separator and the convergence rates can be obtained without the assumptions that $\mathbf{w}(t)$ and $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}(t))$ converge in direction (see Appendix J).*

Now we state the results for the general case of $L$–layer linear network.

**Theorem 4.** *Under assumption 2 and the conditions and notations of Theorem 2, if the SVM support vectors span the data then for any depth $L$ the network equivalent linear predictor $\mathbf{w}(t)$ satisfies:*

$$\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \begin{cases} O\left(\frac{1}{g(t)}\right), & f'(u) = \omega(1) \\ \Theta\left(\frac{1}{g(t)f'(g(t))}\right), & \textit{otherwise} \end{cases}$$

*where $g(t)$ is the asymptotic solution of*

$$\frac{dg(t)}{dt} = -\ell'(g(t))(g(t))^{2\left(1-L^{-1}\right)}. \qquad (9)$$

**Remark 6.** *Importantly, from Assumption 1, $-\ell'(u)$ has super-polynomial tail, which suggests the factor $(g(t))^{2\left(1-L^{-1}\right)}$ only negligibly affects the asymptotic solution of eq. (9). This implies that $\forall L > 1$, and even in the limit $L \to \infty$, the rate predicted by this Theorem 4 will only be slightly smaller than the $L = 1$ case of Theorem 3. This difference will become negligible in the limit $t \to \infty$. For example, for the case of exponential loss, we prove in appendix E.4 that the ODE solution is $g(t) = \log(t) + o(\log(t))$. Thus, in this case, the margin converges as $O(1/\log(t))$ for any depth.*

### 3.3 Faster rates using variable step sizes

Our analysis so far suggests that exponential tails have an optimal convergence rate, and for exponential tail losses with a bounded step size, we have an extremely slow rate of convergence, $O(1/\log t)$. Therefore, the question is can we somehow accelerate this rate using variable unbounded step sizes. Fortunately, at least for linear models trained with exponential loss, the answer is yes and we can indeed show faster rate of convergence by aggressively increasing the step size to compensate for the vanishing gradient. Specially, we examine the following normalized GD algorithm:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \frac{\nabla\mathcal{L}(\mathbf{w}(t))}{\|\nabla\mathcal{L}(\mathbf{w}(t))\|}. \qquad (10)$$

Recall that $\gamma = \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1}\min_n \mathbf{w}^\top \mathbf{x}_n$ is the maximum-margin of the dataset with unit $L_2$ norm separators, and without loss of generality assume $\forall n : \|\mathbf{x}_n\| \leq 1$.

By the triangle inequality, we have that $\|\nabla\mathcal{L}(\mathbf{w}(t))\| = \left\|\sum_n \exp(-\mathbf{w}(t)^\top \mathbf{x}_n)\mathbf{x}_n\right\| \leq \mathcal{L}(\mathbf{w}(t))$. We additionally have the following inequality for all $t$,

$$\begin{aligned}\|\nabla\mathcal{L}(\mathbf{w}(t))\| &= \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \sum_n \exp(-\mathbf{w}(t)^\top \mathbf{x}_n)\mathbf{w}^\top \mathbf{x}_n \\ &\geq \gamma \sum_n \exp(-\mathbf{w}(t)^\top \mathbf{x}_n) = \gamma\mathcal{L}(\mathbf{w}(t))\,.\end{aligned}$$

Thus, for all $\mathbf{w}$, the two-sided bound

$$\gamma\mathcal{L}(\mathbf{w}) \leq \|\nabla\mathcal{L}(\mathbf{w})\| \leq \mathcal{L}(\mathbf{w})$$

holds, and, up to a scaling of step–sizes, the normalized GD in eq. (10) can be alternatively expressed as the following

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \frac{\nabla\mathcal{L}(\mathbf{w}(t))}{\mathcal{L}(\mathbf{w}(t))}\,. \tag{11}$$

We chose to state our results in terms of eq. (11) (normalizing GD by $\mathcal{L}(\mathbf{w}(t))$) so that the stepsize choice $\eta_t$ does not depend on the optimal margin $\gamma$ which is unknown. The following theorem proved in Appendix B.1 shows that using normalized GD can improve the rate of convergence of the margin of the separator to $\log t/\sqrt{t}$ compared to $O(1/\log t)$ for fixed step sizes.

**Theorem 5.** *For any separable data set and any initial point* $\mathbf{w}(0)$*, consider the normalized GD updates in eq.* (11) *with a variable step size* $\eta_t = \frac{1}{\sqrt{t+1}}$ *and exponential loss* $\ell(u) = \exp(-u)$*.*

*Then the margin of the iterates* $\mathbf{w}(t)$ *converges to the max-margin* $\gamma$ *with rate* $t^{-1/2}\log t$*:*

$$\frac{\mathbf{w}(t+1)^T\mathbf{x}_n}{\|\mathbf{w}(t+1)\|} \geq \gamma - \frac{1+\log(t+1)}{\gamma(4\sqrt{t+2}-4)} - \frac{\log\mathcal{L}(\mathbf{w}(0))}{\gamma(2\sqrt{t+2}-2)}.$$

In the appendix we prove a more general version of Theorem 5, which obtains the same rate for any steepest descent algorithm. Also, note that normalized GD as in eq. (10) was analyzed before, but for other purposes. For example, Levy (2016) showed a stochastic version of it can better escape saddle points. Here we study the effect of normalization on the implicit bias of the algorithm.

The observation that aggressive changes in the step size can improve convergence rate is applied in the AdaBoost literature (Schapire and Freund, 2012), where exact line-search is used. We use a slightly less aggressive strategy of decaying step sizes with normalized gradient descent, attaining a rate of $\log(t)/\sqrt{(t)}$. This rate almost matches $1/\sqrt{(t)}$, which is the optimal rate in terms of margin suboptimality for solving hard margin SVM. This rate is achieved by the best known methods.[2] This suggests that gradient descent with a more aggressive step size policy is quite efficient at margin maximization.

We emphasize our goal here is not to develop a faster SVM optimizer, but rather to understand and improve gradient descent and local search in a way that might be applicable also for deep neural networks, as indicated by the numerical results we present next.

# 4 EXPERIMENTS WITH NORMALIZED GRADIENT DESCENT

In the following experiments, we implement the normalized GD in eq. (10) with step sizes separately tuned for each experiment.

## 4.1 Linear Networks on Synthetic Data

First, in Figure 1 we visualize the different rates for GD and normalized GD when training a plain logistic regression model on synthetic data. As expected from Theorem 5, we find that normalized GD converges significantly faster than unnormalized GD.

Additionally, we evaluate experimentally the convergence rates of GD and normalized GD for multi-layer linear models. Networks with $L \in \{1, 2, 3\}$ layers and 10 neurons per hidden layer are trained with GD and normalized GD on a synthetic binary classification dataset composed of 600 points, sampled from two normal distributions (one for each class).

We use a fixed learning rate $\eta = 5 \times 10^{-3}$ chosen through grid-search, and train each network for $5 \times 10^4$ total iterations. Figure 2 shows the margin gaps during training, with normalized GD providing faster convergence rates across models. Appendix I.2 provides details on data generation and training, along with results on ReLU networks.

## 4.2 Image Classification on MNIST

The MNIST dataset is composed of 70,000 grayscale images of 0-9 digits (10 classes total), each having $28 \times 28$ pixels. We use 10,000 images for testing and the rest for training and validation. Unlike harder datasets such as CIFAR-10 and CIFAR-100, MNIST provides a task where simple models can successfully separate the training examples. Hence, we train a 2-layer feedforward network with 5,000 hidden neurons and ReLU activations ($\text{ReLU}(x) = \max(0, x)$) with full-batch GD and normalized GD using the cross-entropy loss, for a total of 3,000 iterations. We decay the learning rate by a factor of 5 at $50\%, 75\%$ and $87.25\%$ of the total number of iterations.

We performed grid-search over initial learning rate values $\{0.1, 0.3, 0.5, 1.0, 2.5, 5.0\}$ using 5,000 images randomly chosen from the training set as validation, and $\eta = 1.0$

[2] The best known method in terms of margin suboptimality, and using vector operations (operations on all training examples), is the aggressive Perceptron, which achieves a rate of $\sqrt{N/t}$. Clarkson et al. (2012) obtained an improved method which they showed is optimal, that does not use vector operations. Clarkson et al. (2012) method achieves a rate of $\sqrt{(N+d)/t}$, where now $t$ is the number of scalar operations.

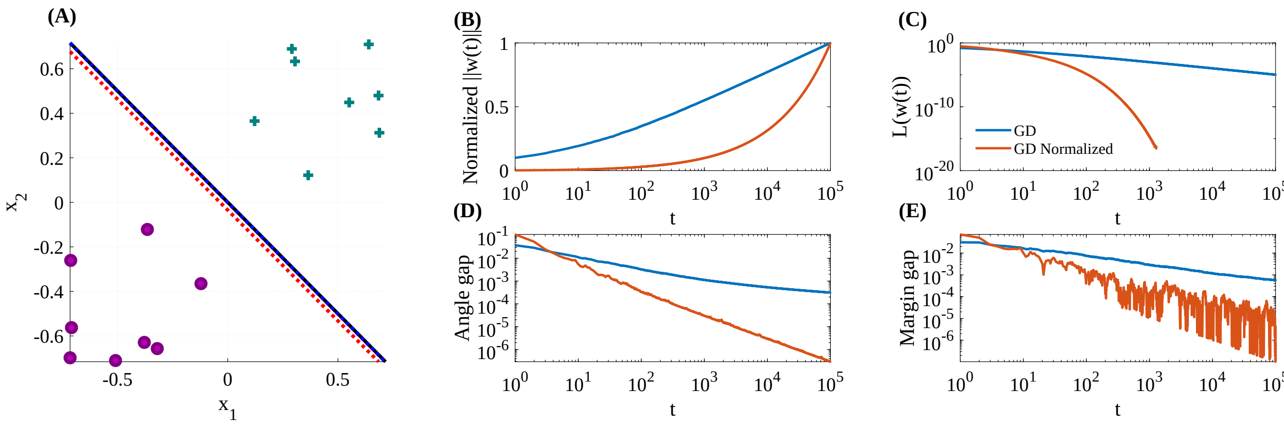



Figure 1: Visualization of the convergence of GD in comparison to normalized GD in a synthetic logistic regression dataset in which the $L_2$ max margin vector $\hat{\mathbf{w}}$ is precisely known. **(A)** The dataset (positive and negatives samples ($y = \pm 1$) are respectively denoted by $'+'$ and $'\circ'$), max margin separating hyperplane (black line), and the solution of GD (dashed red) and normalized GD (dashed blue) after $10^5$ iterations. For both GD and Normalized GD, we show: **(B)** The norm of $\mathbf{w}\,(t)$, normalized so it would equal to 1 at the last iteration, to facilitate comparison; **(C)** The training loss; and **(D&E)** the angle and margin gap of $\mathbf{w}\,(t)$ from $\hat{\mathbf{w}}$. As can be seen in panels **(C-E)**, normalized GD converges to the max-margin separator significantly faster, as expected from our results. More details are given in appendix I.1.

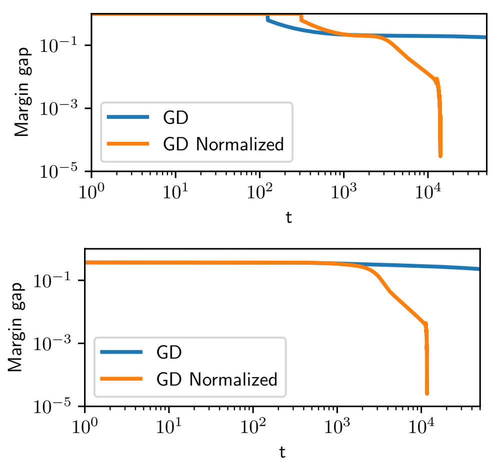



Figure 2: Margin convergence plots for 2 (top) and 3 (bottom) layered linear networks on synthetic clustered data, trained with GD and normalized GD — the latter provides significantly faster convergence.

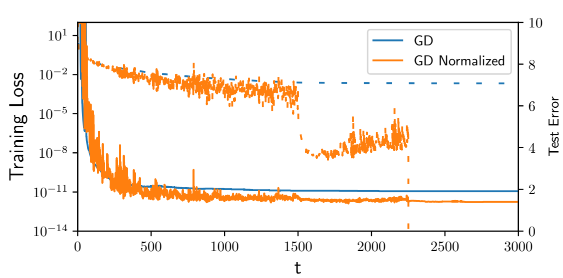



Figure 3: MNIST digit classification with a 2-layer feedforward neural network. Training loss (dashed lines) stagnates with GD once gradients become small, while normalized GD keeps making progress. Normalized GD also achieves lower test error (solid lines).

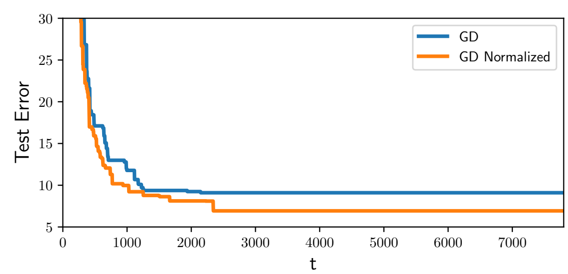



Figure 4: Test performance of a Wide ResNet 28-4 on CIFAR-10, with $\eta = 2.0$, where normalized GD outperforms GD by absolute 2.17%. We plot 'best yet' test error: the lowest error seen up to iteration $t$. Unlike curves reported in Zagoruyko and Komodakis (2016), progress stops early in training: there is no change in the 'best yet' test error after $t = 2350$, even with the decays in learning rate. This suggests that regularization and/or momentum might be required to achieve better results.

yielded better results for both GD and normalized GD. We use no regularization nor data augmentation, since our goal is to observe the contrast between GD and normalized GD as the training loss decreases and gradients become small. Figure 3 shows the training loss and test error at each iteration $t$: while the training loss stagnates early for GD, normalized GD keeps decreasing it. Normalized GD also reaches lower test error: 1.4% compared to 1.91%.

### 4.3 Image Classification on CIFAR-10

The CIFAR-10 dataset (Krizhevsky, 2009) consists of 60,000 colored $32 \times 32$ images belonging to one of 10 possible classes, and is split into 50,000 training and 10,000 test points. The goal of this experiment is to evaluate whether normalized GD can provide advantages for training complex models on more realistic tasks, when using the standard cross-entropy loss.

For that, we train a Wide ResNet 28-4 (Zagoruyko and Komodakis, 2016), a 28-layer convolutional neural network with residual connections and a total of 5.8M parameters. This architecture is capable of reaching less than $4\%$ test error on CIFAR-10 given more features per convolutional layer, making Wide ResNets a strong model baseline to compare the benefits of normalized GD against GD. Following Zagoruyko and Komodakis (2016), we pre-process the dataset by performing channel-wise normalization on each image using statistics computed from the training set. Horizontal flips and random crops are used during training for data-augmentation. We also follow the same learning rate schedule, decaying it by a factor of 5 at $30\%$, $60\%$ and $80\%$ of the total iterations.

To select a learning rate for each method, we train the network for 3,000 iterations with $\eta \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$. Both methods performed better on a validation set of 5000 images with $\eta = 2.0$. Figure 4 shows the test performance when training the model for 7,800 iterations with $\eta = 2.0$, where normalized GD achieves $6.93\%$ test error, while GD yields $9.90\%$.

Note that, while normalized GD outperformed GD in this full-batch setting, its performance is still subpar when compared to the standard optimization for Wide ResNets, which includes SGD with Nesterov momentum and weight decay. To confirm whether momentum and weight decay can have strong positive impacts in a model's performance, we also trained a Wide ResNet 28-4 using SGD, with and without momentum/weight decay. We observed that removing momentum and weight decay resulted in a test error increase from $4.45\%$ to $7.75\%$ (larger error than normalized GD). This suggests an importance in reconciling weight decay, momentum and gradient normalization.

# 5 DISCUSSION

In this work, we have examined the behavior of gradient descent on separable data, in binary linear classification tasks. First, in Theorem 2 we proved the linear classifier resulting from a multilayer linear neural networks converges in direction to the $L_2$ max-margin on almost all linearly separable data — for a wide family of monotone, convex loss functions with super-exponential tails and some technical conditions (Assumption 1). In contrast, polynomially tailed loss function do not lead to convergence to the max-margin. Intuitvely, the reason behind this is that for super-polynomial loss functions the datapoints with the largest margin (i.e., the support vectors) become dominant in the gradient, while for polynomial or heavier tails the contribution of non-support vectors is never negligible.

Next, we examine the convergence rate for a linear classifier with loss within this wide family of loss functions. We prove in Theorem 3 that the exponential tail has the optimal rate. This offers a possible explanation to the empirical preference of the exponentially-tailed loss functions over other losses (e.g. the probit loss): that the exponential loss leads to a faster convergence to the asymptotic (implicitly biased) solution, as we showed here. This result is somewhat surprising, and we do not have an intuitive explanation why this should be true.

In Theorem 4, we extend these results to multilayer linear neural networks, and show similar convergence rates, with only a negligible decrease in the rate with the depth — even when the number of layers is infinite. Note that in this Theorem we already assume convergence of the loss to zero. However, if we do converge, it is somewhat surprising that this rate does not depend much on the depth, as one might expect to have convergence rate issues due to exploding or vanishing gradients.

In Theorem 5 we showed that the convergence of GD for an exponential loss function could be significantly accelerated by simply increasing the learning rate. In fact, GD can also approximate the regularization path in the following sense. Let $R = \|\mathbf{w}_t\|$, and $\mathbf{w}_R = \arg\min_{\|\mathbf{w}\| \leq R} \mathcal{L}(\mathbf{w})$. Then

$$\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}_R) \leq \mathcal{L}(\mathbf{w}(0)) \exp(-c\gamma^2 t) \,. \qquad (12)$$

As a simple implication of this, the normalized GD path starting at $\mathbf{w}_0 = 0$ has $\mathcal{L}(\mathbf{w}(0)) = n$, so after $t \geq \log(n/\epsilon)/\gamma^2$ steps the loss achieved by $\mathbf{w}_t$ is $\epsilon$ close to the best predictor of the same norm. This shows that GD is closely approximating the regularization path.

Finally, we show numerically that normalized GD can significantly improve the convergence speed of GD on synthetic datasets for linear predictors (Figure 1), linear multilayer networks (Figure 2), and even non-linear ReLU multilayer networks (Appendix I.2). Additionally, we show normalized GD can improve the results of GD on standard datasets such as MNIST (by $0.5\%$) and CIFAR-10 (by $3\%$). However, a gap remains from achieving state of the art results. Our experiments indicate the origin of this gap is the use of weight decay and momentum (which are outside the scope of this paper). This suggests that reconciling regularization, momentum and gradient normalization might be of particular interest for future work, possibly reducing the gap between mini-batch and full-batch training.

Recent work explore extensions of the implicit bias result for linear models to non-strictly-separable datasets (Ji and Telgarsky, 2018) and to stochastic gradient descent (Ji and Telgarsky, 2018; Nacson et al., 2018; Xu et al., 2018). It remains to be seen if the results of this work could be also extended to such settings. Additionally, combining our results with the results of a parallel work, Ji and Telgarsky (2019), might enable us to weaken some of the assumptions in this paper. We discuss Ji and Telgarsky (2019) work in appendix A.

**Acknowledgements**

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

**References**

Kenneth L. Clarkson, Elad Hazan, and David P. Woodruff. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):23, 2012.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018a.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint*, 2018b.

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

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

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Kfir Y. Levy. The Power of Normalization: Faster Evasion of Saddle Points. *arXiv*, nov 2016.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796*, 2018.

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

Saharon Rosset, Ji Zhu, and Trevor J Hastie. Margin maximizing loss functions. In *Advances in neural information processing systems*, pages 1237–1244, 2004.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

Daniel Soudry, Elad Hoffer, , Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. *ICLR*, 2018a.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data (journal version). *arXiv preprint: 1710.10345v3*, 2018b.

Matus Telgarsky. Margins, shrinkage and boosting. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pages II–307. JMLR. org, 2013.

Willie Wong. Asymptotic solution for a first order ode. MathOverflow, 2018. URL `https://mathoverflow.net/q/309520`. URL:https://mathoverflow.net/q/309520 (version: 2018-08-31).

Tengyu Xu, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Convergence of sgd in learning relu models with separable data. *arXiv preprint arXiv:1806.04339*, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

# Appendix

## A Additional Related Work

1. Rosset et al. (2004) investigated the connection between the loss function choice and the maximum-margin solution. They considered linear models with monotone loss functions and explicit norm regularization. The authors examined the solutions of the regularized loss function

$$\mathbf{w}(\lambda) = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|_p^p = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n=1}^{N} \ell\left(y_n \mathbf{w}^\top \mathbf{x}_n\right) + \lambda \|\mathbf{w}\|_p^p$$

for some $p$ as the regularization vanishes, i.e., $\lambda \to 0$. We focus here on the euclidean norm, meaning $p = 2$. In this case, they proved that if $\exists T$ (possibly $T = \infty$) so that

$$\forall \epsilon > 0 : \lim_{t \to T} \frac{\ell(t \cdot (1-\epsilon))}{\ell(t)} = \infty\,, \tag{13}$$

then $\mathbf{w}(\lambda)$ converge into the direction of the maximum-margin solution as $\lambda \to 0$. Note, that since in this paper we assume that the loss is strictly decreasing this implies that the last equation can only be satisfied with $T \to \infty$.

We examine the main differences and similarities between Rosset et al. (2004) results and ours.

**Main differences in settings:** In our work, we examine the convergence of GD iterates and its implicit bias, while Rosset et al. (2004) focused on the explicit bias in the limit of the regularization path, as the regularization vanishes. Thus, in our results we take into account the optimization dynamics. In addition, we examine linear fully connected networks while Rosset et al. (2004) only considered linear models.

**Relation between Rosset et al. (2004) and our results:** In eq. (13), Rosset et al. (2004) states a condition on the loss function that guarantees convergence to the maximum-margin separator. In our Theorem 2, the key assumption on the loss function is that $\ell'(u) = -\exp(-f(u)) < 0$ and $f(u) = \omega(u^{-1})$ (Assumption 1). Rosset et al. (2004) condition is weaker than our assumption because $f(u) = \omega(u^{-1})$ implies that $\forall \epsilon > 0 : \lim_{t\to\infty} \exp\left(f(u) - f((1-\epsilon)u)\right) = \infty$ (Lemma 3). This also implies that $\forall \epsilon > 0 : \lim_{t\to\infty} \frac{\ell'(t\cdot(1-\epsilon))}{\ell'(t)} = \infty$. Using L'Hospital's rule we have that $\forall \epsilon \in (0,1)$ : $\lim_{t\to\infty} \frac{\ell(t\cdot(1-\epsilon))}{\ell(t)} = \lim_{t\to\infty} (1-\epsilon) \frac{\ell'(t\cdot(1-\epsilon))}{\ell'(t)} = \infty$. Thus, our assumption 1 implies the assumption in eq. (13). It is still unclear if the opposite direction is also true, *i.e.*, if q. (13) implies Assumption 1.

Furthermore, in Theorem 2 we also make additional assumptions on the the convergence of GD iterate and its gradients. These additional assumption are required in our analyses since analyzing the optimization dynamics in opposed to examining the regularization path limit poses additional technical challenges.

2. After the submission of this paper to AISTATS, another related work appeared Ji and Telgarsky (2019). Ji and Telgarsky (2019) consider fully connected linear networks, separable dataset and strictly decreasing loss functions which are $\beta$-smooth and $G$-Lipschitz. They show that for GD with particular decreasing step sizes and mild assumptions on the initialization, the loss converges to zero. They connect this result to an alignment phenomenon between different layers. In addition, for the logistic loss and under the additional assumption that the SVM support vectors span the all space $\mathbb{R}^d$, they show that the network equivalent linear predictor converge in the direction of the maximum-margin separator. It remains to be seen if combining our results and Ji and Telgarsky (2019), we can weaken the assumptions we required to prove convergence rates for linear neural nets.

## B Adaptive Learning Rate

For learning linear models with exponential loss, Gunasekar et al. (2018b) provide an alternative proof for convergence to max-margin solution when using gradient descent. This result also generalized the characterization of implicit bias for general steepest descent algorithm. While Gunasekar et al. (2018b) do not state a rate of convergence, the technique can be used to establish that the margin converges at the rate of $O(1/\log t)$ as summarized in the following theorem (specialized here only for gradient descent):

**Theorem 6.** *For any separable data set, any initial point* $\mathbf{w}(0)$*, consider gradient descent iterates with a fixed step size* $\eta < \frac{1}{\mathcal{L}(\mathbf{w}(0))}$ *for linear classification with the exponential loss* $\ell(u) = \exp(-u)$*.*
*Then the iterates* $\mathbf{w}(t)$ *satisfy:*

$$\min_n \frac{\mathbf{w}(t)^\top \mathbf{x}_n}{\|\mathbf{w}(t)\|} = \gamma - O\Big(\frac{1}{\log t}\Big)\,,$$

*where* $\gamma = \max_{\mathbf{w}} \min_n \frac{\mathbf{w}^\top \mathbf{x}_n}{\|\mathbf{w}\|} = \frac{1}{\|\hat{\mathbf{w}}\|}$ *is the maximum-margin .*

Note that Theorem 6 ensures the rate of convergence of the margin, but does not specify how quickly $\mathbf{w}(t)$ itself converges to the max-margin predictor $\hat{\mathbf{w}}$.

### B.1 Proof for Theorems 6 and 5

In this section we prove extended versions of Theorems 6 and 5. In this section only, the norm $\|\cdot\|$ is a general norm (not the $L_2$ norm, like in the rest of the paper). First, we state definitions and auxiliary results.

The following lemma is a standard result in convex analysis.

**Lemma 1** (Fenchel Duality)**.** *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, and* $f, g$ *be two closed convex functions. Then*

$$\max_{\mathbf{w}} -f^*(-\mathbf{A}\mathbf{w}) - g^*(\mathbf{w}) \leq \min_{\mathbf{r}} f(\mathbf{r}) + g(\mathbf{A}^\top \mathbf{r}). \tag{14}$$

Let $\mathbf{X} \in \mathbb{R}^{d\times N}$ be the data matrix and without loss of generality $\|\mathbf{x}_n\|_* \leq 1$. Define $\{\mathbf{e}_n \in \mathbb{R}^N\}_{n=1}^N$ denote the standard basis in $\mathbb{R}^d$ and the $\|\cdot\|$–margin as $\gamma = \max_{\|\mathbf{w}\|=1} \min_{n\in[N]} \mathbf{e}_n^\top \mathbf{X}^\top \mathbf{w}$.

In the first auxillary result, we wish to show that $\|\nabla\mathcal{L}(\mathbf{w})\|_* \geq \gamma\mathcal{L}(\mathbf{w})$ for all $\mathbf{w}$, which is an analog of the Polyak condition. Define $r_n(\mathbf{w}) = \exp(-\mathbf{w}^T\mathbf{x}_n)$ and let $\mathbf{r}(w) = [r_n(\mathbf{w})]_{n=1}^N \in \mathbb{R}^N$ denote the vector formed by stacking $r_n(\mathbf{w})$. By noting that $\mathcal{L}(\mathbf{w}) = \|\mathbf{r}(\mathbf{w})\|_1$ and $\nabla\mathcal{L}(\mathbf{w}) = -\mathbf{X}\mathbf{r}(\mathbf{w})$, this can be restated as $\frac{\|\mathbf{X}\mathbf{r}(\mathbf{w})\|_*}{\|\mathbf{r}(\mathbf{w})\|_1} \geq \gamma$. Since we require this for all $\mathbf{w}$, with $r_n(\mathbf{w}) \geq 0$, and norms are homogeneous, this condition follows from showing that

$$\min_{\mathbf{r}\in\Delta_N} \|\mathbf{X}\mathbf{r}\|_* \geq \gamma, \tag{15}$$

where $\Delta_N$ is the $N$-dimensional probability simplex.

**Lemma 2.** *The following duality holds for all* $\mathbf{X}$*:*

$$\min_{\mathbf{r}\in\Delta_N} \|\mathbf{X}\mathbf{r}\|_* \geq \max_{\|\mathbf{w}\|=1} \min_{n\in[N]} \mathbf{e}_n^\top \mathbf{X}^\top \mathbf{w} = \gamma. \tag{16}$$

*This in turn implies that for all* $\mathbf{w} \in \mathbb{R}^d$*,* $\|\nabla\mathcal{L}(\mathbf{w})\|_* \geq \gamma\mathcal{L}(\mathbf{w})$*.*

*Proof.* Let $f(\mathbf{r}) = \mathbf{1}_{\mathbf{r}\in\Delta_N}$ and $g(\mathbf{z}) = \|\mathbf{z}\|_*$. Thus $\min_{\mathbf{r}\in\Delta_N} \|\mathbf{X}\mathbf{r}\|_* = f(\mathbf{r}) + g(\mathbf{X}\mathbf{r})$. The conjugates are $f^*(\mathbf{w}) = \max_{\mathbf{z}\in\Delta_N} \mathbf{w}^\top\mathbf{z} = \max_n \mathbf{e}_n^\top \mathbf{w}$, and $g^*(\mathbf{w}) = \mathbf{1}_{\|\mathbf{w}\|\leq 1}$. The LHS of Lemma 1 is

$$\max_{\mathbf{w}} \big( - f^*(-\mathbf{X}^\top\mathbf{w}) - g^*(\mathbf{w})\big) = \max_{\mathbf{w}} \big( - \max_n -\mathbf{e}_n^\top \mathbf{X}^\top \mathbf{w} - \mathbf{1}_{\|\mathbf{w}\|\leq 1}\big) \tag{17}$$

$$= \max_{\|\mathbf{w}\|\leq 1} - \max_i -\mathbf{e}_n^\top \mathbf{X}^\top \mathbf{w} \tag{18}$$

$$= \max_{\|\mathbf{w}\|\leq 1} \min_i \mathbf{e}_n^\top \mathbf{X}^\top \mathbf{w} = \gamma. \tag{19}$$

Thus the LHS is equal to $\gamma$, since it is precisely the optimization program of $\|\cdot\|$ -SVM. By weak duality (Lemma 1), we have shown that $\min_{\mathbf{r}\in\Delta_N} \|\mathbf{X}\mathbf{r}\|_* \geq \gamma$. □

Using this lower bound we proceed with the optimization analysis which largely follows the standard arguments from the optimization literature on first-order methods and the analysis extends the proof by Telgarsky (2013), where a similar result was derived for the case of $L_1$ margin using AdaBoost (coordinate descent). We prove the theorems for general steepest descent algorithms which includes gradient descent as a special case.

### B.2 Proof of Theorem 6

Consider the steepest descent algorithm described by the updates below:

$$\begin{aligned}\mathbf{w}(t+1) &= \mathbf{w}(t) - \eta\gamma_t\Delta\mathbf{w}(t)\\ \Delta\mathbf{w}(t)^T\nabla\mathcal{L}(\mathbf{w}(t)) &= \|\nabla\mathcal{L}(\mathbf{w}(t))\|_*\\ \|\Delta\mathbf{w}(t)\| &= 1\\ \gamma_t &\triangleq \|\nabla\mathcal{L}(\mathbf{w}(t))\|_*\,.\end{aligned} \tag{20}$$

It is easy to check that when $\|.\|$ is the $L_2$ norm, steepest descent above is simply gradient descent.

Next, we prove the generalized version of Theorem 6, which applies to steepest descent (instead of just gradient descent). For fixed step sizes, the proof is essentially same as in the proof of Theorem 5 in Gunasekar et al. (2018b) where in the end we add the computation of rates.

**Theorem 7. [Generalized Theorem 6]** *For any separable data set and any initial point* $\mathbf{w}(0)$*, consider steepest descent updates with a fixed step size* $\eta$ *and exponential loss* $\ell(u) = \exp(-u)$*.*

*Let us assume that* $\|\mathbf{x}_n\|_* \leq 1$*. If the step size* $\eta \leq \frac{1}{\mathcal{L}(\mathbf{w}(0))}$ *, then the iterates* $\mathbf{w}(t)$ *satisfy:*

$$\min_n \frac{\mathbf{w}(t)^\top\mathbf{x}_n}{\|\mathbf{w}(t)\|} = \max_{\mathbf{w}}\min_n \frac{\mathbf{w}^\top\mathbf{x}_n}{\|\mathbf{w}\|} - O\Big(\frac{1}{\log t}\Big)$$

*In particular, if there is a unique maximum-*$\|.\|$ *margin solution* $\mathbf{w}^\star_{\|.\|} = \arg\max_{\mathbf{w}}\min_n \frac{\mathbf{w}^\top\mathbf{x}_n}{\|\mathbf{w}\|}$*, then* $\lim_{t\to\infty}\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \mathbf{w}^\star_{\|.\|}$*.*

By Lemma 12 of Gunasekar et al. (2018b), $\mathcal{L}(\mathbf{w}(t) - \eta\gamma_t\Delta\mathbf{w}(t)) \leq \mathcal{L}(\mathbf{w}(t))$ for any $\eta < \frac{1}{\mathcal{L}(w(0))}$. Then, we note the following:

$$\mathbf{v}^\top\nabla^2\mathcal{L}(\mathbf{w})\mathbf{v} = \sum_n r_n(\mathbf{x}_n^\top\mathbf{v})^2 \leq \sum_n r_n\|\mathbf{x}_n\|_*^2\|\mathbf{v}\|^2 \leq \mathcal{L}(\mathbf{w})\max_n\|\mathbf{x}_n\|_*^2\|\mathbf{v}\|^2 \leq \mathcal{L}(\mathbf{w})\|\mathbf{v}\|^2\,,$$

where the last inequality follows since we assumed without loss of generality $\|\mathbf{x}_n\| \leq 1$.

Now using Taylor's theorem along with the steepest descent updates in eq. (20) gives the following

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(t) - \eta\gamma_t\|\nabla\mathcal{L}(\mathbf{w}(t)\|_* + \frac{1}{2}\eta^2\gamma_t^2\max_{r\in(0,1)}\Delta\mathbf{w}(t)^\top\nabla^2\mathcal{L}(\mathbf{w}(t) - r\eta\gamma_t\Delta\mathbf{w}(t))\Delta\mathbf{w}(t) \tag{21}$$

$$\leq \mathcal{L}(\mathbf{w}(t)) - \eta\gamma_t^2 + \frac{1}{2}\eta^2\gamma_t^2\mathcal{L}(\mathbf{w}(t)) \tag{22}$$

$$\leq \mathcal{L}(\mathbf{w}(t))\left(1 - \eta\frac{\gamma_t^2}{\mathcal{L}(\mathbf{w}(t))} + \frac{1}{2}\eta^2\gamma_t^2\right) \tag{23}$$

$$\leq \mathcal{L}(\mathbf{w}(t))\exp\left(-\eta\frac{\gamma_t^2}{\mathcal{L}(\mathbf{w}(t))} + \frac{1}{2}\eta^2\gamma_t^2\right). \tag{24}$$

Recursing gives

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(0))\exp\left(-\eta\left(\sum_{i=0}^t\frac{\gamma_i^2}{\mathcal{L}(\mathbf{w}_i)} + \frac{1}{2}\eta\gamma_i^2\right)\right). \tag{25}$$

We show convergence of margin the following steps

Step I. *Lower bound on the un-normalized margin.*

$$\max_{n\in[N]}\exp(-\mathbf{w}(t+1)^T x_n) \leq \mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(0))\exp\Big(-\eta\big(\sum_{i=0}^t\frac{\gamma_i^2}{\mathcal{L}(\mathbf{w}_i)} + \frac{1}{2}\eta\gamma_i^2\big)\Big). \tag{26}$$

By applying $-\log$,

$$\min_{j\in[n]}\mathbf{w}(t+1)^\top\mathbf{x}_n \geq \eta\sum_{i=0}^t\frac{\gamma_i^2}{\mathcal{L}(\mathbf{w}_{(i)})} - \frac{1}{2}\eta^2\sum_{i=0}^t\gamma_i^2 - \log\mathcal{L}(\mathbf{w}_0). \tag{27}$$

Step II. *Upper bound on norm* $\|\mathbf{w}(t+1)\|$.

$$\|\mathbf{w}(t+1)\| = \left\|\sum_{i=0}^{t} \eta\gamma_i \Delta\mathbf{w}(i)\right\| \leq \eta \sum_{i=0}^{t} \gamma_i. \tag{28}$$

Step III. *Convergence of margin.* For every $n \in [N]$, from the above steps we have that

$$\frac{\mathbf{w}(t+1)^\top \mathbf{x}_n}{\|\mathbf{w}(t+1)\|} = \frac{\sum_{i=0}^{t} \frac{\gamma_i^2}{\mathcal{L}(\mathbf{w}_{(i)})}}{\sum_{i=0}^{t} \gamma_i} - \frac{\eta}{2} \frac{\sum_{i=0}^{t} \gamma_i^2}{\sum_{i=0}^{t} \gamma_i} - \frac{\log \mathcal{L}(\mathbf{w}_0)}{\eta \sum_{i=0}^{t} \gamma_i}. \tag{29}$$

Use that $\gamma_i \geq \gamma \mathcal{L}(\mathbf{w}_i)$ by the duality result in Lemma 2,

$$\frac{\mathbf{w}(t+1)^\top \mathbf{x}_n}{\|\mathbf{w}(t+1)\|} \geq \gamma - \frac{\eta}{2} \frac{\sum_{i=0}^{t} \gamma_i^2}{\sum_{i=0}^{t} \gamma_i} - \frac{\log \mathcal{L}(\mathbf{w}_0)}{\eta \sum_{i=0}^{t} \gamma_i} \tag{30}$$

In order to prove the rest of the Theroem, we show that (a) $\sum_{i=0}^{t} \gamma_i^2 < \infty$, and (b) $\sum_{i=0}^{t} \gamma_i = \Omega(\log t)$

(a) Proof that $\sum_{i=0}^{t} \gamma_i^2 < \infty$. From eq. (22),

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(t)) - \eta\gamma_t^2 + \frac{1}{2}\eta^2\gamma_t^2 \mathcal{L}(\mathbf{w}(0)) \leq \mathcal{L}(\mathbf{w}(t)) - \frac{\eta}{2}\gamma_t^2, \tag{31}$$

where the last inequality follows since $\eta \leq \frac{1}{\mathcal{L}(\mathbf{w}(0))}$. Now using telescoping sum gives the following for all $t > 0$,

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(0)) - \frac{\eta}{2} \sum_{i=0}^{t} \gamma_i^2 \tag{32}$$

$$\implies \sum_{i=0}^{t} \gamma_i^2 \leq \frac{\mathcal{L}(\mathbf{w}(0)) - \mathcal{L}(\mathbf{w}(t+1))}{\eta/2} \leq \frac{\mathcal{L}(\mathbf{w}(0))}{\eta/2} < \infty. \tag{33}$$

(b) Next we show that $\eta \sum_{i=0}^{t} \gamma_i = \Omega(\log t)$. From eq. (22) again,

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(t)) - \eta\gamma_t^2 + \frac{1}{2}\eta^2\gamma_t^2 \mathcal{L}(\mathbf{w}(0)). \tag{34}$$

Since we chose $\eta < \frac{1}{\mathcal{L}(\mathbf{w}(0))}$ and $\gamma_t \geq \gamma \mathcal{L}(\mathbf{w}(t))$ from Lemma 2, we have

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(t)) - \frac{\eta}{2}\gamma_t^2 \leq \mathcal{L}(\mathbf{w}(t)) - \frac{\eta}{2}\gamma^2 \mathcal{L}(\mathbf{w}(t))^2. \tag{35}$$

The following claim is proved at the end of this section.

**Claim 1** (Solve Recursion)**.** *For a non-negative sequence* $\{a_t\}_t$*, the recursion* $a_{t+1} \leq a_t - c^2 a_t^2$ *implies*

$$a_{t+1} \leq \frac{1}{(t+1)c^2/(1-c^2a_0) + 1/a_0}. \tag{36}$$

We use Claim 1 with $c^2 = \frac{\eta}{2}\gamma^2$ and $a_t = \mathcal{L}(\mathbf{w}(t))$. Note that $c^2 a_0 \frac{\eta}{2}\gamma^2 \mathcal{L}(\mathbf{w}(0)) \leq \frac{1}{2}$, since $\eta < \frac{1}{\mathcal{L}(\mathbf{w}(0))}$ and $\gamma = \max_{\|\mathbf{w}\|\leq 1} \min_n {\mathbf{x}_n}^\top \mathbf{w} \leq \max_n \|\mathbf{x}_n\|_* \leq 1$, we have $2c^2 \geq \frac{c^2}{1-c^2a_0} > 0$. Thus,

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \frac{1}{\eta\gamma^2(t+1) + 1/\mathcal{L}(\mathbf{w}(0))} \leq \frac{1}{\eta\gamma^2(t+1)} \triangleq q(t+1). \tag{37}$$

We then lower bound $\|\mathbf{w}_{t+1}\|$. Since $\|\mathbf{x}_n\|_* \leq 1$, then

$$q(t) \geq \mathcal{L}(\mathbf{w}(t)) \geq \exp(-\mathbf{w}(t)^\top \mathbf{x}_n) \tag{38}$$

$$\implies \log \frac{1}{q(t)} \leq \mathbf{w}(t)^\top \mathbf{x}_n \leq \|\mathbf{w}(t)\|. \tag{39}$$

From eq. (28),

$$\eta \sum_{i=0}^{t} \gamma_i \geq \|\mathbf{w}(t+1)\| \geq \log \frac{1}{q(t+1)} = \log(\eta\gamma^2(t+1)). \tag{40}$$

Putting together the inequalities from eqs. (30), (33), and (40)

$$\frac{\mathbf{w}(t+1)^\top \mathbf{x}_n}{\|\mathbf{w}(t+1)\|} \geq \gamma - \frac{\eta}{2}\frac{\sum_{i=0}^t \gamma_i^2}{\sum_{i=0}^t \gamma_i} - \frac{\log \mathcal{L}(\mathbf{w}_0)}{\eta \sum_{i=0}^t \gamma_i} \tag{41}$$

$$\geq \gamma - \frac{\eta \mathcal{L}(\mathbf{w}(0)) + \log \mathcal{L}(\mathbf{w}(0))}{\log\left(\eta\gamma^2(t+1)\right)} \tag{42}$$

This completes the proof of Theorem 6, the proof of intermediate Claim 1 is given below.

*Proof of Claim 1.* For a non-negative decreasing sequence satisfying, $a_{t+1} \leq a_t - c^2 a_t^2$, by inversion we have

$$\frac{1}{a_{t+1}} \geq \frac{1}{a_t(1-c^2 a_t)} = \frac{1}{a_t} + \frac{c^2}{1-c^2 a_t} \geq \frac{1}{a_t} + \frac{c^2}{1-c^2 a_0}. \tag{43}$$

Suming from $i = 0, \ldots, t$,

$$\frac{1}{a_{t+1}} \geq \frac{1}{a_0} + (t+1)\frac{c^2}{1-c^2 a_0} \implies a_{t+1} \leq \frac{1}{1/a_0 + (t+1)c^2/(1-c^2 a_0)}. \tag{44}$$

□

## B.3 Proof of Theorem 5

We not look at the steepest descent with varying step sizes algorithm:

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta_t \gamma_t \mathbf{p}_t \tag{45}$$

$$\mathbf{p}_t^\top \nabla \mathcal{L}(\mathbf{w}(t)) = \|\nabla \mathcal{L}(\mathbf{w}(t))\|_* \tag{46}$$

$$\|\mathbf{p}_t\| = 1 \tag{47}$$

$$\gamma_t \triangleq \frac{\|\nabla \mathcal{L}(\mathbf{w}(t))\|_*}{\mathcal{L}(\mathbf{w}(t))}. \tag{48}$$

Note that for quadratic norm normalized steepest descent becomes normalized gradient descent. In this section we will prove the generalized version of Theorem 5, which applies for normalized steepest descent (instead of just normalized gradient descent):

**Theorem 8. [Generalized Theorem 5]** . *For any separable data set, any initial point* $\mathbf{w}(0)$*, consider the normalized steepest descent updates above with a variable step size* $\eta_t = \frac{1}{\sqrt{t+1}}$ *for linear classification with the exponential loss* $\ell(u) = \exp(-u)$.

*The margin of the iterates* $\mathbf{w}(t)$ *converge to max margin* $\gamma$ *at rate* $t^{-1/2}\log t$*:*

$$\frac{\mathbf{w}(t+1)^\top \mathbf{x}_n}{\|\mathbf{w}(t+1)\|} \geq \gamma - \frac{1}{2}\frac{1+\log(t+1)}{\gamma(2\sqrt{t+2}-2)} - \frac{\log \mathcal{L}(\mathbf{w}(0))}{\gamma(2\sqrt{t+2}-2)}. \tag{49}$$

*Proof.* Since $\eta_t / L(w_t) < 1/L(w_t)$, this stepsize choice satisfies the conditions of Lemma 12 in (Gunasekar et al., 2018b), and so the objective function is decreasing.

The progress in one step is

$$\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(t)) \exp\big(-\eta_t \gamma_t^2 + \frac{\eta_t^2}{2}\gamma_t^2\big) \tag{50}$$

$$\leq \mathcal{L}(w_0) \exp\big(-\sum_{i=0}^t \eta_i \gamma_i^2 + \sum_{i=0}^t \frac{\eta_i^2}{2}\gamma_i^2\big) \tag{51}$$

The margin bound is

$$\max_{n\in[N]} \exp(-\mathbf{w}(t+1)^\top \mathbf{x}_n) \leq \mathcal{L}(\mathbf{w}(t+1)) \leq \exp\big(-\sum_{i=0}^t \eta_i \gamma_i^2 + \sum_{i=0}^t \frac{\eta_i^2}{2}\gamma_i^2\big). \tag{52}$$

By applying $-\log$,

$$\min_{n\in[N]} \mathbf{w}(t+1)^\top \mathbf{x}_n \geq \sum_{i=0}^{t} \eta_i \gamma_i^2 - \sum_{i=0}^{t} \frac{\eta_i^2}{2}\gamma_i^2 - \log \mathcal{L}(w_0). \tag{53}$$

The norm growth is

$$\|\mathbf{w}(t+1)\| = \left\| \sum_{i=0}^{t} \eta_i \gamma_i p_i \right\| \leq \sum_{i=0}^{t} \eta_i \gamma_i. \tag{54}$$

Thus the margin of every point $j$ satisfies

$$\frac{\mathbf{w}(t+1)^\top \mathbf{x}_n}{\|\mathbf{w}(t+1)\|} \geq \frac{\sum_{i=0}^t \eta_i \gamma_i^2}{\sum_{i=0}^t \eta_i \gamma_i} - \frac{\sum_{i=0}^t \frac{\eta_i^2}{2}\gamma_i^2}{\sum_{i=0}^t \eta_i \gamma_i} - \frac{\log \mathcal{L}(w_0)}{\sum_{i=0}^t \eta_i \gamma_i}. \tag{55}$$

Choose $\eta_i = \frac{1}{\sqrt{i+1}}$ so that $\sum_{i=0}^t \eta_i \geq 2\sqrt{t+2} - 2$. Since $\gamma_i \geq \gamma$, then

$$\frac{\mathbf{w}(t+1)^\top \mathbf{x}_n}{\|\mathbf{w}(t+1)\|} \geq \gamma \frac{\sum_{i=0}^t \eta_i \gamma_i}{\sum_{i=0}^t \eta_i \gamma_i^2} - \frac{\sum_{i=0}^t \frac{\eta_i^2}{2}\gamma_i^2}{\sum_{i=0}^t \eta_i \gamma_i} - \frac{\log \mathcal{L}(w_0)}{\gamma(2\sqrt{t+2}-2)} \tag{56}$$

$$= \gamma - \frac{\sum_{i=0}^t \frac{\eta_i^2}{2}\gamma_i^2}{\sum_{i=0}^t \eta_i \gamma_i} - \frac{\log \mathcal{L}(w_0)}{\gamma(2\sqrt{t+2}-2)} \tag{57}$$

Assume that $\|\mathbf{x}_n\|_* \leq 1$, so that $\|\nabla \mathcal{L}(\mathbf{w}(t))\|_* \leq \max_{j\in[n]} \|\mathbf{x}_n\|_* \mathcal{L}(\mathbf{w}(t))$. Thus $\gamma_i \leq 1$. Using this

$$\frac{\mathbf{w}(t+1)^\top \mathbf{x}_n}{\|\mathbf{w}(t+1)\|} \geq \gamma - \frac{1}{2}\frac{\sum_{i=0}^t \eta_i^2}{\gamma \sum_{i=0}^t \eta_i} - \frac{\log \mathcal{L}(w_0)}{\gamma(2\sqrt{t+2}-2)} \tag{58}$$

$$\geq \gamma - \frac{1}{2}\frac{1+\log(t+1)}{\gamma(2\sqrt{t+2}-2)} - \frac{\log \mathcal{L}(w_0)}{\gamma(2\sqrt{t+2}-2)} \tag{59}$$

□

# C Tail Analysis – Proof sketch

In this section we describe non-rigorously the main ideas for our proofs on the results on the affect of loss tail on the convergence rate. In later appendix sections we give the complete proofs. Recall we consider strictly monotone losses (Definition 1) with a general tail, given by $-\ell'(u) = \exp(-f(u))$, such that $f(u)$ is a strictly increasing function of $u$.

## C.1 Convergence to the max-margin separator

From Lemma 1 in Soudry et al. (2018a) we know that for linearly separable datasets, and smooth strictly monotonic loss functions, the iterates of GD entail that $\|\mathbf{w}(t)\| \to \infty$ and $\mathcal{L}(\mathbf{w}(t)) \to 0$ as $t \to \infty$, if the learning rate is sufficiently small. Now, if $\lim_{t\to\infty} \mathbf{w}(t)/\|\mathbf{w}(t)\|$ exists, then we can write $\mathbf{w}(t) = \mathbf{w}_\infty g(t) + \boldsymbol{\rho}(t)$ where $\lim_{t\to\infty} g(t) = \infty$, $\forall n: \mathbf{x}_n^\top \mathbf{w}_\infty > 0$ and $\lim_{t\to\infty} \|\boldsymbol{\rho}(t)\|/g(t) = 0$. Using this result, the gradients can be written as:

$$-\nabla \mathcal{L}(\mathbf{w}(t)) = \sum_{n=1}^{N} \exp\left(-f\left(\mathbf{w}(t)^\top \mathbf{x}_n\right)\right)\mathbf{x}_n = \sum_{n=1}^{N} \exp\left(-f\left(g(t)\mathbf{w}_\infty^\top \mathbf{x}_n + \boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)\right)\mathbf{x}_n \tag{60}$$

As $g(t) \to \infty$ the exponents become more negative, since $f(t)$ is an increasing function, $\forall n: \mathbf{x}_n^\top \mathbf{w}_\infty > 0$ and $\|\boldsymbol{\rho}(t)\| = o(g(t))$. Therefore, if $f$ is increasing sufficiently fast, only samples with minimal margin $\mathbf{w}_\infty^\top \mathbf{x}_n$ contribute to the sum. Examining the gradient descent dynamics, this implies that $\mathbf{w}(t)$ and also its scaling $\hat{\mathbf{w}} = \frac{\mathbf{w}(t)}{\min_n \mathbf{w}_\infty^\top \mathbf{x}_n}$ are a linear non negative combination of support vectors:

$$\hat{\mathbf{w}} = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n \;\; \forall n: \left(\alpha_n \geq 0 \text{ and } \hat{\mathbf{w}}^\top \mathbf{x}_n = 1\right) \text{ or } \left(\alpha_n = 0 \text{ and } \hat{\mathbf{w}}^\top \mathbf{x}_n > 1\right) \tag{61}$$

these are exactly the KKT conditions for the SVM problem and we can conclude that $\mathbf{w}_\infty$ is proportional to $\hat{\mathbf{w}}$.

### C.2 Calculation of rates and validity conditions

Next, we aim to find $g(t)$ and $||\boldsymbol{\rho}(t)||$ so we can calculate the convergence rates. Also, we aim to find what are the conditions on $f$ so this calculation would break. To simplify our analysis we examine the continuous time version of GD, in which we take the limit $\eta \to 0$. In this limit

$$\dot{\mathbf{w}}(t) = -\nabla\mathcal{L}(\mathbf{w}(t)) = \sum_{n=1}^{N} \exp\left(-f\left(\mathbf{x}_n^\top \mathbf{w}(t)\right)\right)\mathbf{x}_n, \tag{62}$$

We define $\mathcal{S} = \operatorname{argmin}_n \hat{\mathbf{w}}^\top \mathbf{x}_n$, i.e., the set of indices of support vectors, so $\forall n \in \mathcal{S}$ we have $\hat{\mathbf{w}}^\top \mathbf{x}_n = 1$. From our reasoning above, if $f$ increases fast enough, then we expect that the contribution of the non-support vectors to the gradient would be negligible, and therefore

$$\dot{\mathbf{w}}(t) \approx \sum_{n\in\mathcal{S}} \exp\left(-f\left(\mathbf{x}_n^\top \mathbf{w}(t)\right)\right)\mathbf{x}_n, \tag{63}$$

Additionally, if we assume that $\boldsymbol{\rho}(t)$ converges to some direction $\boldsymbol{a}$, and $\mathbf{b}$ is some vector orthogonal to the support vectors (if such direction exists), then we expect that asymptotic solution to be of the form

$$\mathbf{w}(t) = g(t)\hat{\mathbf{w}} + h(t)\boldsymbol{a} + \mathbf{b},\ \text{s.t. } h(t) = o(g(t)).$$

In order for this to be a valid solution, it must satisfy eq. (63). We verify this by substitution and examining the leading orders

$$\begin{aligned}
\dot{g}(t)\hat{\mathbf{w}} &\approx \sum_{n\in\mathcal{S}} \exp\left(-f\left(g(t)\mathbf{x}_n^\top\hat{\mathbf{w}} + h(t)\mathbf{x}_n^\top\boldsymbol{a}\right)\right)\mathbf{x}_n \\
&\overset{(1)}{\approx} \sum_{n\in\mathcal{S}} \exp\left(-\left(f\left(g(t)\mathbf{x}_n^\top\hat{\mathbf{w}}\right) + f'\left(g(t)\mathbf{x}_n^\top\hat{\mathbf{w}}\right)h(t)\mathbf{x}_n^\top\boldsymbol{a}\right)\right)\mathbf{x}_n \\
&\overset{(2)}{\approx} \exp(-f(g(t)))\sum_{n\in\mathcal{S}} \exp(-f'\left(g(t)\right)h(t)\mathbf{x}_n^\top\boldsymbol{a})\mathbf{x}_n,
\end{aligned}$$

where in (1) we used a Taylor approximation and in (2) we used that $\hat{\mathbf{w}}^\top\mathbf{x}_n = 1$, $\forall n \in \mathcal{S}$. For the last equation to to hold, we require

$$\dot{g}(t) = \exp\left(-f\left(g(t)\right)\right),\ h(t) = \frac{1}{f'\left(g(t)\right)} \tag{64}$$

and $\boldsymbol{a}$ satisfies the equations:

$$\forall n \in \mathcal{S}\ :\ \exp(-\mathbf{x}_n^\top\boldsymbol{a}) = \alpha_n,\ \bar{\mathbf{Q}}\boldsymbol{a} = 0, \tag{65}$$

where we define $\mathbf{Q} \in \mathbb{R}^d$ as the orthogonal projection matrix to the subspace spanned by the support vectors, and $\bar{\mathbf{Q}} = I - \mathbf{Q}$ as the complementary projection matrix. Equation 65 has a unique solution for almost every dataset from Lemma 8 in Soudry et al. (2018a). Specifically, this equation does not have a solution when one of the $\alpha_n$ must be equal to zero (i.e., some support vectors exert "zero force" on the the margin — and this happens only in measure zero cases).

Since we assume that $h(t) = o(g(t))$ we must have $\lim_{t\to\infty} g(t)f'(g(t)) = \lim_{u\to\infty} uf'(u) = \infty$ meaning $f'(t) = \omega(t^{-1})$ which implies $f(t) = \omega(\log(t))$. This condition must hold for this analysis to make sense. Moreover, the differential equation that defines $g(t)$ (eq. (64)) is generally intractable. However, if the condition $\log\left(f'\left(t\right)\right) = o\left(f\left(t\right)\right)$ holds (which is true for many functions), then we can approximate

$$\dot{g}\left(t\right) \approx \exp\left(-f\left(g\left(t\right)\right) - \log\left(f'\left(g\left(t\right)\right)\right)\right)$$

which has a closed form solution

$$g\left(t\right) = f^{-1}\left(\log\left(t + C\right)\right).$$

# D Proof of Theorem 2

## D.1 Auxillary lemma

In order to prove Theorem 2, we need to expand Lemma 8 in Gunasekar et al. (2018b) to general loss functions. To this end, we first prove the following lemma:

**Lemma 3.** *Let $f$ be a differentiable function so that $f'(u) = \omega\left(u^{-1}\right)$ and let $g$ be a function satisfying $\lim_{t\to\infty} g(t) = \infty$. Then, for any $a_1 > a_2$:*

$$\lim_{t\to\infty}\left(f\left(a_1 \cdot g\left(t\right)\right) - f\left(a_2 \cdot g\left(t\right)\right)\right) = \infty$$

*Proof.* We denote $\tilde{f}(t) \triangleq f\left(a_1 \cdot g\left(t\right)\right) - f\left(a_2 \cdot g\left(t\right)\right)$. Let $M > 0$ be any number. We need to prove that $\exists \tilde{t} > 0$ s.t. $\forall t > \tilde{t}$: $\tilde{f}(t) \geq M$. We show this in the following steps:

- First, $f'(u) = \omega\left(u^{-1}\right)$ implies that $\forall M', \exists u_1$ so that $\forall u > u_1 : f'(u) \geq M' u^{-1}$. We choose $M' = \dfrac{M}{\log(a_1/a_2)} > 0$ (since $a_1 > a_2$).
- Secondly, $\lim_{t\to\infty} g(t) = \infty$ implies that $\exists t_2 > 0$ so that $\forall t > t_2 :\ a_2 \cdot g\left(t\right) > u_1$ (the corresponding $u_1$ for the $M'$ we chose).

We define $\tilde{t} = t_2$, so that $\forall t > \tilde{t}$:

$$f\left(a_1 \cdot g\left(t\right)\right) - f\left(a_2 \cdot g\left(t\right)\right) = \int_{a_2 \cdot g(t)}^{a_1 \cdot g(t)} f'(\tau) d\tau \geq \int_{a_2 \cdot g(t)}^{a_1 \cdot g(t)} M' \frac{1}{\tau} d\tau = M' \log\left(\frac{a_1 \cdot g\left(t\right)}{a_2 \cdot g\left(t\right)}\right) = M.$$

□

Using this lemma, we can prove an extended version of Lemma 8 in Gunasekar et al. (2018b), which only applied to exponential tail.

**Lemma 4.** *For almost all datasets $\left\{\mathbf{x}_n, y_n\right\}_{n=1}^N$ which are linearly separable, consider any sequence $\mathbf{w}(t)$ that minimizes the empirical loss in eq. (1), i.e., $\mathcal{L}\left(\mathbf{w}(t)\right) \to 0$ with a strictly monotone loss function $\ell$ satisfying Assumption 1, i.e., $\ell'(u) = -\exp(-f(u))$, with $f'(u) > 0$ and $f'(u) = \omega\left(u^{-1}\right)$.*

*If (a) $\bar{\mathbf{w}}_\infty \triangleq \lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$ exists and has a positive margin, and (b) $\mathbf{z}_\infty \triangleq \lim_{t\to\infty} \frac{-\nabla_\mathbf{w}\mathcal{L}(\mathbf{w}(t))}{\|\nabla_\mathbf{w}\mathcal{L}(\mathbf{w}(t))\|}$ exists , then $\exists \left\{\alpha_n \geq 0\right\}_{n=1}^N$ s.t.*

$$\mathbf{z}_\infty = \lim_{t\to\infty} \frac{-\nabla_\mathbf{w}\mathcal{L}\left(\mathbf{w}(t)\right)}{\left\|\nabla_\mathbf{w}\mathcal{L}\left(\mathbf{w}(t)\right)\right\|} = \lim_{t\to\infty} \frac{-\sum_{n\in\mathcal{S}'} \ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n}{\left\|\sum_{n\in\mathcal{S}'} \ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n\right\|} = \sum_{n\in\mathcal{S}'} \alpha_n y_n \mathbf{x}_n$$

*where $S'$ denotes the indices of support vectors (data points with the smallest margin) of the limit direction $\bar{\mathbf{w}}_\infty$ given by $\mathcal{S}' = \left\{n : y_n \bar{\mathbf{w}}_\infty^\top \mathbf{x}_n = \min_{\bar{n}} y_{\bar{n}} \bar{\mathbf{w}}_\infty^\top \mathbf{x}_{\bar{n}}\right\}$.*

*Proof.* $\left\{\mathbf{x}_n, y_n\right\}_{n=1}^N$ is a linearly separable dataset. We assume that $\forall n :\ y_n = 1$ without loss of generality, since we can always re-define $y_n \mathbf{x}_n$ as $\mathbf{x}_n$.

From the assumption in the lemma, we have that $\mathcal{L}(\mathbf{w}(t)) \to 0$ as $t \to \infty$ and that the loss is a strictly monotone loss function (Definition 1). This implies that $\|\mathbf{w}(t)\| \to \infty$ as $t \to \infty$. Also, since $\dfrac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$ converges in direction to $\bar{\mathbf{w}}_\infty$ we can write $\mathbf{w}(t) = \bar{\mathbf{w}}_\infty g(t) + \boldsymbol{\rho}(t)$ where $\lim_{t\to\infty} g(t) = \infty$ and $\lim_{t\to\infty} \dfrac{\|\boldsymbol{\rho}(t)\|}{g(t)} = 0$. From these conditions we also have that $\bar{\mathbf{w}}_\infty^\top \mathbf{X} > 0$, where recall that $\mathbf{X}$ denotes the data matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$.

We define the following additional notations:

- $\gamma = \min_n \bar{\mathbf{w}}_\infty^\top \mathbf{x}_n > 0$. This is the minimal margin of $\bar{\mathbf{w}}_\infty$.
- $\mathcal{S}' = \left\{n : \bar{\mathbf{w}}_\infty^\top \mathbf{x}_n = \gamma\right\}$.

- $\bar{\gamma} = \min_{n \notin \mathcal{S}} \bar{\mathbf{w}}_{\infty}^{\top} \mathbf{x}_n$. This is the second smallest margin of $\bar{\mathbf{w}}_{\infty}$.
- $\bar{\gamma}_n = \bar{\mathbf{w}}_{\infty}^{\top} \mathbf{x}_n$. This is the margin of the datapoint $\mathbf{x}_n$
- $B = \max_n \|\mathbf{x}_n\|$.

Since $\lim_{t\to\infty} \dfrac{\|\boldsymbol{\rho}(t)\|}{g(t)} = 0$ we have that $\forall \epsilon_1, \epsilon_2 > 0, \exists t_\epsilon > 0$ such that $\forall t > t_\epsilon$, the following holds

$$\max_{n\in\mathcal{S}} \boldsymbol{\rho}(t)^{\top} \mathbf{x}_n \leq \|\boldsymbol{\rho}(t)\| B \leq \epsilon_1 \gamma g(t), \tag{66}$$

$$\min_{n\notin\mathcal{S}} \boldsymbol{\rho}(t)^{\top} \mathbf{x}_n \geq -\|\boldsymbol{\rho}(t)\| B \geq -\epsilon_2 \bar{\gamma} g(t). \tag{67}$$

For the general loss we defined in the lemma the gradients are given by

$$\begin{aligned} -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}(t)) &= \sum_{n=1}^{N} \exp\left(-f\left(g(t)\mathbf{x}_n^{\top}\bar{\mathbf{w}}_{\infty} + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n \\ &= \sum_{n\in\mathcal{S}'} \exp\left(-f\left(g(t)\gamma + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n + \sum_{n\notin\mathcal{S}'} \exp\left(-f\left(g(t)\bar{\gamma}_n + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n \\ &= I(t) + II(t), \end{aligned}$$

where $I(t) = \sum_{n\in\mathcal{S}'} \exp\left(-f\left(g(t)\gamma + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n$ and $II(t) = \sum_{n\notin\mathcal{S}'} \exp\left(-f\left(g(t)\bar{\gamma}_n + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n$.

**Step 1:** Show that $\lim_{t\to\infty} \dfrac{\|II(t)\|}{\|I(t)\|} = 0$.

We define $\mathbf{v}(t) \in \mathbb{R}^{|\mathcal{S}'|}$ as $\forall k = 1, ..., |\mathcal{S}'|$: $\mathbf{v}_k(t) = \exp\left(-f\left(g(t)\gamma + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_{i_k}\right)\right)$ where $i_1, ..., i_{|\mathcal{S}'|} \in \mathcal{S}'$. In addition, with these indices, we define $X_{\mathcal{S}'} = \begin{pmatrix} | & & | \\ \mathbf{x}_{i_1} & \dots & \mathbf{x}_{i_{|\mathcal{S}'|}} \\ | & & | \end{pmatrix}$. We denote $\sigma_{\min}(\mathbf{X}_{\mathcal{S}'})$ as the minimal singular value of $\mathbf{X}_{\mathcal{S}'}$.

For almost all datasets, $\sigma_{\min}(\mathbf{X}_{\mathcal{S}'}) > 0$ (from Claim 1 in Gunasekar et al. (2018b)).

Using these notations, we can lower bound $\|I(t)\|$. $\forall t > t_\epsilon$:

$$\begin{aligned} \|I(t)\| &= \left\| \sum_{n\in\mathcal{S}'} \exp\left(-f\left(g(t)\gamma + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n \right\| \overset{(1)}{\geq} \sigma_{\min}(X'_{\mathcal{S}}) \min_{n\in\mathcal{S}'} \exp\left(-f\left(g(t)\gamma + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right) \\ &\overset{(2)}{=} \sigma_{\min}(X'_{\mathcal{S}}) \exp\left(-f\left(g(t)\gamma + \max_{n\in\mathcal{S}'} \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right) \overset{(3)}{\geq} \sigma_{\min}(X'_{\mathcal{S}}) \exp\left(-f\left(g(t)\gamma(1+\epsilon_1)\right)\right), \end{aligned} \tag{68}$$

where in (1) we used the fact that

$$\left\| \sum_{n\in\mathcal{S}'} \exp\left(-f\left(g(t)\gamma + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n \right\| = \|X'_{\mathcal{S}}\mathbf{v}\| \geq \sigma_{\min}(X'_{\mathcal{S}}) \|\mathbf{v}\| \geq \sigma_{\min}(X'_{\mathcal{S}}) \min_{n\in\mathcal{S}'} \mathbf{v}_n,$$

in (2) we used the fact that $f$ is strictly increasing (or $f'(u) > 0$) and in (3) we used eq. (66).

Next, we upper bound $\|II(t)\|$. $\forall t > t_\epsilon$:

$$\begin{aligned} \|II(t)\| &= \left\| \sum_{n\notin\mathcal{S}'} \exp\left(-f\left(g(t)\bar{\gamma}_n + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right)\mathbf{x}_n \right\| \overset{(1)}{\leq} NB \max_{n\notin\mathcal{S}'} \exp\left(-f\left(g(t)\bar{\gamma}_n + \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right) \\ &\overset{(2)}{=} NB \exp\left(-f\left(\min_{n\notin\mathcal{S}'} g(t)\bar{\gamma}_n + \min_{n\notin\mathcal{S}'} \boldsymbol{\rho}(t)^{\top}\mathbf{x}_n\right)\right) \overset{(3)}{\leq} NB \exp\left(-f\left(g(t)\bar{\gamma}(1-\epsilon_2)\right)\right), \end{aligned} \tag{69}$$

where in (1) we used the triangle inequality along with $\|\mathbf{x}_n\| \leq B$, in (2) we used the fact that $f$ is strictly increasing and in (3) we used eq. (67) and $\forall n \notin \mathcal{S}' : \bar{\gamma}_n \geq \bar{\gamma}$.

Combining equations 68, 69 we have that $\forall\epsilon_1, \epsilon_2 > 0, \exists t_\epsilon > 0$ so that $\forall t > t_\epsilon$:

$$\frac{\|II(t)\|}{\|I(t)\|} \leq \frac{NB}{\sigma_{\min}\left(X'_{\mathcal{S}}\right)} \exp\left(-\left(f\left(g(t)\bar{\gamma}\left(1-\epsilon_2\right)\right)\right) - f\left(g(t)\gamma\left(1+\epsilon_1\right)\right)\right).$$

For the choice $\epsilon_1 = \dfrac{\bar{\gamma}-\gamma}{4\gamma}$, $\epsilon_2 = \dfrac{\bar{\gamma}-\gamma}{4\bar{\gamma}}$ we have that $\bar{\gamma}\left(1-\epsilon_2\right) > \gamma\left(1+\epsilon_1\right)$ and thus, using Lemma 3 and Squeeze theorem we have that $\lim\limits_{t\to\infty}\dfrac{\|II(t)\|}{\|I(t)\|} = 0$.

**Step 2:** Using $\lim\limits_{t\to\infty}\dfrac{\|II(t)\|}{\|I(t)\|} = 0$ show that, $\lim\limits_{t\to\infty}\dfrac{-\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)}{\|\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)\|} = \sum\limits_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n$ for some $\alpha_n \geq 0$ (if the limit exists).

Since $\lim\limits_{t\to\infty}\dfrac{\|II(t)\|}{\|I(t)\|} = 0$, we have $\frac{\|I(t)+II(t)\|}{\|I(t)\|}$ satisfying $1-\frac{\|II(t)\|}{\|I(t)\|} \leq \frac{\|I(t)+II(t)\|}{\|I(t)\|} \leq 1+\frac{\|II(t)\|}{\|I(t)\|}$. Using squeeze theorem, we get $\frac{\|I(t)+II(t)\|}{\|I(t)\|} \to 1$.

Now consdier the limit direction of gradients,

$$\frac{-\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)}{\|\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)\|} = \frac{I(t)}{\|I(t)+II(t)\|} + \frac{II(t)}{\|I(t)+II(t)\|}.$$

- $\left\|\dfrac{II(t)}{\|I(t)+II(t)\|}\right\| = \dfrac{\|II(t)\|}{\|I(t)\|}\dfrac{\|I(t)\|}{I(t)+II(t)} \xrightarrow{t\to\infty} 0.$
- Similarly, $\lim_{t\to\infty}\frac{\|I(t)\|}{\|I(t)+II(t)\|} = \lim_{t\to\infty}\frac{I(t)}{\|I(t)\|}\frac{\|I(t)\|}{I(t)+II(t)}\lim_{t\to\infty}\frac{I(t)}{\|I(t)\|}$.
- Finally, since $I(t) \propto \sum\limits_{n\in\mathcal{S}} v_n(t)\mathbf{x}_n$ for $v_k(t) > 0$, then every limit point of $\dfrac{-\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)}{\|\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)\|}$ converges to $\sum\limits_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n$ for some $\alpha_n \geq 0$.

Summarizing, if $\lim\limits_{t\to\infty}\dfrac{-\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)}{\|\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)\|}$ exists, then

$$\frac{-\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)}{\|\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)\|} = \lim_{t\to\infty}\frac{-\sum\limits_{n\in\mathcal{S}'}\ell'\left(\mathbf{w}(t)^\top\mathbf{x}_n\right)\mathbf{x}_n}{\left\|\sum\limits_{n\in\mathcal{S}'}\ell'\left(\mathbf{w}(t)^\top\mathbf{x}_n\right)\mathbf{x}_n\right\|} = \sum_{n\in\mathcal{S}'}\alpha_n\mathbf{x}_n.$$

This completes the proof for general tail. □

### D.2 Theorem 2 Proof

**Theorem 2.** *For any depth $L$, almost all linearly separable datasets, almost all initialization and any bounded sequence of step sizes $\{\eta_t\}$, consider the sequence $\mathcal{W}(t) = \{\mathbf{W}_l(t)\}_{l=1}^L$ of gradient descent updates in eq. (6) for minimizing the empirical loss $\mathcal{L}_{\mathcal{P}}(\mathcal{W})$ (eq. (5)) with a strictly monotone loss function $\ell$ satisfying Assumption 1, i.e.: $\ell'(u) = -\exp(-f(u)) < 0$, where asymptotically $f'(u) > 0$ and $f'(u) = \omega\left(u^{-1}\right)$.*

*If (a) $\mathcal{W}(t)$ minimizes the empirical loss, i.e. $\mathcal{L}_{\mathcal{P}}\left(\mathcal{W}(t)\right) \to 0$, (b) $\mathcal{W}(t)$, and consequently $\mathbf{w}(t) = \mathcal{P}\left(\mathbf{w}(t)\right)$, converge in direction to yield a separator with positive margin, and (c) the gradients with respect to the linear predictors $\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)$ converge in direction, then the limit direction is given by,*

$$\bar{\mathbf{w}}_\infty = \lim_{t\to\infty}\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|},$$

*where*

$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}\|\mathbf{w}\|^2 \text{ s.t. } \mathbf{w}^\top\mathbf{x}_n \geq 1. \tag{8}$$

The proof for this theorem is very similar to Theorem 1 proof in Gunasekar et al. (2018a). The main exception is that instead of using Lemma 8 in Gunasekar et al. (2018a), which only applies to exponential loss, we use the extended lemma which was proved in the previous section (Lemma 4).

*Proof.* Let $\mathcal{W}(t) = [\mathbf{W}_l(t) \in \mathbb{R}^{D_{l-1}\times D_l}]_{l=1}^L$ denote the iterates of individual matrices $\mathbf{W}_l(t)$ along the gradient descent path, and $\mathbf{w}(t) = \mathbf{W}_1(t)...\mathbf{W}_l(t)$ denote the corresponding sequence of linear predictors.

We first introduce the following notation.

1. Let $\overline{\mathbf{W}}_\infty = \lim\limits_{t\to\infty} \frac{\mathcal{W}(t)}{\|\mathcal{W}(t)\|}$ denote the limit direction of the parameters, with component matrices in each layer denoted as $\overline{\mathbf{W}}_\infty = [\overline{\mathbf{W}}_l^\infty]$. We have that for some $\boldsymbol{\delta}_{\mathbf{W}_l}(t) \to 0$ the following representation of $\mathbf{W}_l(t)$ holds.
$$\mathbf{W}_l(t) = \overline{\mathbf{W}}_l^\infty g(t) + \boldsymbol{\delta}_{\mathbf{W}_l}(t)\, g(t), \tag{70}$$
where $g(t) = \|\mathcal{W}(t)\|$ and $\boldsymbol{\delta}_{\mathbf{W}_l}(t) \to 0$.
2. For $0 < l_1 < l_2 \leq L$, denote $\mathbf{W}_{l_1:l_2}(t) = \mathbf{W}_{l_1}(t)\mathbf{W}_{l_1+1}(t)\dots\mathbf{W}_{l_2}(t)$ and $\overline{\mathbf{W}}_{l_1:l_2}^\infty = \overline{\mathbf{W}}_{l_1}^\infty\overline{\mathbf{W}}_{l_1+1}^\infty\dots\overline{\mathbf{W}}_{l_2}^\infty$. Using eq. (77), we can check by induction on $l_2 - l_1$ that $\exists\, \boldsymbol{\delta}_{\mathbf{W}_{l_1:l_2}}(t) \to 0$ such that the following holds,
$$\mathbf{W}_{l_1:l_2}(t) = \overline{\mathbf{W}}_{l_1:l_2}^\infty\, g(t)^{l_2-l_1+1} + \boldsymbol{\delta}_{\mathbf{W}_{l_1:l_2}}(t)\, g(t)^{l_2-l_1+1}. \tag{71}$$
3. Denote the gradients with respect to linear predictors as $\mathbf{z}(t) = -\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}(t))$. Since we assume that $\mathbf{z}(t)$ converges in direction, let $\bar{\mathbf{z}}^\infty = \lim\limits_{t\to\infty}\frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|}$. Denoting $p(t) = \|\mathbf{z}(t)\|$, for some $\boldsymbol{\delta}_{\mathbf{z}}(t) \to 0$, we can write $\mathbf{z}(t)$ as,
$$\mathbf{z}(t) = \bar{\mathbf{z}}^\infty p(t) + \boldsymbol{\delta}_{\mathbf{z}}(t)\, p(t). \tag{72}$$
4. From Lemma 4, we have that $\exists\{\alpha_n\}_{n\in S_\infty}$ such that $\bar{\mathbf{z}}^\infty = \sum\limits_{n\in S_\infty}\alpha_n\, y_n\mathbf{x}_n$, where $S_\infty$ are support vectors of $\bar{\mathbf{w}}^\infty = \lim\limits_{t\to\infty}\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$.

The proof of Theorem 2 is fairly straight forward from using Lemma 4. In the following arguments we show that a positive scaling $\tilde{\mathbf{w}}_\infty = \gamma\lim\limits_{t\to\infty}\frac{\mathcal{P}(\mathcal{W}(t))}{\|\mathcal{P}(\mathcal{W}(t))\|}$ satisfies the following KKT conditions for the optimality of explicitly regularized convex problem in eq. (8):

$$\begin{aligned}\exists\{\alpha_n\}_{n=1}^N \quad \text{s.t.} \quad &\forall n,\ \langle\mathbf{x}_n,\mathbf{w}\rangle \geq 1, \mathbf{w} = \sum_n \alpha_n\,\mathbf{x}_n,\\ &\forall n, \alpha_n \geq 0 \text{ and } \alpha_n = 0, \forall i \notin S := \{i\in[N] : \langle\mathbf{x}_n,\mathbf{w}\rangle = 1\}.\end{aligned} \tag{73}$$

Since $\bar{\mathbf{w}}_\infty \triangleq \overline{\mathbf{W}}_{1:L}^\infty$ has strictly positive margin, we can scale $\overline{\mathbf{W}}_{1:L}^\infty$ to get $\tilde{\mathbf{w}}^\infty = \gamma\bar{\mathbf{w}}_\infty$ with unit margin, i.e., $\forall n,\ \langle\mathbf{x}_n,\tilde{\mathbf{w}}^\infty\rangle \geq 1$. For the dual variables, we again use a positive scaling of $\alpha_n$ from Lemma 4, such that $\bar{\mathbf{z}}^\infty = \sum_{n\in S_\infty}\alpha_n\,\mathbf{x}_n$. In order to prove the theorem, we need to show that $\tilde{\mathbf{w}}^\infty \propto \bar{\mathbf{z}}^\infty$ or equivalently $\overline{\mathbf{W}}_{1:L}^\infty \propto \bar{\mathbf{z}}^\infty$.

Computing the gradients descent updates for $\mathbf{W}_1(t)$, we have

$$\mathbf{W}_1(t+1) - \mathbf{W}_1(t) = -\eta_t\nabla_{\mathbf{W}_1}\mathcal{L}_{\mathcal{P}}(\mathcal{W}(t)) = -\eta_t\nabla_{\mathbf{W}_1}\mathcal{L}(\mathbf{W}_{1:L}) = -\eta_t\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}(t))\mathbf{W}_{2:L}(t)^\top = \eta_t\mathbf{z}(t)\mathbf{W}_{2:L}(t)^\top.$$

Using eq. 72 we obtain

$$\begin{aligned}\Delta\mathbf{W}_1 \triangleq \mathbf{W}_1(t+1) - \mathbf{W}_1(t) &= \eta_t\Big(\bar{\mathbf{z}}^\infty p(t) + \boldsymbol{\delta}_{\mathbf{z}}(t)\,p(t)\Big)\Big(\overline{\mathbf{W}}_{2:L}^\infty\, g(t)^{L-1} + \boldsymbol{\delta}_{\mathbf{W}_{2:L}}(t)\, g(t)^{L-1}\Big)^\top\\ &\overset{(1)}{=} \left(\eta_t p(t) g(t)^{L-1}\right)\left(\bar{\mathbf{z}}^\infty(\overline{\mathbf{W}}_{2:L}^\infty)^\top + \boldsymbol{\delta}(t)\right),\end{aligned}$$

where in (1) $\boldsymbol{\delta}(t) = \boldsymbol{\delta}_{\mathbf{z}}(t)\Big(\overline{\mathbf{W}}_{2:L}^\infty + \boldsymbol{\delta}_{\mathbf{W}_{2:L}}(t)\Big)^\top + \boldsymbol{\delta}_{\mathbf{z}}(t)\boldsymbol{\delta}_{\mathbf{W}_{2:L}}(t)^\top \to 0$. This implies that $\mathbf{W}_1(t+1) - \mathbf{W}_1(t)$ converge in direction with positive margin (see Claim 1 in Gunasekar et al. (2018a)).

Summing the last equation over $t$ we obtain

$$\mathbf{W}_1(t) - \mathbf{W}_1(0) = \bar{\mathbf{z}}^\infty(\overline{\mathbf{W}}_{2:L}^\infty)^\top\sum_{u<t}\eta_u p(u) g(u)^{L-1} + \sum_{u<t}\eta_u p(u) g(u)^{L-1}\boldsymbol{\delta}(u)\,. \tag{74}$$

From Claim 1 in Gunasekar et al. (2018a) we have that $\left\|\bar{\mathbf{z}}^\infty(\overline{\mathbf{W}}_{2:L}^\infty)^\top\right\| > 0$ and $\sum_{u<t}\eta_t p(t) g(t)^{L-1} \to \infty$. This implies that the sequence $b_t = \sum_{u<t}\eta_u p(u) g(u)^{L-1}$ is monotonic increasing and diverging. Thus, for $a_t =$

$\sum_{u<t} \boldsymbol{\delta}(u)\eta_u p(u)g(u)^{L-1}$, using Stolz-Cesaro theorem (Theorem 11 in Gunasekar et al. (2018a)), we have

$$\lim_{t\to\infty} \frac{a_t}{b_t} = \lim_{t\to\infty} \frac{\sum_{u<t} \boldsymbol{\delta}(u)\eta_u p(u)g(u)^{L-1}}{\sum_{u<t} \eta_u p(u)g(u)^{L-1}} = \lim_{t\to\infty} \frac{a_{t+1}-a_t}{b_{t+1}-b_t} = \lim_{t\to\infty} \boldsymbol{\delta}(t) = 0.$$

$$\implies \text{for } \boldsymbol{\delta}(t)_2 \to 0, \text{ we have } \sum_{u<t} \boldsymbol{\delta}(u)\eta_u p(u)g(u)^{L-1} = \boldsymbol{\delta}_2(t) \sum_{u<t} \eta_u p(u)g(u)^{L-1}. \tag{75}$$

Substituting eq. (75) in eq. (74), we have

$$\mathbf{W}_1(t) = \left[\bar{\mathbf{z}}^\infty (\overline{\mathbf{W}}_{2:L}^\infty)^\top + \boldsymbol{\delta}_3(t)\right] \sum_{u<t} \eta_u p(u)g(u)^{L-1}\,, \tag{76}$$

where we defined $\boldsymbol{\delta}_3(t) \triangleq \boldsymbol{\delta}_2(t) + \mathbf{W}_1(0)/\sum_{u<t} \eta_u p(u)g(u)^{L-1} \to 0$. Note that the last equation implies that $\mathbf{W}_1(t)$ and $\Delta\mathbf{W}_1$ converge in the same direction. Multiplying the last equation from the right with $\overline{\mathbf{W}}_{2:L}^\infty$ we get that

$$\overline{\mathbf{W}}_{1:L}^\infty \propto \bar{\mathbf{z}}^\infty = \sum_{n\in S_\infty} \alpha_n\, y_n \mathbf{x}_n,$$

where $S_\infty$ are support vectors of $\bar{\mathbf{w}}^\infty = \lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$. This concludes the proof for Theorem 2.

We continue our derivation in order to obtain a useful result which will be used in following proofs. Since $\mathbf{W}_1(t)$ and $\Delta\mathbf{W}_1$ converge in the same direction and

$$\mathbf{W}_l(t) = \overline{\mathbf{W}}_l^\infty g(t) + \delta_{\mathbf{W}_l}(t) g(t) \tag{77}$$

we have that

$$\Delta\mathbf{W}_1(t) = \overline{\mathbf{W}}_1^\infty h(t) + \boldsymbol{\delta}_{\Delta\mathbf{W}_1}(t)\, h(t) \tag{78}$$

where $h(t) = \|\Delta\mathcal{W}(t)\|$, $g(t) = \sum_{u<t} \eta_u h(u) \to \infty$, and $\boldsymbol{\delta}_{\Delta\mathbf{W}_l}(t), \boldsymbol{\delta}_{\mathbf{W}_l}[t] \to 0$. Consider the following arguments on $\Delta\mathbf{W}_1(t)\mathbf{W}_{2:L}(t)$,

$$\begin{aligned}
&\Delta\mathbf{W}_1(t)\mathbf{W}_{2:L}(t) = \mathbf{z}(t)\,\|\mathbf{W}_{2:L}(t)\|^2 \\
\overset{(a)}{\implies}\ & \left(\overline{\mathbf{W}}_1^\infty h(t) + \boldsymbol{\delta}_{\Delta\mathbf{W}_1}(t)\, h(t)\right)\left(\overline{\mathbf{W}}_{2:L}^\infty\, g(t)^{L-1} + \boldsymbol{\delta}_{\mathbf{W}_{2:L}}(t)\, g(t)^{L-1}\right) = \mathbf{z}(t)\,\|\mathbf{W}_{2:L}(t)\|^2 \\
\overset{(b)}{\implies}\ & \frac{\mathbf{z}(t)\,\|\mathbf{W}_{2:L}(t)\|^2}{h(t)\, g(t)^{L-1}} = \overline{\mathbf{W}}_{1:L}^\infty + \boldsymbol{\delta}(t) = \bar{\mathbf{w}}_\infty + \tilde{\boldsymbol{\delta}}(t),
\end{aligned} \tag{79}$$

where in $(a)$, we used eqs. 78-77, and in $(b)$ we have $\tilde{\boldsymbol{\delta}}(t) = \boldsymbol{\delta}_{\Delta\mathbf{W}_1}(t)\boldsymbol{\delta}_{\mathbf{W}_{2:L}}(t) + \boldsymbol{\delta}_{\Delta\mathbf{W}_1}(t)\overline{\mathbf{W}}_{2:L}^\infty + \overline{\mathbf{W}}_1^\infty \boldsymbol{\delta}_{\mathbf{W}_{2:L}}(t) \to 0$. Denote $s(t) := \frac{\|\mathbf{z}(t)\|\|\mathbf{W}_{2:L}(t)\|^2}{h(t)\, g(t)^{L-1}}$ From eq. (79) using triangle inequality we have that

$$\|\bar{\mathbf{w}}_\infty\| - \left\|\tilde{\boldsymbol{\delta}}(t)\right\| \le s(t) \le \|\bar{\mathbf{w}}_\infty\| + \left\|\tilde{\boldsymbol{\delta}}(t)\right\|. \tag{80}$$

Since $\tilde{\boldsymbol{\delta}}(t) \to 0$, by squeeze theorem, we have that $\lim_{t\to\infty} s(t) = \|\bar{\mathbf{w}}_\infty\|$. Using this in eq. (79), we get the following:

$$\frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|} = \frac{\bar{\mathbf{w}}_\infty}{s(t)} + \frac{\boldsymbol{\delta}(t)}{s(t)}. \tag{81}$$

□

# E Proof of Theorems 3 and 4

In Theorem 2, we showed that gradient descent on separable dataset converge in direction to the $L_2$ maximum-margin separator for a large family of super polynomial tailed loss functions specified by Assumption 1. Theorems 3 and 4 show rate of convergence at which gradient descent converges to the maximum-margin separator.

We prove these theorems in the following steps:

1. We first give a general result that specifies the weights model using an ordinary differential equation. This is stated in Theorem 9, the proof of which is provided in appendix section F.
2. With the results from Theorems 9, we explicitly calculate the rates for general tails for deep linear networks. The calculation is in appendix section E.1. This completes the proof of Theorem 4.
3. We give an additional result characterizing $\boldsymbol{\rho}(t)$ component that is not in the support vectors span ($\left\|\bar{\mathbf{Q}}\boldsymbol{\rho}(t)\right\|$), for the special case of $L=1$. This result is stated in Theorem 10 and proved in section G.
4. Finally, we special case Theorem 4 and use Theorem 10 to get simplified results for the case of $L=1$, thus proving Theorem 3.

**Theorem 9.** *Under Assumption 2 and the conditions and notations of Theorem 2, the equivalent linear predictor of a depth $L$ linear network will behave as:*

$$\mathbf{w}(t)=\tilde{g}(t)\hat{\mathbf{w}}+\boldsymbol{\rho}(t) \tag{82}$$

*where $\boldsymbol{\rho}(t)=o\left(\tilde{g}(t)\right)$, $\rho(t)^\top\hat{\mathbf{w}}=0$, and $\hat{\mathbf{w}}$ is the $L_2$ max margin separator*

$$\hat{\mathbf{w}}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\left\|\mathbf{w}\right\|^2 \text{ s.t. } \mathbf{w}^\top\mathbf{x}_n\geq 1.$$

*Further, $\tilde{g}(t)$ and $\boldsymbol{\rho}(t)$ are the asymptotic solution of the following,*

$$\lim_{t\to\infty}\frac{L\eta_t\exp\left(-f\left(\tilde{g}\left(t\right)\right)\right)\left(\tilde{g}(t)\right)^{2\left(1-\frac{1}{L}\right)}\phi_1(t)}{\gamma^{1-2/L}\frac{d}{dt}\tilde{g}(t)}=1, \textit{and}$$

$$\left\|\mathbf{Q}\boldsymbol{\rho}(t)\right\|=\phi_2(t)\left(f'\left(\tilde{g}(t)\right)\right)^{-1}+o\left(\left(f'\left(\tilde{g}(t)\right)\right)^{-1}\right).$$

*where $\phi_1(t)=\Theta(1)$ and $\phi_2(t)=\Theta(1)$ are positive functions that depend only on the data set and not on the loss function $\ell$; $\gamma=\min_n\frac{\hat{\mathbf{w}}^\top\mathbf{x}_n}{\|\hat{\mathbf{w}}\|}$ is the maximum-margin attainable for the dataset with unit $L_2$ norm separator; $\mathbf{Q}\in\mathbb{R}^{d\times d}$ is the orthogonal projection matrix to the subspace spanned by the support vectors.*
*If, in addition, the support vector span the dataset then*

$$\left\|\bar{\mathbf{Q}}\boldsymbol{\rho}(t)\right\|=O(1).$$

For the case $L=1$, if we assume that the loss is $\beta$-smooth, $\eta_t<2\beta^{-1}$ and $f'(t)=\Omega\left(\frac{\log^{1+\epsilon}(t)}{t}\right)$ we can omit the requirement that the loss is minimized (this is guaranteed in this case) and that the support vector span the dataset and prove the following Theorem.

**Theorem 10.** *For $L=1$, $\beta$-smooth loss and $\eta<2\beta^{-1}$, if $f'(t)=\Omega\left(\frac{\log^{1+\epsilon}(t)}{t}\right)$ for some $\epsilon>0$ then*

$$\left\|\bar{\mathbf{Q}}\boldsymbol{\rho}(t)\right\|=O(1),$$

*where $\mathbf{Q}\in\mathbb{R}^{d\times d}$ is the orthogonal projection matrix to the subspace spanned by the support vectors and $\bar{\mathbf{Q}}=I-\mathbf{Q}$ is the complementary projection.*

### E.1 Asymptotic rates for depth $L$ linear networks

From Theorem 9, we can write $\mathbf{w}(t)=\hat{\mathbf{w}}g(t)+\boldsymbol{\rho}(t)$ where $\boldsymbol{\rho}(t)=o(g(t))$ and $\rho(t)^\top\hat{\mathbf{w}}=0$. We can use this to calculate the normalized weight vector:

$$\begin{aligned}\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}&=\frac{g(t)\hat{\mathbf{w}}+\boldsymbol{\rho}(t)}{\sqrt{g(t)^2\hat{\mathbf{w}}^\top\hat{\mathbf{w}}+\boldsymbol{\rho}(t)^\top\boldsymbol{\rho}(t)}}=\frac{\hat{\mathbf{w}}+g^{-1}(t)\boldsymbol{\rho}(t)}{\|\hat{\mathbf{w}}\|\sqrt{1+\frac{\|\boldsymbol{\rho}(t)\|^2}{g^2(t)\|\hat{\mathbf{w}}\|^2}}}\\&\overset{(1)}{=}\frac{\hat{\mathbf{w}}+g^{-1}(t)\boldsymbol{\rho}(t)}{\|\hat{\mathbf{w}}\|}\left[1-O\left(\frac{\|\boldsymbol{\rho}(t)\|^2}{\|\hat{\mathbf{w}}\|^2}\frac{1}{g^2(t)}\right)\right]\\&=\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}+\frac{\boldsymbol{\rho}(t)}{g(t)\|\hat{\mathbf{w}}\|}-O\left(\frac{\|\boldsymbol{\rho}(t)\|^2}{\|\hat{\mathbf{w}}\|^2}\frac{1}{g^2(t)}\right)\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}\end{aligned} \tag{83}$$

where in (1) we used $\frac{1}{\sqrt{1+x}} = 1 - \frac{1}{2}x + \frac{3}{4}x^2 + O(x^3)$.

Calculation of the margin:

$$\begin{aligned}\min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} &\overset{(1)}{=} \min_{n\in\mathcal{S}} \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} \\ &= \min_{n\in\mathcal{S}} \mathbf{x}_n^\top \left[\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|} + \frac{\boldsymbol{\rho}(t)}{g(t)\|\hat{\mathbf{w}}\|} - \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|} O\left(\frac{\|\boldsymbol{\rho}(t)\|^2}{\|\hat{\mathbf{w}}\|^2}\frac{1}{g^2(t)}\right)\right] \\ &= \frac{1}{\|\hat{\mathbf{w}}\|} + \frac{\min_{n\in\mathcal{S}} \mathbf{x}_n^\top \boldsymbol{\rho}(t)}{g(t)\|\hat{\mathbf{w}}\|} + O\left(\frac{\|\boldsymbol{\rho}(t)\|^2}{\|\hat{\mathbf{w}}\|^2}\frac{1}{g^2(t)}\right)\frac{1}{\|\hat{\mathbf{w}}\|},\end{aligned} \tag{84}$$

where in (1) we used the fact that $\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$ converge to the maximum-margin separator and thus the minimal value is obtained on the support vectors.
From Theorem 9, we can also characterize $\boldsymbol{\rho}(t)$:

$$\|\mathbf{Q}\boldsymbol{\rho}(t)\| = \begin{cases}\psi_2(t)\left(f'\left(g(t)\right)\right)^{-1} + o\left(\left(f'\left(g(t)\right)\right)^{-1}\right), & \text{if } \left(f'\left(g(t)\right)\right)^{-1} = \Omega(1) \\ O\left(1\right), & \text{otherwise}\end{cases}, \ \left\|\bar{\mathbf{Q}}\boldsymbol{\rho}(t)\right\| = O(1) \tag{85}$$

where $\psi_2(t) = \Theta(1)$, $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is the orthogonal projection matrix to the subspace spanned by the support vectors and $\bar{\mathbf{Q}} = I - \mathbf{Q}$ is the complementary projection.
Substituting eq. (85) into eqs. 118, 120 we get

$$\left|\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|}\right| = \begin{cases} O\left(\frac{1}{g(t)}\right), & f'(u) = \omega(1) \\ C_1 \frac{1}{g(t)f'(g(t))} + O\left(\frac{1}{g(t)}\right), & \text{otherwise}\end{cases}$$

where $C_1$ is a constant independent of $f(t)$.

### E.2 Asymptotic rates for $L = 1$

In this section we want to show that, in the special case of $L = 1$, the optimal margin convergence rate is obtained for exponential loss.

Using Theorem 9 with general tail and $L = 1$ and without assuming the support vectors span the data we have that $\tilde{g}(t)$ is the asymptotic solution of

$$\exp\left(-f\left(g(t)\right)\right) = \frac{1}{\eta_t \psi_1(t)} \frac{d}{dt} g(t)$$

and

$$\|\mathbf{Q}\boldsymbol{\rho}(t)\| = \begin{cases}\psi_2(t)\left(f'\left(g(t)\right)\right)^{-1} + o\left(\left(f'\left(g(t)\right)\right)^{-1}\right), & \text{if } \left(f'\left(g(t)\right)\right)^{-1} = \Omega(1) \\ O\left(1\right), & \text{otherwise}\end{cases}$$

for positive functions $\psi_1(t) = \Theta(1),\ \psi_2(t) = \Theta(1)$ independent of $f$.

Additionally, From Theorem 10 we have that

$$\left\|\bar{\mathbf{Q}}\boldsymbol{\rho}(t)\right\| = O(1).$$

We note that under Theorem 10 assumptions ($L = 1$, $\beta$-smooth loss and $\eta < 2\beta^{-1}$) we have $\lim_{t\to\infty} \mathcal{L}(\mathbf{w}(t)) = 0$ from Lemma 6, so we can use Theorem 9 without this assumption.

We denote $\tilde{\psi}_1(t) = \eta_t \psi_1 = \Theta(1)$. We define $u(t) = \int_0^t \tilde{\psi}_1(x)dx = H(t) \Rightarrow t = H^{-1}(u)$ (this is well defined since $H(t)$ is monotonic increasing) and $\hat{g}(u) = g\left(H^{-1}(u)\right) = g(t)$. Using these definition we have

$$\frac{d}{du}\hat{g}(u) = \frac{1}{\tilde{\psi}_1(t)} \frac{d}{dt} g(t) = \exp\left(-f\left(\hat{g}(u)\right)\right). \tag{86}$$

Since $\tilde{\psi}_1(t) = \Theta(1)$ we know that exists positive constants $C_L, C_U, t_1$ so that $\forall t > t_1 : C_L \leq \tilde{\psi}_1(t) \leq C_U \Rightarrow C_L t \leq H(t) \leq C_U t$. This implies $\hat{g}(C_L t) \leq g(t) = \hat{g}(H(t)) \leq \hat{g}(C_U t)$ (since $\hat{g}(u)$ is an increasing function). This will enable us to characterize $\tilde{g}(t)$ asymptotic behaviour using $\hat{g}(u)$.

For functions with tight exponential tail ($f(t) = \Theta(t)$) the margin convergence rate is $O(\frac{1}{\log(t)})$ and we know that this bound is tight (this result was proved in Soudry et al. (2018a)).

For $f(t) = \omega(t)$ (the tail goes to zero faster than exponential tail) the margin convergence rates are proportional to $1/g(t)$ (from the calculation in the previous section and Theorems 9 and 10 results). Additionally, For functions with $f(t) = \omega(t)$, the asymptotic solution for eq. (86) is $\hat{g}(u) = f^{-1}(\log(u))$, i.e. $\lim_{u\to\infty} \frac{\hat{g}(u)}{f^{-1}(\log(u))} = 1$ (this result is proved in Wong (2018)). This implies slower convergence rates than the rates obtained with exponential tail ($1/\log(t)$) since in this case $f^{-1}(\log(t)) = o(\log(t))$.

For $f(t) = o(t)$ we first prove the following claim.

**Claim 2.** *For a strictly concave function $f$ that satisfies $f'(t) > 0,\ f'(t) = o(1)$ and $f'(t) = \Omega\left(t^{-1}\log^{1+\epsilon}(t)\right)$, $\exists x'$ so that $\forall x > x'$:*

$$\frac{1}{f^{-1}(x)\, f'\left(f^{-1}(x)\right)} > \frac{1}{x}. \tag{87}$$

*Proof.* We denote $h(x) = f^{-1}(x)$. $h(x)$ is strictly convex since $f$ is strictly increasing and strictly concave. Substituting $h(x)$ and $h'(x) = \frac{1}{f'(f^{-1}(x))}$ into the equation, we need to show that $\exists x_1$ so that $\forall x > x_1$

$$\frac{h'(x)}{h(x)} > \frac{1}{x}. \tag{88}$$

From the gradient inequality, $\forall x > x' > 0$:

$$h'(x)(x - x') > h(x) - h(x')$$
$$h'(x) > \frac{h(x) - h(x')}{x - x'}.$$

Additionally, since $h(t) = \omega(t)$ (from definition and $f(t) = o(t)$) $\exists x''$ so that $\forall x > x''$:

$$h(x) > \frac{h(x')}{x'}x \Leftrightarrow -xh(x') > -x'h(x) \Leftrightarrow \frac{h(x) - h(x')}{x - x'} > \frac{h(x)}{x}.$$

Thus, for $x > \max(x', x'')$

$$h'(x) > \frac{h(x)}{x} \Leftrightarrow \frac{h'(x)}{h(x)} > \frac{1}{x}.$$

□

For $f(t) = o(t)$ we have

$$\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{C_1}{g(t) f'(g(t))} + o\left(\frac{1}{g(t) f'(g(t))}\right),$$

where $C_1$ is a constant independent of $f$. In order to show that the optimal rate is obtained for exponential loss we need to show that, asymptotically,

$$\frac{1}{g_1(t) f'\left(g_1(t)\right)} > \frac{1}{g_2(t)}. \tag{89}$$

where $g_1(t)$ is the solution of the following equation

$$\exp\left(-f\left(g_1(t)\right)\right) = \frac{1}{\eta_t \psi_1(t)} \frac{d}{dt} g_1(t)$$

for $f(t) = o(t)$ and $g_2(t)$ is the solution of this equation for exp tail $f(u) = u$ (asymptotically), i.e.

$$\exp\left(-g_2(t)\right) = \frac{1}{\eta_t \psi_1(t)} \frac{d}{dt} g_2(t).$$

Substituting $t = H^{-1}(u)$ (time rescaling - as explained above) to eq. (89) we obtain the equivalent equation

$$\frac{1}{\hat{g}_1(t) f'(\hat{g}_1(t))} > \frac{1}{\hat{g}_2(t)}, \tag{90}$$

where $\hat{g}_1(u) = g_1(H^{-1}(u))$ and $\hat{g}_2(u) = g_2(H^{-1}(u))$. The obtained ODE (as explained in eq. (86)) for $\hat{g}_1(t)$ and $\hat{g}_2(t)$ are

$$\exp(-f(\hat{g}_1(t))) = \frac{d}{dt}\hat{g}_1(t)\,;\, \exp(-\hat{g}_2(t)) = \frac{d}{dt}\hat{g}_2(t) \Rightarrow \hat{g}_2(t) = \log(t+C).$$

Thus, we need to show that

$$\frac{1}{\hat{g}_1(t) f'(\hat{g}_1(t))} > \frac{1}{\log(t)}$$

(the constant $C$ only contributes an $o(1/log(t))$ term and we are interested in the leading term for characterizing the rates). From claim 3 we have that $\hat{g}_1(t) = f^{-1}(\log(t)) + o\left(f^{-1}(\log(t))\right)$. Thus, since we are only interested in the leading term, we need to show that $\exists t'$ so that $\forall t > t'$:

$$\frac{1}{f^{-1}(\log(t)) f'(f^{-1}(\log(t)))} > \frac{1}{\log(t)}. \tag{91}$$

Using claim 2 with $x = \log(t)$ we get the desired result.

### E.3 $g(t) = f^{-1}(\log(t)) + o\left(f^{-1}(\log(t))\right)$ for $L = 1$

**Claim 3.** *If*

$$\frac{d}{dt}g(t) = \exp(-f(g(t)))\ ,$$

*where $f$ is concave, $f'(t) = o(1)$, $f'(t) = \Omega\left(t^{-1}\log^{1+\epsilon}(t)\right)$ then*

$$g(t) = f^{-1}(\log(t)) + o\left(f^{-1}(\log(t))\right)\ .$$

*Proof.* Note that $g(t)$ is increasing since $\forall t : \dot{g}(t) = \exp(-f(g(t))) > 0$ and unbounded (as we will prove next) and therefore, $\lim_{t\to\infty} g(t) = \infty$. In order to prove that $g(t)$ is unbounded we assume in contradiction that $\exists M, t_0$ such that $\forall t > t_0 : g(t) \leq M$. Thus, $\forall t > t_0 : \dot{g}(t) = \exp(-f(g(t))) \geq \exp(-f(M))$ (since $f$ is increasing) which implies $g(t) \geq \exp(-f(M))\, t + c \to \infty$ in contradiction to our assumption.
We want to show that $g(t) = f^{-1}(\log(t)) + o\left(f^{-1}(\log(t))\right)$.

**First step: proving that** $g(t) \leq f^{-1}(\log(t + C_1))$

$$\dot{g}(t) = \exp(-f(g(t))) \overset{(1)}{\leq} \exp(-f(g(t)))\frac{1}{f'(g(t))}\ ,$$

where in (1) we used $f'(t) = o(1)$ and $\lim_{t\to\infty} g(t) = \infty$. Thus, we have

$$\begin{aligned}\exp(f(g(t))) f'(g(t))\,\dot{g}(t) \leq 1 \Rightarrow \\ \exp(f(g(t))) \leq t + C_1 \Leftrightarrow \\ g(t) \leq f^{-1}(\log(t + C_1))\ ,\end{aligned}$$

where in the last transition we used the fact that $f$ is increasing (and so does $f^{-1}$).

**Second step: proving that** $g(t) \geq f^{-1}(\log(t + C_2)) + o\left(f^{-1}(\log(t + C_2))\right)$

$$\dot{g}(t) = \exp(-f(g(t))) \overset{(1)}{\geq} \exp(-f(g(t)))\frac{1}{g(t) f'(g(t))}\ ,$$

where in (1) we used $f'(t) = \omega\left(t^{-1}\right)$. We have

$$\exp(f(g(t))) f'(g(t))\,\dot{g}(t) g(t) \geq 1 \tag{92}$$

In addition

$$\int \left[\exp\left(f\left(g\left(t\right)\right)\right) f'\left(g\left(t\right)\right) \dot{g}(t) g(t)\right] dt \overset{(1)}{=} \int u \exp\left(f\left(u\right)\right) f'\left(u\right) du$$
$$= \int u \frac{d}{du}\left[\exp\left(f\left(u\right)\right)\right] du$$
$$= u \exp\left(f\left(u\right)\right) - \int \exp\left(f\left(u\right)\right) du$$
$$\leq u \exp\left(f\left(u\right)\right) \ , \tag{93}$$

where in $(1)$ we defined $u = g(t)$ and used $du = g'(t)dt$.
Combining the last two equations we obtain

$$\begin{aligned} g(t) \exp\left(f\left(g(t)\right)\right) &\geq t + C_2 \\ f\left(g(t)\right) + \log\left(g\left(t\right)\right) &\geq \log\left(t + C_2\right) \\ f\left(g(t)\right) &\geq \log\left(t + C_2\right) - \log\left(g\left(t\right)\right) \\ &\geq \log\left(t + C_2\right) - \log\left(f^{-1}\left(\log\left(t + C_1\right)\right)\right) \\ &\geq \log\left(t + C_2\right) - h\left(t\right), \end{aligned}$$

where we defined $h\left(t\right) = \log\left(f^{-1}\left(\log\left(t + C_1\right)\right)\right)$.
Thus,

$$g\left(t\right) \geq f^{-1}\left(\log\left(t + C_2\right) - h\left(t\right)\right) = f^{-1}\left(\log\left(t + C_2\right)\right) + o\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)$$

since

$$\lim_{t\to\infty} \frac{f^{-1}\left(\log\left(t + C_2\right) - h\left(t\right)\right)}{f^{-1}\left(\log\left(t + C_2\right)\right)} = 1.$$

This is true from the Squeeze Theorem since we have

$$f^{-1}\left(\log\left(t + C_2\right) - h\left(t\right)\right) \leq f^{-1}\left(\log\left(t + C_2\right)\right)$$

because $f$ is increasing and $h(t) \geq 0$ (for sufficiently large $t$), and also, from the gradient inequality (we recall that $f^{-1}$ is convex since $f$ is concave and increasing)

$$\begin{aligned} f^{-1}\left(\log\left(t + C_2\right) - h\left(t\right)\right) &\geq f^{-1}\left(\log\left(t + C_2\right)\right) - \frac{1}{f'\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)} h\left(t\right) \\ &= f^{-1}\left(\log\left(t + C_2\right)\right) - o\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right) \ , \end{aligned}$$

where in the last equality we used

$$\begin{aligned} &\lim_{t\to\infty} \frac{h\left(t\right)}{f^{-1}\left(\log\left(t + C_2\right)\right) f'\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)} \\ &= \lim_{t\to\infty} \frac{\log\left(f^{-1}\left(\log\left(t + C_1\right)\right)\right)}{f^{-1}\left(\log\left(t + C_2\right)\right) f'\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)} \\ &= \lim_{t\to\infty} \underbrace{\frac{\log\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)}{\log^{1+\epsilon}\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)}}_{\to 0} \cdot \underbrace{\frac{\log\left(f^{-1}\left(\log\left(t + C_1\right)\right)\right)}{\log\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)}}_{\to 1} \cdot \underbrace{\frac{\log^{1+\epsilon}\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)}{f^{-1}\left(\log\left(t + C_2\right)\right) f'\left(f^{-1}\left(\log\left(t + C_2\right)\right)\right)}}_{O(1)} \\ &= 0 \, , \end{aligned}$$

where in the last transition we used $\frac{\log\left(f^{-1}(\log(t+C_1))\right)}{\log(f^{-1}(\log(t+C_2)))} \to 1$ from the next claim.

**Claim 4.** $\forall C_1, C_2$: $\lim_{t\to\infty} \frac{\log\left(f^{-1}(\log(t+C_1))\right)}{\log(f^{-1}(\log(t+C_2)))} = 1$

*Proof.* We assume WLOG $C_2 > C_1$. Thus,

$$\log\left(f^{-1}\left(\log\left(t+C_1\right)\right)\right) = \log\left(f^{-1}\left(\log\left(t+C_2\right) + \log\left(\frac{t+C_1}{t+C_2}\right)\right)\right) \leq \log\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right)$$

and

$$\begin{aligned}
&\log\left(f^{-1}\left(\log\left(t+C_2\right) + \log\left(\frac{t+C_1}{t+C_2}\right)\right)\right) \\
&\geq \log\left(f^{-1}\left(\log\left(t+C_2\right)\right) + \frac{\log\left(\frac{t+C_1}{t+C_2}\right)}{f'\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right)}\right) \\
&= \log\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right) + \log\left(1 + \frac{\log\left(\frac{t+C_1}{t+C_2}\right)}{f^{-1}\left(\log\left(t+C_2\right)\right) f'\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right)}\right) \\
&= \log\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right) + o\left(1\right)
\end{aligned}$$

□

**Third step: proving that** $g(t) = f^{-1}\left(\log(t)\right) + o\left(f^{-1}\left(\log(t)\right)\right)$

We have

$$f^{-1}\left(\log\left(t+C_2\right)\right) + o\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right) \leq g\left(t\right) \leq f^{-1}\left(\log\left(t+C_1\right)\right).$$

Using Claim 5, this eq. also implies

$$f^{-1}\left(\log\left(t+C_2\right)\right) + o\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right) \leq g\left(t\right) \leq f^{-1}\left(\log\left(t+C_2\right)\right) + o\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right).$$

Therefore, we have

$$\lim_{t\to\infty} \frac{g\left(t\right)}{f^{-1}\left(\log\left(t\right)\right)} = \frac{g\left(t\right)}{f^{-1}\left(\log\left(t+C_2\right)\right)} = 1.$$

**Claim 5.** $\forall C_1, C_2$: $\lim_{t\to\infty} \frac{f^{-1}(\log(t+C_1))}{f^{-1}(\log(t+C_2))} = 1$

*Proof.* We assume WLOG $C_2 > C_1$. Thus,

$$f^{-1}\left(\log\left(t+C_1\right)\right) = f^{-1}\left(\log\left(t+C_2\right) + \log\left(\frac{t+C_1}{t+C_2}\right)\right) \leq f^{-1}\left(\log\left(t+C_2\right)\right)$$

and

$$\begin{aligned}
f^{-1}\left(\log\left(t+C_2\right) + \log\left(\frac{t+C_1}{t+C_2}\right)\right) &\geq f^{-1}\left(\log\left(t+C_2\right)\right) + \frac{\log\left(\frac{t+C_1}{t+C_2}\right)}{f'\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right)} \\
&= f^{-1}\left(\log\left(t+C_2\right)\right) + o\left(f^{-1}\left(\log\left(t+C_2\right)\right)\right)
\end{aligned}$$

□

□

### E.4 $g(t)=\log(t)+o(\log(t))$ proof for $L>1$ and f(u)=u

We have that

$$\frac{dg\left(t\right)}{dt}=\exp\left(-g\left(t\right)\right)g^{b}(t),\ b\in\left[1,2\right]. \tag{94}$$

We can write

$$g(t)=\log(t)+b\log(g(t))+h(t). \tag{95}$$

First step: we want to show that $h(t)\leq\log(\log(t)+C_2)$. Substituting eq. (95) into eq. (94) we get:

$$\begin{aligned}t^{-1}+b\frac{\dot{g}(t)}{g(t)}+h'(t)&=t^{-1}\exp(-h(t))\\ t^{-1}+h'(t)&=t^{-1}\exp(-h(t))\left(1-\frac{b}{g(t)}\right).\end{aligned}$$

Since $\frac{b}{g(t)}<0$ and $t^{-1}>0$ we get

$$h'(t)\leq t^{-1}\exp(-h(t)).$$

Integrating both sides we get:

$$\begin{aligned}\exp(h(t))&\leq\log(t)+C\\ h(t)&\leq\log\left(\log(t)+C\right)\\ &=\log\left(\left(1+\frac{C}{\log(t)}\right)\log(t)\right)\\ &=\log\left(\log(t)\right)+\log\left(1+\frac{C}{\log(t)}\right)\end{aligned}$$

and thus, $\exists t_2,C_2>1$ so that $\forall t>t_2$ :

$$h(t)\leq\log\left(\log(t)+C_2\right).$$

Step 2: Showing that $g(t)\leq C_4\log(t)$

$$g(t)=log(t)+b\log(g(t))+h(t)\leq log(t)+b\log(g(t))+\log\left(\log(t)+C_2\right).$$

Since $g(t)-b\log(g(t))=\Theta(g(t))$, $\exists t_3,C_3,C_4$ so that $\forall t>t_3$:

$$g(t)\leq C_3\left(\log(t)+\log\left(\log(t)+C_2\right)\right)\leq C_4\log(t).$$

Step 3: Showing that $g(t)\geq\log(t)$

We define $s(t)=\exp(g(t))\Rightarrow\dot{s}(t)=\exp(g(t))\dot{g}(t)=g^b(t)=\left[\log(s(t))\right]^b$. Note that $g(t)\rightarrow\infty$ implies $s(t)\rightarrow\infty$. We have

$$\lim_{t\rightarrow\infty}\frac{t}{s(t)}=\lim_{t\rightarrow\infty}\frac{1}{\left[\log(s(t))\right]^b}=0$$

and therefore $s(t)=\omega(t)$. This implies that $\exists t_4$ so that $\forall t>t_4$:

$$\begin{aligned}s(t)&\geq t\\ \exp(g(t))&\geq t\\ g(t)&\geq\log(t).\end{aligned}$$

Combining the results from steps 2 and 3 we obtain $g(t) = \theta\left(\log(t)\right)$.

$$1 \leq \frac{g(t)}{\log(t)} = \frac{g(t) - b\log\left(g(t)\right)}{\log(t)} + \frac{b\log\left(g(t)\right)}{log(t)} \leq \frac{\log(t) + \log\left(\log(t) + C_2\right)}{\log(t)} + \frac{b\log\left(C_3\left(\log(t) + \log\left(\log(t) + C_2\right)\right)\right)}{\log(t)} \to 1$$

From the Squeeze Theorem $\lim_{t\to\infty}\frac{g(t)}{log(t)} = 1$. Thus, $g(t) = \log(t) + o\left(\log(t)\right)$.

### E.5 Proof that $\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \Omega\left(\frac{1}{\log(t)}\right)$

In this section, we need to prove that if $f'(u) = \omega(1)$ and $\left|\frac{f'(u)}{f(u)}\right| = O(u^{-1})$ then $\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \Omega\left(\frac{1}{\log(t)}\right)$. From the calculation in appendix sections E.1, E.2, we have that exists $C > 0$ so that $\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = C\frac{1}{g(t)f'(g(t))}$ where $g(t) = f^{-1}(\log(t)) + o(f^{-1}(\log(t)))$ from Wong (2018). From $\left|\frac{f'(u)}{f(u)}\right| = O(u^{-1})$ we have that

$$f^{-1}(\log(t))f'(f^{-1}(\log(t))) = O\left(f^{-1}(\log(t))\frac{f\left(f^{-1}(\log(t))\right)}{f^{-1}(\log(t)}\right) = O(\log(t)) \,.$$

Thus, combining this result with $g(t) = f^{-1}(\log(t)) + o(f^{-1}(\log(t)))$ and $\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = C\frac{1}{g(t)f'(g(t))}$ we get that

$$\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \Omega\left(\frac{1}{\log(t)}\right)$$

as required.

## F Proof of Theorem 9

### F.1 Preliminaries and Auxiliary Lemma

Recall Assumption 2

**Assumption 2.** *$f(u)$ is real analytic on $\mathbb{R}_{++}$ and satisfies* $\forall k \in \mathbb{N}: \left|\frac{f^{(k+1)}(u)}{f'(u)}\right| = O\left(u^{-k}\right)$.

**Claim 6.** *For any function $f(u)$ that satisfies assumption 2, and any two functions $g(t),\ h(t)$ such that $h(t) = o(g(t))$ and $\lim_{t\to\infty} g(t) = \infty$, $\exists t_1$ so that $\forall t > t_1$:*

$$f\left(g(t) + h(t)\right) = f\left(g(t)\right) + f'\left(g(t)\right)h(t) + R(t),$$

*where $R(t) = o\left(f'\left(g(t)\right)h(t)\right)$.*

*Proof.* Since $h(t) = o(g(t))$ and $\lim_{t\to\infty} g(t) = \infty$ we have that $\exists t_1$ so that $\forall t > t_1 : g(t) + h(t) > 0$. From our assumption that $f$ is real analytic on $\mathbb{R}_{++}$ we get that $\forall t > t_1$:

$$f\left(g(t) + h(t)\right) = f\left(g(t)\right) + f'\left(g(t)\right)h(t) + \sum_{k=2}^{\infty}\frac{1}{k!}f^{(k)}(g(t))h^k(t).$$

We denote $R(t) = \sum_{k=2}^{\infty}\frac{1}{k!}f^{(k)}(g(t))h^k(t)$. We need to show that $\lim_{t\to\infty}\frac{R(t)}{f'(g(t))h(t)} = 0$.

Since $\forall k \in \mathbb{N}: \left|\frac{f^{(k+1)}(t)}{f'(t)}\right| = O\left(t^{-k}\right)$ we have that $\exists t_2 > t_1$ and positive constants $\{C_k\}_{k=2}^{\infty}$, so that $\forall t > t_2$:

$$\begin{aligned} 0 \leq \left|\frac{R(t)}{f'\left(g(t)\right)h(t)}\right| &\leq \left|\frac{1}{f'\left(g(t)\right)h(t)}\sum_{k=2}^{\infty}C_k f'(g(t))\frac{h^k(t)}{\left(g(t)\right)^{k-1}}\right| \\ &= \left|\sum_{k=2}^{\infty}C_k\left(\frac{h(t)}{g(t)}\right)^{k-1}\right| \leq \max_k C_k\left|\frac{\frac{h(t)}{g(t)}}{1 - \frac{h(t)}{g(t)}}\right| \xrightarrow{t\to\infty} 0 \end{aligned} \tag{96}$$

where in the last transition we used $h(t) = o(g(t))$. Thus, by the squeeze theorem $\lim_{t\to\infty} \frac{R(t)}{f'(g(t))\, h(t)} = 0$. □

The following lemma will be useful in characterizing $\mathbf{w}(t)$ asymptotic behaviour.

**Lemma 5.** *If* $\lim_{t\to\infty} \sum_{n=1}^{N} f_n(t)\mathbf{x}_n = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n$ *where* $\mathbf{x}_n$ *are linearly independent vectors then* $\forall n \in \{1, ..., N\}$ : $\lim_{t\to\infty} f_n(t) = \alpha_n$.

*Proof.* $\lim_{t\to\infty} \sum_{n=1}^{N} f_n(t)\mathbf{x}_n = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n$ implies that $\forall \epsilon > 0$, $\exists t'$ so that $\forall t > t'$ :

$$\left\| \sum_{n=1}^{N} f_n(t)\mathbf{x}_n - \sum_{n=1}^{N} \alpha_n \mathbf{x}_n \right\| < \epsilon.$$

Since $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent, $\forall k\ \exists \mathbf{u}$ such that $\forall n \neq k : \langle \mathbf{x}_n, \mathbf{u} \rangle = 0$ and $\langle \mathbf{x}_k, \mathbf{u} \rangle = 1$. Using Cauchy-Schwarz inequality we have that, $\forall k = 1, ..., n$ and $\forall t > t'$:

$$|f_k(t) - \alpha_k| = \left| \left\langle \sum_{n=1}^{N} f_n(t)\mathbf{x}_n - \sum_{n=1}^{N} \alpha_n \mathbf{x}_n, \mathbf{u} \right\rangle \right| \leq \left\| \sum_{n=1}^{N} f_n(t)\mathbf{x}_n - \sum_{n=1}^{N} \alpha_n \mathbf{x}_n \right\| \|\mathbf{u}\| < \|\mathbf{u}\| \epsilon.$$

□

Using Theorem 2, Lemma 4, Lemma 5 and the claim 6 we will prove Theorem 9 that characterizes $\mathbf{w}(t)$ asymptotic behavior.

### F.2 Proof of Theorem 9

**Theorem 9.** *Under Assumption 2 and the conditions and notations of Theorem 2, the equivalent linear predictor of a depth $L$ linear network will behave as:*

$$\mathbf{w}(t) = \tilde{g}(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t) \tag{82}$$

*where $\boldsymbol{\rho}(t) = o(\tilde{g}(t))$, $\rho(t)^\top \hat{\mathbf{w}} = 0$, and $\hat{\mathbf{w}}$ is the $L_2$ max margin separator*

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \text{ s.t. } \mathbf{w}^\top \mathbf{x}_n \geq 1.$$

*Further, $\tilde{g}(t)$ and $\boldsymbol{\rho}(t)$ are the asymptotic solution of the following,*

$$\lim_{t\to\infty} \frac{L\eta_t \exp\left(-f\left(\tilde{g}\left(t\right)\right)\right) \left(\tilde{g}(t)\right)^{2\left(1-\frac{1}{L}\right)} \phi_1(t)}{\gamma^{1-2/L} \frac{d}{dt}\tilde{g}(t)} = 1, \textit{and}$$

$$\|\mathbf{Q}\boldsymbol{\rho}(t)\| = \phi_2(t) \left(f'\left(\tilde{g}(t)\right)\right)^{-1} + o\left(\left(f'\left(\tilde{g}(t)\right)\right)^{-1}\right).$$

*where $\phi_1(t) = \Theta(1)$ and $\phi_2(t) = \Theta(1)$ are positive functions that depend only on the data set and not on the loss function $\ell$; $\gamma = \min_n \frac{\hat{\mathbf{w}}^\top \mathbf{x}_n}{\|\hat{\mathbf{w}}\|}$ is the maximum-margin attainable for the dataset with unit $L_2$ norm separator; $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is the orthogonal projection matrix to the subspace spanned by the support vectors.*
*If, in addition, the support vector span the dataset then*

$$\left\|\bar{\mathbf{Q}}\boldsymbol{\rho}(t)\right\| = O(1).$$

*Proof.* From Theorem 2 we have that $\lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}$ where $\hat{\mathbf{w}}$ is the maximum-margin separator. In addition, from lemma 4 we have that $\lim_{t\to\infty} \|\mathbf{w}(t)\| = \infty$. Combining these two results we can write

$$\mathbf{w}(t) = \tilde{g}(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t), \tag{97}$$

where $\lim_{t\to\infty}\tilde{g}(t)=\infty$, $\|\boldsymbol{\rho}(t)\|=o\left(\tilde{g}(t)\right)$ and $\boldsymbol{\rho}(t)^\top\hat{\mathbf{w}}=0$.

Using Claim 6 we have,

$$\forall n\in\mathcal{S}:\ f\left(\tilde{g}(t)+\boldsymbol{\rho}(t)^\top\mathbf{x}_n\right)=f\left(\tilde{g}(t)\right)+f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top\mathbf{x}_n+R_n(t), \tag{98}$$

where $R_n(t)=o\left(f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top\mathbf{x}_n\right)$. We denote $\mathcal{S}=\left\{n:\hat{\mathbf{w}}^\top\mathbf{x}_n=1\right\}$ (the indices of the support vectors) and recall that the maximum-margin separator can be expresses as a linear combination of the support vectors $\hat{\mathbf{w}}=\sum\limits_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n$. In order to calculate the convergence rates we need to characterize the asymptotic behavior of $\|\boldsymbol{\rho}(t)\|$ and $g(t)$.

**Step 1:** Proving that $\forall n\in\mathcal{S}:\ \psi_n(t)\triangleq f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top\mathbf{x}_n+R_n(t)=O(1)$.

We want to prove that $\forall n\in\mathcal{S}$ $\psi_n(t)$ is asymptotically bounded, i.e. $\exists t_1>0,m,M$ so that $\forall t>t_1: m\leq\psi_n(t)\leq M$.

We assume in contradiction that $\exists k\in\mathcal{S}$ so that $\psi_k(t)$ is not asymptotically bounded from above. Using Bolzano-Weierstrass theorem we have that $\exists\left\{\bar{t}_i\right\}_{i=1}^{\infty}$ so that the sequence $\psi_k(\bar{t}_i)\xrightarrow{i\to\infty}\infty$.

From Theorem 2 and Lemma 4 $\exists\left\{\tilde{\alpha}_n\geq0\ :\ n\in\mathcal{S}\right\}$ which are a positive scaling of $\alpha_n$ so that

$$\frac{\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\mathbf{w}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n}{\left\|\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\mathbf{w}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n\right\|}\to\sum_{n\in\mathcal{S}}\tilde{\alpha}_n\mathbf{x}_n. \tag{99}$$

Substituting eq. (97) into eq. (99) we get that

$$\begin{aligned}\frac{\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\mathbf{w}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n}{\left\|\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\mathbf{w}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n\right\|}&\overset{(1)}{=}\frac{\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\tilde{g}(t)\hat{\mathbf{w}}^\top\mathbf{x}_n+\boldsymbol{\rho}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n}{\left\|\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\tilde{g}(t)\hat{\mathbf{w}}^\top\mathbf{x}_n+\boldsymbol{\rho}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n\right\|}\\ \overset{(2)}{=}\frac{\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\tilde{g}(t)\right)-\psi_n(t)\right)\mathbf{x}_n}{\left\|\sum\limits_{n\in\mathcal{S}}\exp\left(-f\left(\tilde{g}(t)\right)-\psi_n(t)\right)\mathbf{x}_n\right\|}&=\frac{\sum\limits_{n\in\mathcal{S}}\exp\left(-\psi_n(t)\right)\mathbf{x}_n}{\left\|\sum\limits_{n\in\mathcal{S}}\exp\left(-\psi_n(t)\right)\mathbf{x}_n\right\|}\to\sum_{n\in\mathcal{S}}\tilde{\alpha}_n\mathbf{x}_n\ ,\end{aligned} \tag{100}$$

where in (1) we used eq. (97), in (2) we used $\forall n\in\mathcal{S}:\hat{\mathbf{w}}^\top\mathbf{x}_n=1$ and eq. (98). From the last equation we also have that

$$\frac{\sum\limits_{n\in\mathcal{S}}\exp\left(-\psi_n(\bar{t}_i)\right)\mathbf{x}_n}{\left\|\sum\limits_{n\in\mathcal{S}}\exp\left(-\psi_n(\bar{t}_i)\right)\mathbf{x}_n\right\|}\xrightarrow{i\to\infty}\sum_{n\in\mathcal{S}}\tilde{\alpha}_n\mathbf{x}_n. \tag{101}$$

Combining the last equation with the facts that for almost every dataset $\forall n\in\mathcal{S}:\tilde{\alpha}_n>0$ (as a positive scaling of $\alpha_n$) and $\mathbf{x}_n$ are linearly independent (from Lemma 8 in Soudry et al. (2018a)) and Lemma 5 we have that $\forall n\in\mathcal{S}$ : $\exp\left(-\psi_n(\bar{t}_i)\right)=\Theta\left(\left\|\sum_{m=1}^{N}\exp\left(-\psi_m(\bar{t}_i)\right)\mathbf{x}_m\right\|\right)$.

Thus, if for some $k\in\mathcal{S}$, $\psi_k(\bar{t}_i)\xrightarrow{i\to\infty}\infty$, then this implies that $\forall n\in\mathcal{S}$ : $\psi_n(\bar{t}_i)\xrightarrow{i\to\infty}\infty$. In addition, since $\psi_n(t)\triangleq f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top\mathbf{x}_n+R_n(t)$ where $R_n(t)=o\left(f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top\mathbf{x}_n\right)$ we get that $\forall n\in\mathcal{S}:f'\left(\tilde{g}(\bar{t}_i)\right)\boldsymbol{\rho}(\bar{t}_i)^\top\mathbf{x}_n\to\infty$. However, this implies

$$0\overset{(1)}{=}f'\left(\tilde{g}(\bar{t}_i)\right)\hat{\mathbf{w}}^\top\boldsymbol{\rho}(\bar{t}_i)\overset{(2)}{=}\sum_{n\in\mathcal{S}}\alpha_n f'\left(\tilde{g}(\bar{t}_i)\right){\mathbf{x}_n}^\top\boldsymbol{\rho}(\bar{t}_i)\xrightarrow[i\to\infty]{(3)}\infty$$

where in (1) we used $\forall t:\ \hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)=0$, in (2) we used $\hat{\mathbf{w}}=\sum\limits_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n$ and in (3) we used that $\alpha_n>0$. We got a contradiction and thus our contradiction assumption must be false $\Rightarrow\forall k\in\mathcal{S}:\psi_k(t)$ is bounded from above. Similarly, $\forall k\in\mathcal{S}:\psi_k(t)$ is bounded from below since $\tilde{\alpha}_n>0$ for all $n\in\mathcal{S}$. Combining these results we have that $\forall n\in\mathcal{S}$ : $\psi_n(t)=O(1)$.

**Step 2:** Characterizing $\boldsymbol{\rho}(t)$ and $\tilde{g}(t)$ asymptotic behavior.

From the previous step we have that $\forall n \in \mathcal{S} : \psi_n(t) = O(1)$. We recall $\psi_n$ definition $\psi_n(t) \triangleq f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n + R_n(t)$ where $R_n(t) = o\left(f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)$. This implies that $\forall n \in \mathcal{S} : f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n = O(1)$.

Since this is true $\forall n \in \mathcal{S}$ we have that $\boldsymbol{\rho}(t)$ components that are in subspace spanned by the support vectors are bounded, i.e. $f'(\tilde{g}(t)) \|\mathbf{Q}_1\boldsymbol{\rho}(t)\| = O(1) \Rightarrow \|\mathbf{Q}_1\boldsymbol{\rho}(t)\| = O\left(\left(f'(\tilde{g}(t))\right)^{-1}\right)$. In the next steps we will further characterize $\|\mathbf{Q}_1\boldsymbol{\rho}(t)\|$ behaviour. We denote $\mathbf{z}(t) \triangleq -\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}(t)) = -\sum_{n=1}^{N} \ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n$.

**Step 2.a:** Showing that $\lim_{t\to\infty} \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|} = \lim_{t\to\infty} \frac{1}{\|\mathbf{z}(t)\|} \exp\left(-f\left(\tilde{g}(t)\right)\right) \sum_{n\in\mathcal{S}} \exp\left(-f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n$

From Theorem 2 we have that $\lim_{t\to\infty} \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|}$ exists (and finite). In addition,

$$\begin{aligned}
\lim_{t\to\infty} \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|} &\overset{(1)}{=} \lim_{t\to\infty} \frac{-\sum_{n\in\mathcal{S}} \ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n}{\|\mathbf{z}(t)\|} \\
&\overset{(2)}{=} \lim_{t\to\infty} \frac{\sum_{n\in\mathcal{S}} \exp\left(-f\left(\tilde{g}(t)\hat{\mathbf{w}}^\top \mathbf{x}_n + \boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)\right)\mathbf{x}_n}{\|\mathbf{z}(t)\|} \\
&\overset{(3)}{=} \lim_{t\to\infty} \frac{\exp\left(-f\left(\tilde{g}(t)\right)\right) \sum_{n\in\mathcal{S}} \exp\left(-f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n - R_n(t)\right)\mathbf{x}_n}{\|\mathbf{z}(t)\|} \\
&\overset{(4)}{=} \lim_{t\to\infty} \frac{\exp\left(-f\left(\tilde{g}(t)\right)\right) \sum_{n\in\mathcal{S}} \exp\left(-f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n}{\|\mathbf{z}(t)\|}\,,
\end{aligned}$$

where in (1) we used $\mathbf{z}(t)$ definition and Lemma 4, in (2) we used $-\ell'(u) = -\exp\left(-f(u)\right)$ and $\mathbf{w}(t) = \tilde{g}(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t)$ (eq. (97)), in 3 we used eq. (98) and in 4 we used the fact that $\forall n \in \mathcal{S}:\ \psi_n(t) \triangleq f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n + R_n(t) = O(1)$ and $R_n(t) = o\left(f'(\tilde{g}(t))\, \boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)$ which implies $R_n(t) \to 0$.

**Step 2.b:** Showing that for a constant $C$: $\lim_{t\to\infty} \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|} = \lim_{t\to\infty} C \frac{\frac{d}{dt}\tilde{g}(t)}{\|\mathbf{z}(t)\| \left(\tilde{g}(t)\right)^{\frac{2}{L}(L-1)}} \sum_{n\in\mathcal{S}} \alpha_n \mathbf{x}_n$

From eq. (81) we have that

$$\lim_{t\to\infty} \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|} = \lim_{t\to\infty} \frac{\bar{\mathbf{w}}_\infty h(t) g^{L-1}(t)}{\|\mathbf{z}(t)\| \left\|\mathbf{W}_{2:L}(t)\right\|^2}. \tag{102}$$

In this equation, $\mathbf{W}_{2:L}^{(t)} \triangleq \mathbf{W}_2(t)\mathbf{W}_3(t)...\mathbf{W}_L(t)$ and $h(t)$ and $g(t)$ were used to model each layer weights:

$$\begin{aligned}
\Delta\mathbf{W}_l(t) &= \overline{\mathbf{W}}_l^\infty h(t) + \delta_{\Delta\mathbf{W}_l}(t)h(t), \\
\mathbf{W}_l(t) &= \overline{\mathbf{W}}_l^\infty g(t) + \delta_{\mathbf{W}_l}(t)g(t),
\end{aligned}$$

where $g(t) = \sum_{u<t} \eta_u h(u)$, $\Delta\mathbf{W}_l(t) \triangleq \eta_t^{-1}\left(\mathbf{W}_l(t+1) - \mathbf{W}_l(t)\right)$, $\overline{\mathbf{W}}_\infty = \lim_{t\to\infty} \frac{\mathcal{W}(t)}{\|\mathcal{W}(t)\|}$ and $\overline{\mathbf{W}}_\infty = \left[\overline{\mathbf{W}}_l^\infty\right]$. From definition we have $h(t) = \eta_t^{-1}\left(g(t) - g(t-1)\right)$. Additionally, we can define $\bar{g}(u) = \bar{g}\left(\frac{t}{c}\right) = g(t),\ u = \frac{t}{c}$ (time rescaling) to show that $\lim_{t\to\infty} \frac{g(t)-g(t-1)}{g'(t)} = 1$ since:

$$\lim_{c\to\infty} \frac{\bar{g}(u+\frac{1}{c}) - \bar{g}(u)}{\frac{1}{c}} = \bar{g}'(u) \implies \lim_{c\to\infty} \frac{\bar{g}(u+\frac{1}{c}) - \bar{g}(u)}{\frac{1}{c}\bar{g}'(u)} = 1.$$

$$\text{Also, } \frac{d}{dt}g(t) = \frac{1}{c}\bar{g}'(u) \implies 1 = \lim_{c\to\infty} \frac{\bar{g}(u+\frac{1}{c}) - \bar{g}(u)}{\frac{1}{c}\bar{g}'(u)} = \lim_{t\to\infty} \frac{g(t+1)-g(t)}{g'(t)}$$

where in the last transition we used $\bar{g}(u)$ definition and $t = cu$. Combining this result with eq. (102) (using the fact that both limits exist and finite) we have:

$$\lim_{t\to\infty} \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|} = \lim_{t\to\infty} \frac{\bar{\mathbf{w}}_\infty \eta_t^{-1} g'(t) g^{L-1}(t)}{\|\mathbf{z}(t)\| \left\|\mathbf{W}_{2:L}(t)\right\|^2}. \tag{103}$$

In addition, since we defined $\mathbf{w}(t) = \tilde{g}(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t)$ we have that $\tilde{g}(t) = \gamma\left(g(t)\right)^L$ where $\bar{\mathbf{w}}_\infty = \overline{\mathbf{W}}_1^\infty ... \overline{\mathbf{W}}_L^\infty$ and $\gamma \triangleq \min_n \bar{\mathbf{w}}_\infty^\top \mathbf{x}_n$. Using $\tilde{g}(t)$ eq. (103) can be written as

$$\lim_{t\to\infty} \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|} = \lim_{t\to\infty} \frac{\frac{\bar{\mathbf{w}}_\infty}{\gamma L \eta_t} \cdot \frac{d}{dt}\tilde{g}(t)}{\|\mathbf{z}(t)\|\left(\gamma^{-1}\tilde{g}(t)\right)^{\frac{2}{L}(L-1)}} = \lim_{t\to\infty} \frac{\gamma^{1-2/L}}{L\eta_t} \cdot \frac{\frac{d}{dt}\tilde{g}(t)\sum_{n\in\mathcal{S}}\alpha_n \mathbf{x}_n}{\|\mathbf{z}(t)\|\left(\tilde{g}(t)\right)^{\frac{2}{L}(L-1)}}, \tag{104}$$

where in the last step we used $\frac{\bar{\mathbf{w}}_\infty}{\gamma} = \hat{\mathbf{w}} = \sum_{n\in\mathcal{S}} \alpha_n \mathbf{x}_n$.

**Step 2.c:**

Combining the result from the previous steps (using the fact that both limits exist and finite, $\mathbf{x}_n$ are linearly independent and Lemma 5), we get, $\forall n \in \mathcal{S}$:

$$\lim_{t\to\infty} \frac{\eta_t L \exp\left(-f\left(\tilde{g}\left(t\right)\right)\right)\left(\tilde{g}(t)\right)^{\frac{2}{L}(L-1)}}{\gamma^{1-2/L}\frac{d}{dt}\tilde{g}(t)} \cdot \frac{\exp\left(-f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)}{\alpha_n} = 1.$$

We denote $\zeta_n(t) \triangleq \dfrac{1}{\alpha_n}\exp\left(-f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top\mathbf{x}_n\right) > 0$. $\zeta_n(t) = \Theta(1)$ since $f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top \mathbf{x}_n = O(1)$. Substituting this into the limit we get $\forall n \in \mathcal{S}$

$$\lim_{t\to\infty} \frac{\exp\left(-f\left(\tilde{g}\left(t\right)\right)\right)\left(\tilde{g}(t)\right)^{\frac{2}{L}(L-1)}}{\frac{d}{dt}\tilde{g}(t)} \cdot \frac{L\eta_t\zeta_n(t)}{\gamma^{1-2/L}} = 1,$$

which implies $\forall n_1, n_2 \in \mathcal{S} : \frac{\zeta_{n_1}(t)}{\zeta_{n_2}(t)} \to 1$. Therefore, we can write

$$\forall n \in \mathcal{S}: \ \exp\left(-f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top \mathbf{x}_n\right) = \alpha_n\left(\zeta(t) + \zeta(t)\tilde{\delta}_n(t)\right),$$

where $\zeta(t) \triangleq \zeta_1(t)$ and $\forall n \in \mathcal{S}: \ \tilde{\delta}_n(t) \to 0$.

**Claim 7.** *For a given depth L, $\zeta(t)$ asymptotic behaviour is independent of $f$, i.e., $\lim_{t\to\infty}\dfrac{\zeta^{(1)}(t)}{\zeta^{(2)}(t)} = 1$ where $\zeta^{(1)}(t)$ and $\zeta^{(2)}(t)$ correspond to two different function $f_1, f_2$.*

*Proof.* Let $f_1(t), f_2(t)$ be different functions with $\zeta^{(1)}(t), \zeta^{(2)}(t) > 0$ so that $\forall n \in \mathcal{S}$ :

$$\exp\left(-f_1'\left(\tilde{g}_1(t)\right)\boldsymbol{\rho}_1(t)^\top \mathbf{x}_n\right) = \alpha_n\left(\zeta^{(1)}(t) + \zeta^{(1)}(t)\tilde{\delta}_n^{(1)}(t)\right) \tag{105}$$

$$\exp\left(-f_2'\left(\tilde{g}_2(t)\right)\boldsymbol{\rho}_2(t)^\top \mathbf{x}_n\right) = \alpha_n\left(\zeta^{(2)}(t) + \zeta^{(2)}(t)\tilde{\delta}_n^{(2)}(t)\right). \tag{106}$$

We need to show that $\lim_{t\to\infty}\dfrac{\zeta^{(1)}(t)}{\zeta^{(2)}(t)} = 1$. Since we know $\zeta(t) > 0$, $\exists t_2$ such that $\forall t > t_2: \ \zeta^{(2)}(t) + \zeta^{(2)}(t)\tilde{\delta}_n^{(2)}(t) > 0$ and $\zeta^{(1)}(t) + \zeta^{(1)}(t)\tilde{\delta}_n^{(1)}(t) > 0$. We define $\forall n \in \mathcal{S} : \tilde{\zeta}_n(t) = \log\left(\dfrac{\zeta^{(2)}(t) + \zeta^{(2)}(t)\tilde{\delta}_n^{(2)}(t)}{\zeta^{(1)}(t) + \zeta^{(1)}(t)\tilde{\delta}_n^{(1)}(t)}\right)$ (this is well defined $\forall t > t_2$). From the last two equations we have:

$$\forall n \in \mathcal{S}: \ \left(f_1'\left(\tilde{g}_1(t)\right)\boldsymbol{\rho}_1(t) - f_2'\left(\tilde{g}_2(t)\right)\boldsymbol{\rho}_2(t)\right)^\top \mathbf{x}_n = \tilde{\zeta}_n(t).$$

Additionally, since $\boldsymbol{\rho}_1(t)^\top\hat{\mathbf{w}} = \boldsymbol{\rho}_2(t)^\top\hat{\mathbf{w}} = 0$ (from definition) we get $\forall t > t_2$:

$$\begin{aligned}
0 &= \left(f_1'\left(\tilde{g}_1(t)\right)\boldsymbol{\rho}_1(t) - f_2'\left(\tilde{g}_2(t)\right)\boldsymbol{\rho}_2(t)\right)^\top \hat{\mathbf{w}} = \sum_{n\in\mathcal{S}}\alpha_n\left(f_1'\left(\tilde{g}_1(t)\right)\boldsymbol{\rho}_1(t) - f_2'\left(\tilde{g}_2(t)\right)\boldsymbol{\rho}_2(t)\right)^\top \mathbf{x}_n \\
&= \sum_{n\in\mathcal{S}}\alpha_n\tilde{\zeta}_n(t) = \sum_{n\in\mathcal{S}}\alpha_n\log\left(\frac{\zeta^{(2)}(t) + \zeta^{(2)}(t)\tilde{\delta}_n^{(2)}(t)}{\zeta^{(1)}(t) + \zeta^{(1)}(t)\tilde{\delta}_n^{(1)}(t)}\right) = \sum_{n\in\mathcal{S}}\alpha_n\log\left(\frac{\zeta^{(2)}(t)}{\zeta^{(2)}(t)}\cdot\frac{1+\tilde{\delta}_n^{(2)}(t)}{1+\tilde{\delta}_n^{(1)}(t)}\right) \\
&= \log\left(\frac{\zeta^{(2)}(t)}{\zeta^{(2)}(t)}\right)\sum_{n\in\mathcal{S}}\alpha_n + \sum_{n\in\mathcal{S}}\alpha_n\log\left(\frac{1+\tilde{\delta}_n^{(2)}(t)}{1+\tilde{\delta}_n^{(1)}(t)}\right).
\end{aligned}$$

Additionally, $\lim_{t\to\infty}\sum_{n\in\mathcal{S}}\alpha_n\log\left(\frac{1+\tilde{\delta}_n^{(2)}(t)}{1+\tilde{\delta}_n^{(1)}(t)}\right)=0$ since $\forall n\in\mathcal{S}:\ \tilde{\delta}_n^{(1)}(t)\to 0,\tilde{\delta}_n^{(2)}(t)\to 0$. Combining both results we get $\log\left(\frac{\zeta^{(2)}(t)}{\zeta^{(2)}(t)}\right)\sum_{n\in\mathcal{S}}\alpha_n\to 0$. Since for almost every dataset $\forall n\in\mathcal{S}:\alpha_n>0$, this implies $\log\left(\frac{\zeta^{(2)}(t)}{\zeta^{(2)}(t)}\right)\to 0\Rightarrow \lim_{t\to\infty}\frac{\zeta^{(1)}(t)}{\zeta^{(2)}(t)}=1$. □

Summarizing, we have that $\boldsymbol{\rho}(t)$ and $\tilde{g}(t)$ are the asymptotic solutions of:

$$\frac{\exp\left(-f\left(\tilde{g}\left(t\right)\right)\right)\left(\tilde{g}(t)\right)^{\frac{2}{L}(L-1)}}{\frac{d}{dt}\tilde{g}(t)}\cdot\frac{L\eta_t\zeta(t)}{\gamma^{1-2/L}}=1,\ \text{and} \tag{107}$$

$$\forall n\in\mathcal{S}:\ \exp\left(-f'\left(\tilde{g}(t)\right)\boldsymbol{\rho}(t)^\top\mathbf{x}_n\right)=\alpha_n\left(\zeta(t)+\zeta(t)\tilde{\delta}_n(t)\right). \tag{108}$$

where $\zeta(t)=\Theta(1)$ is independent of $f$ and $\tilde{\delta}_n(t)\to 0$. Thus, for almost every dataset[3], $\exists\tau(t)=\Theta(1)$, that is only dependent on the data set and $L$ (specifically, it is not dependent on the loss function) such that $\|\mathbf{Q}\boldsymbol{\rho}(t)\|=\tau(t)(f'\left(\tilde{g}(t)\right))^{-1}+o\left(\left(f'\left(\tilde{g}(t)\right)\right)^{-1}\right)$ where $\mathbf{Q}\in\mathbb{R}^{d\times d}$ is the projection matrix to the subspace spanned by the support vectors. This completes our proof with $\phi_1(t)=\zeta(t)$ and $\phi_2(t)=\tau(t)$. □

## G Proof that non-support vectors direction converge for $L=1$

**Theorem 10.** *For $L=1$, $\beta$-smooth loss and $\eta<2\beta^{-1}$, if $f'(t)=\Omega\left(\frac{\log^{1+\epsilon}(t)}{t}\right)$ for some $\epsilon>0$ then*

$$\left\|\bar{\mathbf{Q}}\boldsymbol{\rho}(t)\right\|=O(1),$$

*where $\mathbf{Q}\in\mathbb{R}^{d\times d}$ is the orthogonal projection matrix to the subspace spanned by the support vectors and $\bar{\mathbf{Q}}=I-\mathbf{Q}$ is the complementary projection.*

*Proof.* From Theorem 9 we have that

$$\mathbf{r}(t)=\mathbf{w}(t)-\hat{\mathbf{w}}g(t)$$

where $\mathbf{r}(t)=o(g(t))$. We aim to show that $\left\|\bar{\mathbf{Q}}\mathbf{r}(t)\right\|$ is bounded. We have

$$\left\|\bar{\mathbf{Q}}\mathbf{r}(t+1)\right\|^2=\left\|\bar{\mathbf{Q}}\mathbf{r}(t+1)-\bar{\mathbf{Q}}\mathbf{r}(t)\right\|^2+2\left(\bar{\mathbf{Q}}\mathbf{r}(t+1)-\bar{\mathbf{Q}}\mathbf{r}(t)\right)^\top\bar{\mathbf{Q}}\mathbf{r}(t)+\left\|\bar{\mathbf{Q}}\mathbf{r}(t)\right\|^2. \tag{109}$$

1.
$$\left\|\bar{\mathbf{Q}}\mathbf{r}(t+1)-\bar{\mathbf{Q}}\mathbf{r}(t)\right\|^2=\left\|-\eta\bar{\mathbf{Q}}\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\right\|^2\le\eta^2\left\|\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\right\|^2 \tag{110}$$

Additionally, From Lemma 7 we have

$$\sum_{u=1}^{\infty}\left\|\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\right\|^2<\infty. \tag{111}$$

[3]Note that this excludes the degenerate case in which $\forall n\in\mathcal{S}:\alpha_n=1$. We can show this similarly to Lemma 12 in Soudry et al. (2018b).

2.

$$
\begin{aligned}
&\left(\bar{\mathbf{Q}}\mathbf{r}(t+1)-\bar{\mathbf{Q}}\mathbf{r}(t)\right)^\top\bar{\mathbf{Q}}\mathbf{r}(t)\\
&\overset{(1)}{=}\eta\sum_{n\notin\mathcal{S}}\exp\left(-f\left(\mathbf{w}^\top\mathbf{x}_n\right)\right)\mathbf{x}_n{}^\top\bar{\mathbf{Q}}\mathbf{r}(t)\\
&=\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}\exp\left(-f\left(g(t)\hat{\mathbf{w}}^\top\mathbf{x}_n+\mathbf{r}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n{}^\top\bar{\mathbf{Q}}\mathbf{r}(t)\\
&\overset{(2)}{\leq}\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}\exp\left(-f\left(g(t)\theta+\mathbf{r}(t)^\top\mathbf{x}_n\right)\right)\mathbf{x}_n{}^\top\bar{\mathbf{Q}}\mathbf{r}(t)\\
&\overset{(3)}{=}\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}\exp\left(-f\left(g(t)\theta\right)-f'\left(\theta g(t)\right)\mathbf{r}(t)^\top\mathbf{x}_n\left(1+\delta_n(t)\right)\right)\mathbf{x}_n{}^\top\bar{\mathbf{Q}}\mathbf{r}(t)\\
&=\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}\exp\left(-f\left(g(t)\theta\right)-\left(1+\delta_n(t)\right)f'\left(\theta g(t)\right)\mathbf{Q}\mathbf{r}(t)^\top\mathbf{x}_n-\left(1+\delta_n(t)\right)f'\left(\theta g(t)\right)\bar{\mathbf{Q}}\mathbf{r}(t)^\top\mathbf{x}_n\right)\mathbf{x}_n{}^\top\bar{\mathbf{Q}}\mathbf{r}(t)\\
&\overset{(4)}{\leq}\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}\exp\left(-f\left(g(t)\theta\right)+\left(1+\delta_n(t)\right)f'\left(\theta g(t)\right)\left\|\mathbf{Q}\mathbf{r}(t)\right\|\left\|\mathbf{x}_n\right\|\right)\left\|\mathbf{x}_n{}^\top\bar{\mathbf{Q}}\right\|\left\|\mathbf{r}(t)\right\|\\
&\overset{(5)}{\leq}\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}\exp\left(-f\left(g(t)\theta\right)+\left(1+\delta_n(t)\right)f'\left(\theta g(t)\right)\left\|\mathbf{Q}\mathbf{r}(t)\right\|\left\|\mathbf{x}_n\right\|\right)g(t)\\
&\overset{(6)}{\leq}\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}C\exp\left(-f\left(\theta g(t)\right)\right)g(t)=\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\bar{\mathbf{Q}}\mathbf{r}(t)>0}}C\exp\left(-f\left(\theta g(t)\right)+\log\left(g(t)\right)\right)\\
&\leq\eta NC\exp\left(-f\left(\theta g(t)\right)+\log\left(g(t)\right)\right),\forall t>t_2,
\end{aligned}
\tag{112}
$$

where in 1 we used eq. (2) (gradient descent dynamic) and the fact that $\forall n\in\mathcal{S}:\bar{\mathbf{Q}}\mathbf{x}_n=0$, in 2 we used the fact that $f(t)$ is monotonically decreasing and $\theta=\min_{n\notin\mathcal{S}}\hat{\mathbf{w}}^\top\mathbf{x}_n>1$. In 3 we define $\delta_n\to 0$ and used Claim 6 and the fact that $\left|\mathbf{r}(t)^\top\mathbf{x}_n\right|=o(g(t))$. In 4 we used Cauchy-Schwarz and the fact that $\exists t_1$ so that $\forall t>t_1$: $\left(1+\delta_n(t)\right)f'\left(\theta g(t)\right)\bar{\mathbf{Q}}\mathbf{r}(t)^\top\mathbf{x}_n\geq 0$. In 5 we used $\|\mathbf{r}(t)\|=o\left(g\left(t\right)\right)$. In 6, we used the fact that $\left(1+\delta_n(t)\right)f'\left(\theta g(t)\right)\left\|\mathbf{Q}\mathbf{r}(t)\right\|\left\|\mathbf{x}_n\right\|=\Theta(1)$ since $\left\|\mathbf{Q}\mathbf{r}(t)\right\|=\Theta\left(\frac{1}{f'(g(t))}\right)$ and $\frac{f'(\theta g(t)}{f'(g(t))}=\Theta(1)$ ($f'(\theta g(t))=f'(g(t))+(\theta-1)f''(g(t))g(t)+o\left(f''(g(t))g(t)\right)=O\left(f'(g(t))\right)$ from assumption 2). This implies that $\exists C_1,t_2$ so that $\forall t>t_2>t_1$: $\left(1+\delta_n(t)\right)f'\left(\theta g(t)\right)\left\|\mathbf{Q}\mathbf{r}(t)\right\|\left\|\mathbf{x}_n\right\|\leq C_1$ and we define $C=\exp(C_1)$.

3.

$$
\begin{aligned}
\left\|\bar{\mathbf{Q}}\mathbf{r}(t)\right\|^2-\left\|\bar{\mathbf{Q}}\mathbf{r}(t_1)\right\|^2&=\sum_{u=t_1}^{t-1}\left[\left\|\bar{\mathbf{Q}}\mathbf{r}(u+1)\right\|^2-\left\|\bar{\mathbf{Q}}\mathbf{r}(u)\right\|^2\right]\\
&\overset{(1)}{=}\sum_{u=t_1}^{t-1}\left[\left\|\bar{\mathbf{Q}}\mathbf{r}(u+1)-\bar{\mathbf{Q}}\mathbf{r}(u)\right\|^2+2\left(\bar{\mathbf{Q}}\mathbf{r}(u+1)-\bar{\mathbf{Q}}\mathbf{r}(u)\right)^\top\bar{\mathbf{Q}}\mathbf{r}(t)\right]\\
&\leq\eta^2\sum_{u=t_1}^{t-1}\left\|\nabla\mathcal{L}\left(\mathbf{w}(u)\right)\right\|^2+2\eta NC\sum_{u=t_1}^{t-1}\exp\left(-f\left(\theta g(u)\right)+\log\left(g(u)\right)\right)\overset{(3)}{<}\infty
\end{aligned}
$$

where in (1) we used eq. (109), in (2) we used eqs. 110 and 112 and in (3) we used the last transition we used eq. (111) and Claim 8.

□

### G.1 Integrability proof

**Claim 8.** *If* $\exists\epsilon > 0$ *so that* $f'(u) = \Omega\left(\frac{\log^{1+\epsilon}(u)}{u}\right)$ *and* $g(t)$ *satisfies the following equation*

$$\frac{dg(t)}{dt} = \exp\left(-f(g(t))\right) . \tag{113}$$

*then*

$$\int_0^\infty \exp\left(-f(g(t)c) + \log(g(t)\right) dt < \infty,$$

*where* $c > 1$.

*Proof.* We have

$$\begin{aligned}
&\int_0^\infty \left[\exp\left(-f(g(t)c) + \log(g(t))\right] dt \\
&\overset{(1)}{=} \int_{g(0)}^\infty \exp\left(-\left(f(gc) - f(g)\right) + \log(g)\right) dg \\
&\overset{(2)}{\leq} C + \int_{t_0}^\infty \exp\left(-\frac{C_1}{1+\epsilon_1}\left(\log^{2+\epsilon}(gc) - \log^{2+\epsilon}(g)\right) + \log(g)\right) dg \\
&\overset{(3)}{\leq} C + \int_{t_0}^\infty \exp\left(-\frac{C_1}{1+\epsilon_1}(2+\epsilon)\log(c)\log^{1+\epsilon}(g) + \log(g)\right) dg \\
&= C + \int_{t_0}^\infty \exp\left(-\log(g)\left(\frac{C_1}{1+\epsilon_1}(2+\epsilon)\log^{\epsilon}(t) - 1\right)\right) dg \\
&\overset{(4)}{\leq} C + \int_{t_0}^\infty \exp\left(-C'\log(g)\right) dg < \infty
\end{aligned}$$

where in (1) we used variable change, $g(t) \to \infty$, eq. (113) and $\frac{d}{dx}f^{-1}(x) = \left[f'\left(f^{-1}(x)\right)\right]^{-1}$, in (2) we used $g(t) \to \infty$ and $f'(u) = \Omega\left(\frac{\log^{1+\epsilon}(u)}{u}\right)$ for some $\epsilon > 0$ and therefore $\exists t_1$ so that $\forall t > t_1$:

$$\begin{aligned}
&f(gc) - f(g) \\
&= \int_{g(t)}^{cg(t)} f'(u) du \\
&\geq \int_{g(t)}^{cg(t)} C_1 \frac{\log^{1+\epsilon}(u)}{u} du \\
&= \frac{C_1}{1+\epsilon}\left(\log^{2+\epsilon}(cg(t)) - \log^{2+\epsilon}(g(t))\right) .
\end{aligned}$$

Additionally, we defined $C = \int_0^{t_0} \exp\left(-\left(f(gc) - f(g)\right)\right) dg$. In (3) we used the fact that $\forall a \geq 1$:

$$(\log(gc))^a = (\log(c) + \log(g))^a = \log^a(g)\left(1 + \frac{\log(c)}{\log(g)}\right)^a \geq \log^a(g) + a\log(c)\log^{a-1}(g)$$

from Bernoulli's inequality (since $\frac{\log(c)}{\log(g)} \geq -1$ for sufficently large t). In (4) we used the fact that $\exists C' > 1$ since for sufficently large t $\frac{C_1}{1+\epsilon_1}(2+\epsilon)\log^{\epsilon}(t) - 1 > 1$. ☐

## H Examples

We recall Theorem 2, regarding general tails:

**Theorem 2.** *For any depth* $L$*, almost all linearly separable datasets, almost all initialization and any bounded sequence of step sizes* $\{\eta_t\}$*, consider the sequence* $\mathcal{W}(t) = \{\mathbf{W}_l(t)\}_{l=1}^L$ *of gradient descent updates in eq. (6) for* *minimizing the empirical loss $\mathcal{L}_{\mathcal{P}}(\mathcal{W})$ (eq. (5)) with a strictly monotone loss function $\ell$ satisfying Assumption 1, i.e.: $\ell'(u) = -\exp(-f(u)) < 0$, where asymptotically $f'(u) > 0$ and $f'(u) = \omega\left(u^{-1}\right)$.*

*If (a) $\mathcal{W}(t)$ minimizes the empirical loss, i.e. $\mathcal{L}_{\mathcal{P}}\left(\mathcal{W}(t)\right) \to 0$, (b) $\mathcal{W}(t)$, and consequently $\mathbf{w}(t) = \mathcal{P}\left(\mathbf{w}(t)\right)$, converge in direction to yield a separator with positive margin, and (c) the gradients with respect to the linear predictors $\nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{w}(t)\right)$ converge in direction, then the limit direction is given by,*

$$\bar{\mathbf{w}}_{\infty} = \lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|} \,,$$

*where*

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \;\text{ s.t. } \mathbf{w}^\top \mathbf{x}_n \geq 1. \tag{8}$$

In this section we demonstrate this Theorem using two examples.

### H.1 Example: non-convergence to the max-margin separator

Theorem 2 assumes that $f'(t) = \omega\left(t^{-1}\right)$ which implies $f(t) = \omega(\log(t))$, and therefore, if $f(t) = O(\log(t))$, we may not converge to the max margin separator, i.e. $\lim_{t\to\infty} \mathbf{w}(t)/\|\mathbf{w}(t)\| \neq \hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|$. Next, we give an example for such a case. Consider optimization with a power-law tailed loss

$$\ell\left(u\right) = \begin{cases} u^{-1} & ,u > 1 \\ 2-u & ,u \leq 1 \end{cases} ,$$

with two data points $\mathbf{x}_1 = (1, 0)$ and $\mathbf{x}_2 = (0, 2)$. In this case $\hat{\mathbf{w}} = (1, 0.5)$ and $\|\hat{\mathbf{w}}\| = \sqrt{5}/2$. We take the limit $\eta \to 0$, and obtain the continuous time version of GD:

$$\dot{w}_1\left(t\right) = \frac{1}{w_1^2\left(t\right)} \;;\; \dot{w}_2\left(t\right) = \frac{0.5}{w_2^2\left(t\right)}.$$

We can analytically integrate these equations to obtain

$$\frac{1}{3}w_1^3\left(t\right) = t + C \;;\; \frac{1}{3}w_2^3\left(t\right) = 0.5t + C.$$

so

$$w_1\left(t\right) = \sqrt[3]{3t + w_1^3\left(0\right)} \;;\; w_2\left(t\right) = \sqrt[3]{1.5t + w_2^3\left(0\right)}.$$

therefore, as $t \to \infty$

$$w_1\left(t\right)/w_2\left(t\right) \to \sqrt[3]{2} \neq 2 = \hat{w}_1/\hat{w}_2.$$

In this case, asymptotically we have that $f(t) = 2\log(t)$ which is not $\omega(\log(t))$, and therefore in this case Theorem 2 should not apply. Thus, we expect that $h(t)$ will not be $o(g(t))$, as we assumed, and this will break the analysis in appendix section C (specifically, $g(t) \asymp t^{\frac{1}{3}}$ and $h(t) = 1/f'(g(t)) \asymp t^{\frac{1}{3}}$). Using this example, it is easy to verify that we do not converge to the max-margin separator whenever $-\ell'(u)$ is polynomial.

In contrast, it is straightforward to verify, that similar analysis on the same example, only with poly-exponential tails, does yield convergence to the max-margin, as expected. For example, with exponential loss we obtain $-\ell'\left(u\right) = e^{-u}$

$$\lim_{t\to\infty} w_1\left(t\right)/w_2\left(t\right) = 2 = \hat{w}_1/\hat{w}_2$$

In this case, $f(u) = u$, $g(t) \asymp \log(t)$, and $h(t) = f'(g(t)) = 1$, and so these results are consistent with Theorem 1.

### H.2 Example: sub-poly-exponential tails that converge to the max margin separator

Theorem 2 implies that if $-\ell'(u)$ has a tail that decays faster than any polynomial tail, we will still converge to the max margin. To demonstrate this we analyze the same example as before, only with

$$-\ell'(u) = \begin{cases} \exp\left(-\log^{\epsilon}(u) - \log\left(\epsilon \log^{\epsilon-1}(u)\right) + \log(u)\right) & , u > 2 \\ \exp\left(-\log^{\epsilon}(2) - \log\left(0.5\epsilon \log^{\epsilon-1}(2)\right)\right) & , u \le 2 \end{cases}$$

for constant $\epsilon > 1$. In this case $f(t) = \Theta(\log^{\epsilon}(t)) = \omega(\log(t))$. We get:

$$\dot{w}_1(t) = \exp\left(-\log^{\epsilon}(w_1(t)) - \log\left(\epsilon \log^{\epsilon-1}(w_1(t))\right) + \log(w_1(t))\right)$$
$$\dot{w}_2(t) = 2\exp\left(-\log^{\epsilon}(2w_2(t)) - \log\left(\epsilon \log^{\epsilon-1}(2w_2(t))\right) + \log(2w_2(t))\right)$$

We can analytically integrate these equations to obtain

$$\exp\left(\log^{\epsilon} w_1(t)\right) = t + C \ ; \ \exp\left(\log^{\epsilon}(2w_2(t))\right) = 4t + C.$$

so

$$w_1(t) = \exp\left(\log^{\epsilon^{-1}}(t + \tilde{C}_1)\right) \ ; \ w_2(t) = \frac{1}{2}\exp\left(\log^{\epsilon^{-1}}(4t + \tilde{C}_2)\right),$$

where

$$\tilde{C}_1 = \exp\left(\log^{\epsilon}(w_1(0))\right) \ ; \ \tilde{C}_2 = \exp\left(\log^{\epsilon}(2w_2(0))\right),$$

therefore, as $t \to \infty$

$$w_1(t)/w_2(t) \to 2 = \hat{w}_1/\hat{w}_2$$

However, we note that for $\epsilon < 1$ $w_1(t)/w_2(t) \to 0$ and for $\epsilon = 1$ $w_1(t)/w_2(t) \to 0.5$, meaning that for $\epsilon \le 1$ we do not converge to the max margin separator, which is consistent with the conjecture, since then $f(t) = O(\log(t))$.

### H.3 Example: Demonstrating that the upper bound in Theorem 3.1 is not always obtained

We analyze the same example as before, only with

$$\ell'(u) = -\frac{1}{\nu}\exp\left(-u^{\nu} - (\nu - 1)\log(u)\right),$$

for some $\nu > 1$. In this case we get:

$$w_1(t) = \log^{\frac{1}{\nu}}(t + C_1) \ ; \ w_2(t) = \frac{1}{2}\log^{\frac{1}{\nu}}(2t + C_2),$$

where $C_1 = \exp\left(w_1(0)^{\nu}\right)$ and $C_2 = \exp\left((2w_2(0))^{\nu}\right)$. Therefore, as $t \to \infty$

$$w_1(t)/w_2(t) \to 2 = \hat{w}_1/\hat{w}_2.$$

Recall that the max-margin solution for this case is $\hat{\mathbf{w}} = (1, 0.5)$ and $\|\hat{\mathbf{w}}\| = \sqrt{5}/2$. We can write

$$\mathbf{w}(t) = \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix} = \left(\log^{\frac{1}{\nu}}(t + C_1) + \frac{1}{4}\log^{\frac{1}{\nu}}(2t + C_2)\right) \cdot \frac{4}{5}\hat{\mathbf{w}} + \left(\log^{\frac{1}{\nu}}(t + C_1) - \log^{\frac{1}{\nu}}(2t + C_2)\right) \cdot \frac{2}{5}\hat{\mathbf{w}}^{\perp},$$

where $\hat{\mathbf{w}}^{\perp} = \begin{bmatrix} \frac{1}{2} \\ -1 \end{bmatrix}$, $\left\langle \hat{\mathbf{w}}, \hat{\mathbf{w}}^{\perp} \right\rangle = 0$. Therefore, in this example $\mathbf{w}(t) = g(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t)$ where $g(t) = \left(\log^{\frac{1}{\nu}}(t + C_1) + \frac{1}{4}\log^{\frac{1}{\nu}}(2t + C_2)\right) \cdot \frac{4}{5}$ and $\boldsymbol{\rho}(t) = \left(\log^{\frac{1}{\nu}}(t + C_1) - \log^{\frac{1}{\nu}}(2t + C_2)\right) \cdot \frac{2}{5}\hat{\mathbf{w}}^{\perp} \approx C\frac{1}{\nu}\log^{\frac{1}{\nu}-1}(t) \to 0$ for some constant $C > 0$ independent of $\nu$. This implies that the margin convergence rate is proportional to $\frac{1}{\nu\log(t)}$, i.e., we obtain the same asymptotic rate as exponential loss, only with better constants.

### H.4 Example: Demonstrating that the upper bound in Theorem 3.1 is tight

Next, we give an example to show that the rate upper bound $O\left(\frac{1}{f^{-1}(\log(t))}\right)$ for $f'(u) = \omega(1)$ is tight. Consider optimization with a loss that satisfies

$$-\ell'(u) = \exp(-f(u))$$

for some function $f'(u) = \omega(1)$ with one data point $\mathbf{x}_1 = (1, 0)$. In this case $\hat{\mathbf{w}} = (1, 0)$. We take the limit $\eta \to 0$, and obtain the continuous time version of GD:

$$\dot{w}_1(t) = \exp(-f(w_1(t))) \;\; ; \;\; \dot{w}_2(t) = 0.$$

From Wong (2018) we have that $w_1(t) = \Theta\left(f^{-1}(\log(t)\right)$ and from integrating the right equation we obtain $w_2(t) = w_2(0)$. Thus, using this example with $w_2(0) > 0$, we see that the above upper bound is tight.

## I Numerical results: additional details

### I.1 Implementation details of Figure 1

The original dataset included four support vectors: $\mathbf{x}_1 = (0.5, 1.5), \mathbf{x}_2 = (1.5, 0.5)$ with $y_1 = y_2 = 1$, and $\mathbf{x}_3 = -\mathbf{x}_1$, $\mathbf{x}_4 = -\mathbf{x}_2$ with $y_3 = y_4 = -1$. The $L_2$ normalized max margin vector in this case was $\hat{\mathbf{w}} = \frac{1}{2}(1, 1)$ with margin equal to $\sqrt{2}$. Additional 6 random data points were added from each class. These additional points are sufficiently far from the origin so they are not support vectors. Lastly, we re-scaled all datapoints so that $\max_n ||\mathbf{x}_n|| < 1$, according to our assumption.

For training, we initialized $\mathbf{w}(0) \sim \mathcal{N}(0, \mathbf{I}_d)$, and used the optimal $\eta = 1/\beta$ for GD, and the same as initial step size for normalized GD.

Note that, in panel C, the training error $\mathcal{L}(\mathbf{w}(t))$ of normalized GD converges to zero (much faster than GD) — until it disappears when reaching the lowest numerical precision level. Also, the margin gap figure for normalized GD appears less stable for Normalized GD. We suspect that this is because the index of the datapoint with the smallest margin rapidly switched due to the aggressive learning rate used.

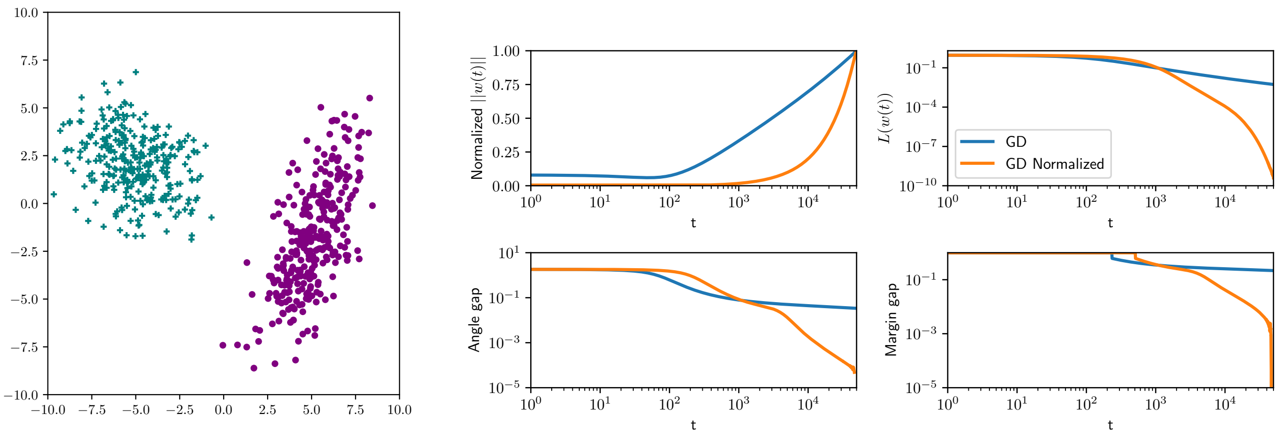
### I.2 Neural Networks on a Toy Dataset

In what follows we compare GD to Normalized GD on linear and non-linear neural networks. For this purpose, we generate a 2-dimensional synthetic dataset composed of 600 data points, where positive and negative samples are generated from $\mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma}^+)$ and $\mathcal{N}(\boldsymbol{\mu}^-, \boldsymbol{\Sigma}^-)$, respectively, with $\boldsymbol{\mu}^+ = (-5, 2)$, $\boldsymbol{\Sigma}^+ = \left[\begin{smallmatrix} 3 & -1 \\ -1 & 3 \end{smallmatrix}\right]$, $\boldsymbol{\mu}^- = (5, -2)$ and $\boldsymbol{\Sigma}^- = \left[\begin{smallmatrix} 2 & 3 \\ 3 & 9 \end{smallmatrix}\right]$. Once the dataset was generated, the same points were used for all the following experiments.

We use a learning rate $\eta$ of $0.005$, which was empirically chosen so that optimization is stable but not slow. Larger learning rates would often result in both GD and Normalized GD presenting convergence issues, as in difficulty to reach (or stay at) a solution that separates the data. The weights were initialized from $\mathcal{N}(0, 0.1)$, and gradients were normalized together: $\sqrt{\sum_{i=1}^{d} \|\nabla_{\mathbf{W}_i} \mathcal{L}\|_F^2}$ was used to normalize each parameter's gradient, where $\mathbf{W}_i$ denotes the weight matrix of the $i$'th layer and $d$ the total number of layers of the network. Finally, each hidden layer contains 10 hidden neurons.

Figure 5 shows the dataset and the convergence of GD and Normalized GD on logistic regression. We can see that Normalized GD converges significantly faster, similarly to Figure 1. To compute angle and margin gaps, we obtain the $L_2$ max margin vector $\hat{\mathbf{w}}$ from a SVM solver, along with the max margin itself.

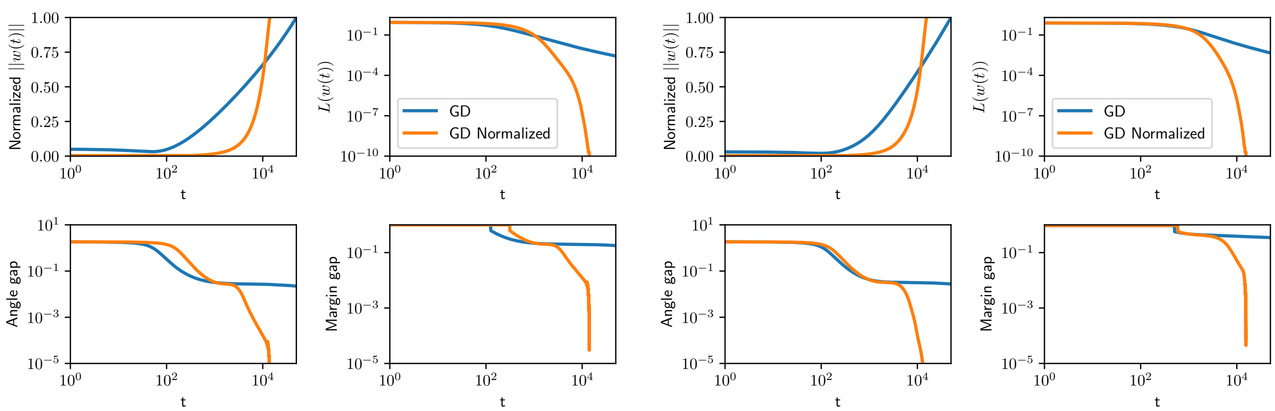
In Figure 6 we see the convergence of GD and Normalized GD for 2-layer neural networks, with and without a ReLU non-linearity. We can observe that there is little difference between all plots, suggesting that our results might translate to more complex models, at least in well-behaved settings such as when the data is linearly separable. Note that for the non-linear network, the angle and margin were computed using $\mathbf{w} = \mathbf{W}_1 \mathbf{W}_2 \ldots \mathbf{W}_d$, as if the model was a linear network. The same observation can be drawn from Figure 7, which depicts convergence for 3-layered networks.




(a) Dataset

(b) Logistic Regression (1 layer Linear Network)

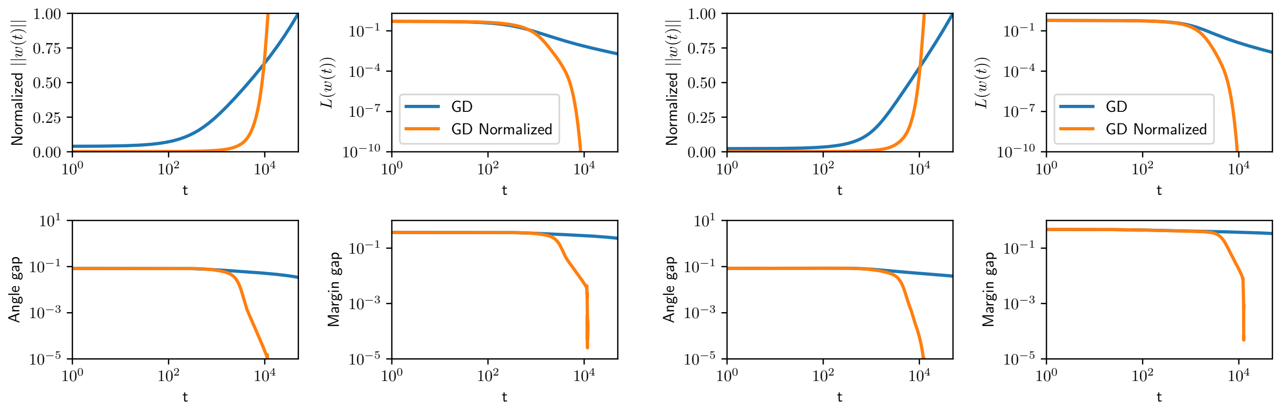
Figure 5: a) Visualization of the synthetic dataset composed of 600 points: 300 labeled positive and 300 negative, again respectively denoted by $'+'$ and $'\circ'$. b) Convergence plots for a logistic regression trained with GD and Normalized GD for $5 \times 10^4$ epochs. Similarly to what is observed in Figure 1, Normalized GD converges significantly faster to the max-margin solution.




(a) 2 layer Linear Network

(b) 2 layer ReLU Network

Figure 6: Convergence plots for 2-layered neural networks with a $2 \times 10 \times 1$ architecture, trained for $5 \times 10^4$ epochs with GD and Normalized GD. (a,b): networks with linear / ReLU activations, respectively. We can observe that the plots for linear and ReLU networks look similar, and for both models Normalized GD still converged noticeably faster to the max margin solution. Additionally, we can see that Normalized GD converged faster in the 2-layer setting when compared to Figure 5, achieving $0$ numerical loss in roughly $10^4$ epochs.




(a) 3 layer Linear Network (b) 3 layer ReLU Network

Figure 7: Convergence plots for 3-layered neural networks with a $2 \times 10 \times 10 \times 1$ architecture, trained for $5 \times 10^4$ epochs with GD and Normalized GD. (a,b): networks with linear / ReLU activations, respectively. As in Figure 6, we can observe that the plots of linear and ReLU networks are similar.

# J Losses with poly-exponential tails

In Theorem 2 we assume that the gradient descent iterates $\mathcal{W}(t)$ minimizes the objective, i.e., $\mathcal{L}_{\mathcal{P}}(\mathcal{W}(t)) \to 0$, and that the incremental updates $\mathcal{W}(t+1) - \mathcal{W}(t)$ converge in direction. In this section we show that in the case of a single layer, $L = 1$, and for a specific type of loss function these assumptions can be omitted.

**Definition 3.** *A function $f(u)$ has a "tight poly-exponential tail", if there exist positive constants $\mu_+, \mu_-, \nu$, and $\bar{u}$ such that $\forall u > \bar{u}$:*

$$(1 - \exp(-\mu_- u^\nu))e^{-u^\nu} \le f(u) \le (1 + \exp(-\mu_+ u^\nu))e^{-u^\nu}$$

**Theorem 11.** *For almost all datasets that are linearly separable and any $\beta$-smooth $\mathcal{L}$, with strictly monotone loss function $\ell$ (Definition 1) for which $-\ell'(u)$ has a tight poly-exponential tail (Definition 3) with $\nu > 0.25$, given step size $\eta < 2\beta^{-1}$ and any initialization $\mathbf{w}(0)$, the iterates of gradient descent in eq. (2) will behave as:*

$$\mathbf{w}(t) = \hat{\mathbf{w}} g(t) + \boldsymbol{\rho}(t), \tag{114}$$

*where $\hat{\mathbf{w}}$ is the following $L_2$ max margin separator:*

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \ \text{ s.t. } \mathbf{w}^\top \mathbf{x}_n \ge 1, \tag{115}$$

*and for a constant $\boldsymbol{a}$ independent of $\nu$,*

$$g(t) = \log^{\frac{1}{\nu}}(t) + \frac{1}{\nu} \log(\nu \log^{1-\frac{1}{\nu}}(t)) \log^{\frac{1}{\nu}-1}(t) \tag{116}$$

$$\|\boldsymbol{\rho}(t)\| = \begin{cases} O(1), & \textit{if } \nu > 1. \\ \frac{1}{\nu} g^{1-\nu}(t) \|\boldsymbol{a}\| + o(g^{1-\nu}(t)), & \textit{if } \frac{1}{4} < \nu \le 1, \end{cases} \tag{117}$$

As we show in the next section Theorem 11 implies that $\mathbf{w}(t)/\|\mathbf{w}(t)\|$ converges to the normalized max margin separator $\hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|$ for poly-exponential tails with $\nu > 0.25$, but with a different rate than exponential loss. In Appendix section K we show that Theorem 11 implies the convergence rates specified in Table 1. From this table, we can see that the optimal convergence rate for poly-exponential tails is achieved at $\nu = 1$. Moreover, this rate becomes slower as $|\nu - 1|$ increases, at least in the range $\nu > 0.25$.

Theorem 11 is proved in appendix section K.1. This Theorem is a generalization of Theorem 1, and therefore builds the ideas of Soudry et al. (2018a), as described non-rigorously in appendix section C. The main proof is rather long, as we calculate exact asymptotic behavior, including constants in some cases, and do not assume the existence of limits.

| | $\nu \geq 1$ | $\frac{1}{4} < \nu \leq 1$ |
|---|---|---|
| $\left\lVert \frac{\mathbf{w}(t)}{\lVert\mathbf{w}(t)\rVert} - \frac{\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert} \right\rVert$ or $\gamma - \min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\lVert\mathbf{w}(t)\rVert}$ | $O\left(\log^{-\frac{1}{\nu}}(t)\right)$ | $\frac{C_1}{\nu}\log^{-1}(t) + o(\log^{-1}(t))$ |
| $1 - \frac{\mathbf{w}(t)^\top\hat{\mathbf{w}}}{\lVert\mathbf{w}(t)\rVert\lVert\hat{\mathbf{w}}\rVert}$ | $O\left(\log^{-\frac{2}{\nu}}(t)\right)$ | $\frac{C_2}{\nu^2}\log^{-2}(t) + o(\log^{-2}(t))$ |

Table 1: Summary of convergence rates for Theorem 11 for loss functions with exponential tail, when $-\ell'(u) \asymp \exp(-u^\nu)$. The first line is the convergence rate for both the distance and the suboptimality of the margin (with $C_3$ instead of $C_1$). The second line is the angle convergence rate. The constants are:
$C_1 = \left\lVert \left(I - \frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^\top}{\lVert\hat{\mathbf{w}}\rVert^2}\right)\frac{\boldsymbol{a}}{\lVert\hat{\mathbf{w}}\rVert}\right\rVert,\ C_2 = \left(\frac{1}{4} - \left(\frac{\boldsymbol{a}^\top\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert\lVert\boldsymbol{a}\rVert}\right)^2\right)\frac{2\lVert\boldsymbol{a}\rVert^2}{\lVert\hat{\mathbf{w}}\rVert^2},\ C_3 = \frac{1}{\lVert\hat{\mathbf{w}}\rVert}\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{a}}{\lVert\hat{\mathbf{w}}\rVert^2} - \min_n \mathbf{x}_n^\top\boldsymbol{a}\right)$.

# K Calculation of convergence rates for poly-exponential tails

From Theorem 11, we can write $\mathbf{w}(t) = \hat{\mathbf{w}}g(t) + \boldsymbol{\rho}(t)$ where $\boldsymbol{\rho}(t) = o(g(t))$. We can use this to calculate the normalized weight vector:

$$\begin{aligned}
\frac{\mathbf{w}(t)}{\lVert\mathbf{w}(t)\rVert} &= \frac{g(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t)}{\sqrt{g(t)^2\hat{\mathbf{w}}^\top\hat{\mathbf{w}} + \boldsymbol{\rho}(t)^\top\boldsymbol{\rho}(t) + 2g(t)\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}} = \frac{\hat{\mathbf{w}} + g^{-1}(t)\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert\sqrt{1 + 2\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)\lVert\hat{\mathbf{w}}\rVert^2} + \frac{\lVert\boldsymbol{\rho}(t)\rVert^2}{g^2(t)\lVert\hat{\mathbf{w}}\rVert^2}}} \\
&\overset{(1)}{=} \frac{\hat{\mathbf{w}} + g^{-1}(t)\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert}\left[1 - \frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)\lVert\hat{\mathbf{w}}\rVert^2} + \left[\frac{3}{4}\left(2\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert^2}\right)^2 - \frac{\lVert\boldsymbol{\rho}(t)\rVert^2}{2\lVert\hat{\mathbf{w}}\rVert^2}\right]\frac{1}{g^2(t)} + O\left(\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)}\right)^3\right)\right] \\
&= \frac{\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert} + \left(\frac{\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert} - \frac{\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert}\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert^2}\right)\frac{1}{g(t)} + O\left(\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)}\right)^2\right)\frac{\hat{w}}{\lVert\hat{w}\rVert} \\
&= \frac{\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert} + \left(I - \frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^\top}{\lVert\hat{\mathbf{w}}\rVert^2}\right)\frac{1}{\lVert\hat{\mathbf{w}}\rVert}\frac{\boldsymbol{\rho}(t)}{g(t)} + O\left(\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)}\right)^2\right) \qquad (118)
\end{aligned}$$

where in (1) we used $\frac{1}{\sqrt{1+x}} = 1 - \frac{1}{2}x + \frac{3}{4}x^2 + O(x^3)$.
We use eq. (118) to calculate the angle:

$$\begin{aligned}
&\frac{\mathbf{w}(t)^\top\hat{\mathbf{w}}}{\lVert\mathbf{w}(t)\rVert\lVert\hat{\mathbf{w}}\rVert} \\
&= \frac{\hat{\mathbf{w}}^\top}{\lVert\hat{\mathbf{w}}\rVert^2}\left(\hat{\mathbf{w}} + g^{-1}(t)\boldsymbol{\rho}(t)\right)\left[1 - \frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)\lVert\hat{\mathbf{w}}\rVert^2} + \left[\frac{3}{4}\left(2\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert^2}\right)^2 - \frac{\lVert\boldsymbol{\rho}(t)\rVert^2}{2\lVert\hat{\mathbf{w}}\rVert^2}\right]\frac{1}{g^2(t)} + O\left(\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)}\right)^3\right)\right] \\
&= 1 + \frac{2}{\lVert\hat{\mathbf{w}}\rVert^2}\left[\left(\frac{\boldsymbol{\rho}(t)^\top\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert\lVert\boldsymbol{\rho}(t)\rVert}\right)^2 - \frac{1}{4}\right]\frac{\lVert\boldsymbol{\rho}(t)\rVert^2}{g^2(t)} + O\left(\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)}\right)^3\right) \qquad (119)
\end{aligned}$$

Calculation of the margin:

$$\begin{aligned}
\min_n \frac{\mathbf{x}_n^\top\mathbf{w}(t)}{\lVert\mathbf{w}(t)\rVert} &\overset{(1)}{=} \min_{n\in\mathcal{S}} \frac{\mathbf{x}_n^\top\mathbf{w}(t)}{\lVert\mathbf{w}(t)\rVert} \\
&= \min_{n\in\mathcal{S}} \mathbf{x}_n^\top\left[\frac{\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert} + \left(\frac{\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert} - \frac{\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert}\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert^2}\right)\frac{1}{g(t)} + O\left(\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)}\right)^2\right)\right] \\
&= \frac{1}{\lVert\hat{\mathbf{w}}\rVert} + \frac{1}{\lVert\hat{\mathbf{w}}\rVert}\left(\min_{n\in\mathcal{S}}\mathbf{x}_n^\top\boldsymbol{\rho}(t) - \frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{\lVert\hat{\mathbf{w}}\rVert^2}\right)\frac{1}{g(t)} + O\left(\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{\rho}(t)}{g(t)}\right)^2\right), \qquad (120)
\end{aligned}$$

where in (1) we used the fact that $\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$ converge to the maximum-margin separator and thus the minimal value is obtained on the support vectors. From equations 118, 119 we have:

$$\left\|\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}-\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}\right\|=\begin{cases}\left\|\left(I-\frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^\top}{\|\hat{\mathbf{w}}\|^2}\right)\boldsymbol{a}\right\|\frac{1}{\|\hat{\mathbf{w}}\|}\frac{\psi_2(t)}{g(t)f'(g(t))}, & \text{if } \frac{1}{f'(g(t))}=\Omega(1)\\ O(g^{-1}(t)), & \text{otherwise}\end{cases} \tag{121}$$

$$1-\frac{\mathbf{w}(t)^\top\hat{\mathbf{w}}}{\|\mathbf{w}(t)\|\|\hat{\mathbf{w}}\|}=\begin{cases}\left(\frac{1}{4}-\left(\frac{\boldsymbol{a}^\top\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|\|\boldsymbol{a}\|}\right)^2\right)\frac{2\|\boldsymbol{a}\|^2}{\|\hat{\mathbf{w}}\|^2}\frac{\psi_2^2(t)}{(g(t)f'(g(t)))^2}, & \text{if } \frac{1}{f'(g(t))}=\Omega(1)\\ O(g^{-2}(t)), & \text{otherwise}\end{cases} \tag{122}$$

Additionally, from Theorem 11, we can write $\mathbf{w}(t)=\hat{\mathbf{w}}g(t)+\boldsymbol{\rho}(t)$, where:

$$g(t)=\log^{\frac{1}{\nu}}(t)+\frac{1}{\nu}\log(\nu\log^{1-\frac{1}{\nu}}(t))\log^{\frac{1}{\nu}-1}(t) \tag{123}$$

$$\boldsymbol{\rho}(t)=\begin{cases}O(1), & \text{if } \nu>1.\\ \frac{1}{\nu}g^{1-\nu}(t)\boldsymbol{a}+o(g^{1-\nu}(t)), & \text{if } \frac{1}{4}<\nu<1,\end{cases} \tag{124}$$

and $\boldsymbol{a}$ is not dependent on $\nu$.
We can obtain the normalized weight vector convergence to normalized max margin vector in $L_2$ norm from substituting eqs. 123, 124 into eq. (121) :

$$\left\|\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}-\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}\right\|=\begin{cases}\left\|\left(I-\frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^\top}{\|\hat{\mathbf{w}}\|^2}\right)\frac{\boldsymbol{a}}{\|\hat{\mathbf{w}}\|}\right\|\frac{1}{\nu\log(t)}+o(\log^{-1}(t)), & \frac{1}{4}<\nu\le 1\\ O(\log^{-\frac{1}{\nu}}(t)), & \nu\ge 1\end{cases} \tag{125}$$

We can also obtain the angle convergence from substituting eqs. 123, 124 into eq. (122):

$$1-\frac{\mathbf{w}(t)^\top\hat{\mathbf{w}}}{\|\mathbf{w}(t)\|\|\hat{\mathbf{w}}\|}=\begin{cases}\left(\frac{1}{4}-\left(\frac{\boldsymbol{a}^\top\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|\|\boldsymbol{a}\|}\right)^2\right)\frac{2\|\boldsymbol{a}\|^2}{\|\hat{\mathbf{w}}\|^2}\frac{1}{\nu^2\log^2(t)}, & \frac{1}{4}<\nu\le 1\\ O(\log^{-\frac{2}{\nu}}(t)), & \nu\ge 1\end{cases} \tag{126}$$

We obtain the margin convergence from substituting eqs. 123, 124 into eq. (120):

$$\frac{1}{\|\hat{\mathbf{w}}\|}-\min_n\frac{\mathbf{x}_n^\top\mathbf{w}(t)}{\|\mathbf{w}(t)\|}=\begin{cases}\frac{1}{\|\hat{\mathbf{w}}\|}\left(\frac{\hat{\mathbf{w}}^\top\boldsymbol{a}}{\|\hat{\mathbf{w}}\|^2}-\min_n\mathbf{x}_n^\top\boldsymbol{a}\right)\frac{1}{\nu\log(t)}, & \frac{1}{4}<\nu\le 1\\ O(\log^{-\frac{1}{\nu}}(t)), & \nu\ge 1\end{cases} \tag{127}$$

We can see that in the case of $\nu<1$ the rates are smaller for larger $\nu$ and that the optimal rates are achieved for $\nu=1$.

### K.1 Proof of Theorem 11

In the following proofs, we define $\hat{\mathbf{w}}$ as the $L_2$ max margin vector, which satisfies eq. (4):

$$\hat{\mathbf{w}}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\|\mathbf{w}\|^2 \text{ s.t. } \mathbf{w}^\top\mathbf{x}_n\ge 1$$

Let $\mathcal{S}=\{n:\hat{\mathbf{w}}^\top\mathbf{x}_n=1\}$ denote indices of support vectors of $\hat{\mathbf{w}}$. From the KKT optimality conditions, we have for some $\alpha_n\ge 0$,

$$\hat{\mathbf{w}}=\sum_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n \tag{128}$$

Let $\tilde{\mathbf{w}}$ be a vector which satisfies the equations:

$$\forall n\in\mathcal{S}\ :\ \eta\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})=\alpha_n,\ \bar{\mathbf{P}}_1\tilde{\mathbf{w}}=0, \tag{129}$$

where we recall that we defined $\mathbf{P}_1 \in \mathbb{R}^d$ as the orthogonal projection matrix to the subspace spanned by the support vectors, and $\bar{\mathbf{P}}_1 = I - \mathbf{P}_1$ as the complementary projection matrix. Equation 129 has a unique solution for almost every dataset from Lemma 8 in Soudry et al. (2018a). Furthermore, let $C_i, \epsilon_i, t_i$ ($i \in \mathbb{N}$) be various positive constants which are independent of $t$, and denote,

$$\theta = \min_{n \notin \mathcal{S}} \mathbf{x}_n^\top \hat{\mathbf{w}} > 1 \tag{130}$$

The following lemmata were proved in Soudry et al. (2018a) [Lemma 1 and Lemma 5].

**Lemma 6.** *Let $\mathbf{w}(t)$ be the iterates of gradient descent (eq. (2)) on a $\beta$-smooth $\mathcal{L}$ and any starting point $\mathbf{w}(0)$. If the data is linearly separable, $\ell$ is a strict monotone loss (Definition 1), and $\eta < 2\beta^{-1}$ then we have: (1) $\lim_{t\to\infty} \mathcal{L}(\mathbf{w}(t)) = 0$, (2) $\lim_{t\to\infty} \|\mathbf{w}(t)\| = \infty$, and (3) $\forall n: \lim_{t\to\infty} \mathbf{w}(t)^\top \mathbf{x}_n = \infty$.*

**Lemma 7.** *Let $\mathcal{L}(\mathbf{w})$ be a $\beta$-smooth non-negative objective. If $\eta < 2\beta^{-1}$, then for any $\mathbf{w}(0)$, with the GD sequence*

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla \mathcal{L}(\mathbf{w}(t)) \tag{131}$$

*we have that $\sum_{u=0}^{\infty} \|\nabla \mathcal{L}(\mathbf{w}(u))\|^2 < \infty$ and therefore $\lim_{u\to\infty} \|\nabla \mathcal{L}(\mathbf{w}(u))\|^2 = 0$.*

From Lemma 6, $\forall n : \lim_{t\to\infty} \mathbf{w}(t)^\top \mathbf{x}_n = \infty$. In addition, we assume that the negative loss derivative $-\ell'(u)$ has a poly-exponential tail $e^{-u^\nu}$. Combining both facts, we have positive constants $\mu_+, \mu_-$ and $\bar{t}$ such that $\forall n, \forall t > \bar{t}$:

$$-\ell'(\mathbf{w}(t)^\top \mathbf{x}_n) \le (1 + \exp(-\mu_+ (\mathbf{w}(t)^\top \mathbf{x}_n)^\nu)) \exp(-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu) \tag{132}$$

$$-\ell'(\mathbf{w}(t)^\top \mathbf{x}_n) \ge (1 - \exp(-\mu_- (\mathbf{w}(t)^\top \mathbf{x}_n)^\nu)) \exp(-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu) \tag{133}$$

### K.2 Case: $\nu > 1$

#### K.2.1 Definitions and auxiliary calculations

In the following proofs, for any solution $\mathbf{w}(t)$, we define

$$\mathbf{r}(t) = \mathbf{w}(t) - g(t)\hat{\mathbf{w}} - g^{1-\nu}(t)\tilde{\mathbf{w}} \tag{134}$$

and the following functions:

$$g(t) = \log^{\frac{1}{\nu}}(t) + \frac{1}{\nu} \log(\nu \log^{1-\frac{1}{\nu}}(t)) \log^{\frac{1}{\nu}-1}(t), \tag{135}$$

$$h(t) = \left(1 - \frac{1}{\nu}\right) \log^{-1}(t) \left(1 - \log\left(\nu \log^{1-\frac{1}{\nu}}(t)\right)\right). \tag{136}$$

We notice that $\exists t_h$ such that $\forall t > t_h : h(t) < 0$.
Additionally, $g(t)$ has the following properties, as can be shown using basic analysis:

$$g(t+1) - g(t) = \Theta(t^{-1} \log^{\frac{1}{\nu}-1}(t)) \tag{137}$$

$$g^{1-\nu}(t) - g^{1-\nu}(t+1) = \log^{\frac{1}{\nu}-1}(t) - \log^{\frac{1}{\nu}-1}(t+1) = \Theta(t^{-1} \log^{\frac{1}{\nu}-2}(t)) \tag{138}$$

$$\frac{1}{\nu t} \log^{\frac{1}{\nu}-1}(t) \left[1 + h(t)\right] - (g(t+1) - g(t)) = o(t^{-2}) \tag{139}$$

We denote $\bar{C}_1 = \frac{1-\nu}{\nu}$. $\exists m_1(t) = o(\log^{\epsilon}(t)), m_2(t) = o(\log^{\epsilon}(t))$ such that $\forall \epsilon > 0$:

$$\left(g^{1-\nu}(t)\right)' = (1-\nu) g^{-\nu}(t) g'(t) = \bar{C}_1 \frac{1}{t} \log^{\frac{1}{\nu}-2}(t) + \frac{1}{t} \log^{\frac{1}{\nu}-3}(t) m_1(t) \tag{140}$$

$$\left(g^{1-\nu}(t)\right)'' = \frac{1}{t^2} \log^{\frac{1}{\nu}-2}(t) m_2(t) \tag{141}$$

In addition:

$$g^{\nu}(t)=\left(\log^{\frac{1}{\nu}}(t)+\frac{1}{\nu}\log(\nu\log^{1-\frac{1}{\nu}}(t))\log^{\frac{1}{\nu}-1}(t)\right)^{\nu}$$
$$=\left(\log^{\frac{1}{\nu}}(t)\left(1+\frac{1}{\nu}\log(\nu\log^{1-\frac{1}{\nu}}(t))\log^{-1}(t)\right)\right)^{\nu}\geq\log(t)+\log(\nu\log^{1-\frac{1}{\nu}}(t)),\ \forall t>t_B\,,$$

where in the last transition we used $\nu\geq 1$, $\exists t_B$ such that

$$\forall t>t_B:\ \frac{1}{\nu}\log(\nu\log^{1-\frac{1}{\nu}}(t))\log^{-1}(t)\geq 0$$

and Bernoulli's inequality:

$$\forall r\geq 1,x\geq -1:(1+x)^r\geq 1+rx \tag{142}$$

Therefore, $\forall a>0$

$$\exp(-ag^{\nu}(t))\leq\exp\left(-a\left(\log(t)+\log(\nu\log^{1-\frac{1}{\nu}}(t))\right)\right)=t^{-a}\left(\frac{1}{\nu}\log^{\frac{1}{\nu}-1}(t)\right)^{a}. \tag{143}$$

### K.2.2 Proof Of Theorem 11 for $\nu>1$

Our goal is to show that $\|\mathbf{r}(t)\|$ is bounded, and therefore $\boldsymbol{\rho}(t)=\mathbf{r}(t)+g^{1-\nu}(t)\tilde{\mathbf{w}}$ is bounded. To show this, we will upper bound the following equation

$$\|\mathbf{r}(t+1)\|^2=\|\mathbf{r}(t+1)-\mathbf{r}(t)\|^2+2\left(\mathbf{r}(t+1)-\mathbf{r}(t)\right)^{\top}\mathbf{r}(t)+\|\mathbf{r}(t)\|^2 \tag{144}$$

First, we note that the first term in this equation can be upper bounded by

$$\begin{aligned}
&||\mathbf{r}(t+1)-\mathbf{r}(t)||^2\\
&\overset{(1)}{=}||\mathbf{w}(t+1)-g(t+1)\hat{\mathbf{w}}-g^{1-\nu}(t+1)\tilde{\mathbf{w}}-\mathbf{w}(t)+g(t)\hat{\mathbf{w}}+g^{1-\nu}(t)\tilde{\mathbf{w}}||^2\\
&\overset{(2)}{=}||-\eta\nabla L(\mathbf{w}(t))-\hat{\mathbf{w}}(g(t+1)-g(t))-\tilde{\mathbf{w}}(g^{1-\nu}(t+1)-g^{1-\nu}(t))||^2\\
&=\eta^2||\nabla L(\mathbf{w}(t))||^2+||\hat{\mathbf{w}}||^2(g(t+1)-g(t))^2+||\tilde{\mathbf{w}}||^2(g^{1-\nu}(t+1)-g^{1-\nu}(t))^2\\
&+2\eta(g(t+1)-g(t))\hat{\mathbf{w}}^{\top}\nabla L(\mathbf{w}(t))+2\eta(g^{1-\nu}(t+1)-g^{1-\nu}(t))\tilde{\mathbf{w}}^{\top}\nabla L(\mathbf{w}(t))\\
&+2(g(t+1)-g(t))(g^{1-\nu}(t+1)-g^{1-\nu}(t))\hat{\mathbf{w}}^{\top}\tilde{\mathbf{w}}\\
&\overset{(3)}{\leq}\eta^2||\nabla L(\mathbf{w}(t))||^2+o(t^{-1}\log^{\frac{1}{\nu}-2}(t)),
\end{aligned} \tag{145}$$

where in (1) we used eq. (134), in (2) we used eq. (2) and in (3) we used eq. (137), eq. (138), Lemma 7, and also that

$$\hat{\mathbf{w}}^{\top}\nabla\mathcal{L}(\mathbf{w}(t))=\sum_{n=1}^{N}\ell'(\mathbf{w}(t)^{\top}\mathbf{x}_n)\hat{\mathbf{w}}^{\top}\mathbf{x}_n\leq 0 \tag{146}$$

since $\hat{\mathbf{w}}^{\top}\mathbf{x}_n\geq 1$ from the definition of $\hat{\mathbf{w}}$ and $\ell'(u)\leq 0$.
Also, from Lemma 7 we know that:

$$\|\nabla\mathcal{L}(\mathbf{w}(u))\|^2=o(1)\text{ and }\sum_{u=0}^{\infty}\|\nabla\mathcal{L}(\mathbf{w}(u))\|^2<\infty \tag{147}$$

Substituting eq. (147) into eq. (145), and recalling that the power series $t^{-1}\log^{-v}(t)$ converges for any $\nu>1$, we can find $C_0$ such that

$$||\mathbf{r}(t+1)-\mathbf{r}(t)||^2=o(1)\text{ and }\sum_{t=0}^{\infty}||\mathbf{r}(t+1)-\mathbf{r}(t)||^2=C_0<\infty \tag{148}$$

This equation also implies that

$$\forall\epsilon_0,\ \exists t_0\ :\ \forall t>t_0\ :\ |||\mathbf{r}(t+1)-\mathbf{r}(t)|||<\epsilon_0 \tag{149}$$

Next, we would like to bound the second term in eq. (144). To do so, will use the following Lemma, which we will prove in appendix K.2.3

**Lemma 8.** *We have*

$$\exists C_1, t_1 \ : \ \forall t > t_1 \ : \ \left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq C_1 t^{-1} \left(\log(t)\right)^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)} \tag{150}$$

*Additionally,* $\forall \epsilon_1 > 0,\ \exists C_2, t_2$*, such that* $\forall t > t_2$*, if*

$$\|\mathbf{P}_1 \mathbf{r}(t)\| \geq \epsilon_1 \tag{151}$$

*then the following improved bounds holds*

$$\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq -C_2 t^{-1} \log^{\frac{1}{\nu}-1}(t) < 0 \tag{152}$$

From eq. (150) in Lemma 8, we can find $C_1, t_1$ such that $\forall t > t_1$:

$$\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq C_1 t^{-1} \left(\log(t)\right)^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)} \tag{153}$$

Thus, by combining eqs. 153 and 148 into eq. (144), we find

$$\begin{aligned}
&\|\mathbf{r}(t)\|^2 - \|\mathbf{r}(t_1)\|^2 \\
&= \sum_{u=t_1}^{t-1} \left[\|\mathbf{r}(u+1)\|^2 - \|\mathbf{r}(u)\|^2\right] \\
&\leq C_0 + 2\sum_{u=t_1}^{t-1} C_1 t^{-1} \left(\log(t)\right)^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)}
\end{aligned}$$

which is bounded, since $\nu > 1$. Therefore, $\|\mathbf{r}(t)\|$ is bounded.

### K.2.3 Proof Of Lemma 8

Recall that we defined $\mathbf{r}(t) = \mathbf{w}(t) - g(t)\hat{\mathbf{w}} - g^{1-\nu}(t)\tilde{\mathbf{w}}$. $\hat{\mathbf{w}}$ and $\tilde{\mathbf{w}}$ were defined in section K.1.

**Lemma 8.** *We have*

$$\exists C_1, t_1 \ : \ \forall t > t_1 \ : \ \left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq C_1 t^{-1} \left(\log(t)\right)^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)} \tag{150}$$

*Additionally,* $\forall \epsilon_1 > 0,\ \exists C_2, t_2$*, such that* $\forall t > t_2$*, if*

$$\|\mathbf{P}_1 \mathbf{r}(t)\| \geq \epsilon_1 \tag{151}$$

*then the following improved bounds holds*

$$\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq -C_2 t^{-1} \log^{\frac{1}{\nu}-1}(t) < 0 \tag{152}$$

We examine the expression we wish to bound, recalling that $\mathbf{r}(t) = \mathbf{w}(t) - g(t)\hat{\mathbf{w}} - g^{1-\nu}(t)\tilde{\mathbf{w}}$:

$$
\begin{aligned}
&(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \\
&= (-\eta \nabla L(\mathbf{w}(t)) - \hat{\mathbf{w}}(g(t+1) - g(t)) - \tilde{\mathbf{w}}(g^{1-\nu}(t+1) - g^{1-\nu}(t)))^\top \mathbf{r}(t) \\
&= -\eta \sum_{n=1}^{N} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\mathbf{x}_n^\top \mathbf{r}(t) - \hat{\mathbf{w}}^\top \mathbf{r}(t)(g(t+1) - g(t)) - \tilde{\mathbf{w}}^\top \mathbf{r}(t)(g^{1-\nu}(t+1) - g^{1-\nu}(t)) \\
&\overset{(1)}{=} \hat{\mathbf{w}}^\top \mathbf{r}(t) \left[\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) - (g(t+1) - g(t))\right] - \eta \sum_{n \notin \mathcal{S}} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\mathbf{x}_n^\top \mathbf{r}(t) \\
&- \eta \sum_{n \in \mathcal{S}} \left[\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) \exp(-\nu \mathbf{x}_n^\top \tilde{\mathbf{w}}) + \ell'(\mathbf{w}^\top(t)\mathbf{x}_n)\right] \mathbf{x}_n^\top \mathbf{r}(t) \\
&- \tilde{\mathbf{w}}^\top \mathbf{r}(t)(g^{1-\nu}(t+1) - g^{1-\nu}(t) - \left(g^{1-\nu}(t)\right)') - \tilde{\mathbf{w}}^\top \mathbf{r}(t) \left(g^{1-\nu}(t)\right)' \\
&\overset{(2)}{\leq} \hat{\mathbf{w}}^\top \mathbf{r}(t) \left[\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) - (g(t+1) - g(t))\right] - \eta \sum_{n \notin \mathcal{S}} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\mathbf{x}_n^\top \mathbf{r}(t) \\
&- \eta \sum_{n \in \mathcal{S}} \left[\gamma_n t^{-1} \log^{\frac{1}{\nu}-2} h_3(t) + \frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) \exp(-\nu \mathbf{x}_n^\top \tilde{\mathbf{w}}) \right.\\
&\left. + \ell'(\mathbf{w}^\top(t)\mathbf{x}_n)\right] \mathbf{x}_n^\top \mathbf{r}(t) - \tilde{\mathbf{w}}^\top \mathbf{r}(t)(g^{1-\nu}(t+1) - g^{1-\nu}(t) - \left(g^{1-\nu}(t)\right)') \\
&\overset{(3)}{\leq} \hat{\mathbf{w}}^\top \mathbf{r}(t) \left[\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) - (g(t+1) - g(t))\right] - \eta \sum_{n \notin \mathcal{S}} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\mathbf{x}_n^\top \mathbf{r}(t) \\
&- \eta \sum_{n \in \mathcal{S}} \left[\gamma_n t^{-1} \log^{\frac{1}{\nu}-2} h_3(t) + \frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) \exp(-\nu \mathbf{x}_n^\top \tilde{\mathbf{w}}) \right.\\
&\left. + \ell'(\mathbf{w}^\top(t)\mathbf{x}_n)\right] \mathbf{x}_n^\top \mathbf{r}(t) + o(t^{-1} \log^{\frac{1}{\nu}-2}(t)),
\end{aligned} \tag{154}
$$

where in (1) we used eqs. 128, 129 to obtain $\hat{\mathbf{w}} = \sum_{n \in \mathcal{S}} \alpha_n \mathbf{x}_n = \sum_{n \in \mathcal{S}} \exp(-\tilde{\mathbf{w}}^\top \mathbf{x}_n)\mathbf{x}_n$. In (2) we defined $h_3(t) = o(\log^{\epsilon}(t)), \forall \epsilon > 0$ and used eq. (140) and the fact that from $\tilde{\mathbf{w}}$ definition (eq. (129)) we can find $\{\gamma_n\}_{n=1}^{d}$ such that:

$$
\tilde{\mathbf{w}} = -\eta \sum_{n \in \mathcal{S}} \gamma_n \mathbf{x}_n \tag{155}
$$

In (3) we used eq. (141) and the fact that $\tilde{\mathbf{w}}^\top \mathbf{r}(t) = o(t)$ since

$$
\begin{aligned}
\tilde{\mathbf{w}}^\top \mathbf{r}(t) &= \tilde{\mathbf{w}}^\top \left(\mathbf{w}(0) - \eta \sum_{u=0}^{t-1} \nabla \mathcal{L}(\mathbf{w}(u)) - g(t)\hat{\mathbf{w}} - g^{1-\nu}(t)\tilde{\mathbf{w}}\right) \\
&\leq -\eta t \min_{0 \leq u \leq t} \tilde{\mathbf{w}}^\top \nabla \mathcal{L}(\mathbf{w}(u)) + O(g(t)) = o(t),
\end{aligned}
$$

where in the last line we used that $\nabla \mathcal{L}(\mathbf{w}(t)) = o(1)$, from Lemma 7.
We examine the three terms in eq. (205). The first term can be upper bounded $\forall t > t_1'$ by

$$
\begin{aligned}
&\hat{\mathbf{w}}^\top \mathbf{r}(t) \left[\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) - (g(t+1) - g(t))\right] \\
&\overset{(1)}{\leq} \max\left[\hat{\mathbf{w}}^\top \mathbf{P}_1 \mathbf{r}(t), 0\right] C_1 t^{-2} \log^{\frac{1}{\nu}-1}(t) \\
&\overset{(2)}{\leq} \begin{cases} ||\hat{\mathbf{w}}||\epsilon_1 C_3 t^{-2} \log^{\frac{1}{\nu}-1}(t) & \text{, if } ||\mathbf{P}_1 \mathbf{r}(t)|| \leq \epsilon_1 \\ o(t^{-1} \log^{\frac{1}{\nu}-1}(t)) & \text{, if } ||\mathbf{P}_1 \mathbf{r}(t)|| > \epsilon_1. \end{cases}
\end{aligned} \tag{156}
$$

where in (1) we used the fact that

$$
\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t)\,(1 + h(t)) - (g(t+1) - g(t)) = \Theta(t^{-2} \log^{\frac{1}{\nu}-1}(t))
$$

and therefore $\exists t_1', C_3$ such that

$$\forall t > t_1' : \frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \left(1 + h(t)\right) - \left(g(t+1) - g(t)\right) \leq C_3 t^{-2} \log^{\frac{1}{\nu}-1}(t).$$

In (2) we used $\hat{\mathbf{w}}^\top \mathbf{r}(t) = o(t)$.
Next, we want to upper bound the second term in eq. (154). If ${\mathbf{x}_n}^\top \mathbf{r}(t) \geq 0$ then we can show that $\exists t_c$ so that $\forall t > t_c$

$$(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu \geq (\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu g^\nu(t) + \nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{\mathbf{w}}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{r}(t). \tag{157}$$

*Proof.*

$$\begin{aligned}
&(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu \overset{(1)}{=} -\left(g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{r}(t)\right)^\nu \\
&= (g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu \left(1 + \frac{g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{r}(t)}{g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n}\right)^\nu \\
&\overset{(2)}{\geq} (g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu \left(1 + \nu \frac{g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{r}(t)}{g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n}\right) \\
&= (\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu g^\nu(t) + \nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{\mathbf{w}}^\top \mathbf{x}_n + \nu g^{\nu-1}(t) (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(3)}{\geq} (\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu g^\nu(t) + \nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{\mathbf{w}}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{r}(t), \forall t > t_c
\end{aligned}$$

where in (1) we used $\mathbf{r}(t)$ definition (eq. (134)), in (2) we used $\nu \geq 1$, Bernoulli's inequality (eq. (142)) and the fact that $\exists t'$ such that $\forall t > t'$:

$$g^{-\nu}(t) \frac{\tilde{\mathbf{w}}^\top \mathbf{x}_n}{\hat{\mathbf{w}}^\top \mathbf{x}_n} \geq -1$$

and therefore:

$$\frac{g^{-(\nu-1)}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_n + \mathbf{r}(t)^\top \mathbf{x}_n}{g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n} = \frac{g^{-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_n}{\hat{\mathbf{w}}^\top \mathbf{x}_n} + \frac{\mathbf{r}(t)^\top \mathbf{x}_n}{g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n} \geq -1.$$

In (3) we used the fact that $\exists t''$ such that $\forall t > t'' : \nu g^{\nu-1}(t) (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \geq 1$ (since $\nu \geq 1$). We define $t_c = \max(t', t_B, t'')$. ☐

Using this result, we upper bound the second term in eq. (154), $\forall t > t_3$:

$$\begin{aligned}
&-\eta \sum_{n \notin \mathcal{S}} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(1)}{\leq} \eta \sum_{\substack{n \notin \mathcal{S} \\ \mathbf{x}_n^\top \mathbf{r}(t) \geq 0}} \left(1 + \exp(-\mu_+ (\mathbf{w}(t)^\top \mathbf{x}_n)^\nu)\right) \exp(-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(2)}{\leq} \eta \sum_{\substack{n \notin \mathcal{S} \\ \mathbf{x}_n^\top \mathbf{r}(t) \geq 0}} 2 \exp(-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(3)}{\leq} 2\eta \sum_{\substack{n \notin \mathcal{S} \\ \mathbf{x}_n^\top \mathbf{r}(t) \geq 0}} \exp(-(\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu g^\nu(t)) \exp\left(-\nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \exp(-\mathbf{x}_n^\top \mathbf{r}(t)) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(4)}{\leq} 2\eta \sum_{\substack{n \notin \mathcal{S} \\ \mathbf{x}_n^\top \mathbf{r}(t) \geq 0}} \nu^{-\theta^\nu} t^{-\theta^\nu} (\log(t))^{(\frac{1}{\nu}-1)\theta^\nu} \exp\left(-\nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \exp(-\mathbf{x}_n^\top \mathbf{r}(t)) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(5)}{\leq} 2\eta N \nu^{-\theta^\nu} t^{-\theta^\nu} (\log(t))^{(\frac{1}{\nu}-1)\theta^\nu} \exp\left(-\nu \min_{n \notin \mathcal{S}} \left((\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\right) \\
&\leq \tilde{C}_2 t^{-\theta^\nu} (\log(t))^{(\frac{1}{\nu}-1)\theta^\nu}, \ \forall t > t_3,
\end{aligned} \tag{158}$$

where in (1) we used eq. (132), in (2) we used the fact that $\mathbf{w}^\top \mathbf{x}_n \to \infty$ (from Lemma 6) and thus $\exists t_L$ so that $\forall t > t_L$ : $(1 + \exp(-\mu_+(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu)) \leq 2$, in (3) we used eq. (157), in (4) we used $\theta = \min_{n \notin \mathcal{S}} \hat{\mathbf{w}}^\top \mathbf{x}_n > 1$ and eq. (143) and in (5) we used $\mathbf{x}_n^\top \mathbf{r} \geq 0$ and also that $x \exp(-x) \leq 1$. We define $t_3 = \max(\bar{t}, t_L, t_c)$ and

$$\tilde{C}_2 = 2\eta N \nu^{-\theta^\nu} \exp\left(-\nu \min_{n \notin \mathcal{S}} \left((\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\right)$$

Lastly, we will aim to bound the third term in eq. (154)

$$\eta \sum_{n \in \mathcal{S}} \left[-\gamma_n t^{-1} \log^{\frac{1}{\nu}-2} h_3(t) - \frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \left(1 + h(t)\right) \exp(-\nu \mathbf{x}_n^\top \tilde{\mathbf{w}}) - \ell'(\mathbf{w}^\top(t) \mathbf{x}_n)\right] \mathbf{x}_n^\top \mathbf{r}(t) \tag{159}$$

We examine each term $k$ in this sum, and divide into two cases, depending on the sign of $\mathbf{x}_k^\top \mathbf{r}(t)$.
First, if $\mathbf{x}_k^\top \mathbf{r}(t) \geq 0$ then term $k$ in eq. (159) can be upper bounded $\forall t > \bar{t}$ using eq. (132) by

$$\begin{aligned}
&\left[ \left(1 + \left[\nu^{-1} t^{-1} (\log(t))^{\left(\frac{1}{\nu}-1\right)}\right]^{\mu_+} \exp\left(-\mu_+ \nu \tilde{\mathbf{w}}^\top \mathbf{x}_k\right)\right) \exp\left(-\nu g^{\nu-1}(t) \mathbf{x}_k^\top \mathbf{r}(t)\right) \right.\\
&\left. - (1 + h(t) + f(t))\right] \left(\eta \frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \exp(-\nu \mathbf{x}_k^\top \tilde{\mathbf{w}})\right) \mathbf{x}_k^\top r(t),
\end{aligned} \tag{160}$$

where we defined $f(t) = \gamma_k \nu \exp(\nu \mathbf{x}_k^\top \tilde{\mathbf{w}}) \log^{-1}(t) h_3(t)$.
We further divide into cases:
1. If $|\mathbf{x}_k^\top r| \leq C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}(t)$ then we can upper bound eq. (160) with

$$2\eta C_0 \frac{1}{\nu} t^{-1} \log^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)}(t) \exp(-\nu \mathbf{x}_k^\top \tilde{\mathbf{w}}) \tag{161}$$

2. If $|\mathbf{x}_k^\top r| > C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}(t)$ then we can upper bound eq. (160) with zero since:

$$\begin{aligned}
&\left[ \left(1 + \left[\nu^{-1} t^{-1} (\log(t))^{\left(\frac{1}{\nu}-1\right)}\right]^{\mu_+} \exp\left(-\mu_+ \nu \tilde{\mathbf{w}}^\top \mathbf{x}_k\right)\right) \exp\left(-C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}(t)\right) - (1 + h(t) + f(t))\right]\\
&\overset{(1)}{\leq} \left[ \left(1 + \left[\nu^{-1} t^{-1} (\log(t))^{\left(\frac{1}{\nu}-1\right)}\right]^{\mu_+} \exp\left(-\mu_+ \nu \tilde{\mathbf{w}}^\top \mathbf{x}_k\right)\right) \left(1 - C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}(t) + C_0^2 \log^{-1-\frac{1}{\nu}}(t)\right)\right.\\
&\left. - (1 + h(t) + f(t))\right]\\
&\leq \left[\left(1 - C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}(t) + C_0^2 \log^{-1-\frac{1}{\nu}}(t)\right) \left[\nu^{-1} t^{-1} (\log(t))^{\left(\frac{1}{\nu}-1\right)}\right]^{\mu_+} \exp\left(-\mu_+ \nu \tilde{\mathbf{w}}^\top \mathbf{x}_k\right)\right.\\
&\left. - C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}(t) + C_0^2 \log^{-1-\frac{1}{\nu}}(t) - h(t) - f(t)\right] \overset{(2)}{\leq} 0,\ \forall t > t_4,
\end{aligned} \tag{162}$$

where in (1) we used the fact that $e^{-x} \leq 1 - x + x^2$ for $x > 0$ and in (2) we defined $t_4$ so that $\forall t > t_4 > \bar{t}$ the previous expression is negative - this is possible because $\log^{-\frac{1}{2}-\frac{1}{2\nu}}(t)$ decreases slower than $h(t)$ and $f(t)$ ($\forall \nu > 1 :\ -\frac{1}{2} - \frac{1}{2\nu} > -1$).
3. If, in addition, $|\mathbf{x}_k^\top \mathbf{r}| \geq \epsilon_2$, then we can find $C_4$ so that we can upper bound eq. (160) with

$$-C_4 \eta \frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \exp(-\nu \max_n (\mathbf{x}_n^\top \tilde{\mathbf{w}})) \epsilon_2,\ \forall t > \bar{t} \tag{163}$$

Second, if $\mathbf{x}_k^\top \mathbf{r}(t) < 0$, we again further divide into cases:
1. If $|\mathbf{x}_k^\top \mathbf{r}| \leq C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}$, then, since $-\ell'(\mathbf{w}(t)^\top \mathbf{x}_k) > 0$, we can upper bound term $k$ in equation 159 by

$$C_0 \eta \frac{1}{\nu} t^{-1} \log^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)}(t) \exp(-\nu \min_n (\mathbf{x}_n^\top \tilde{\mathbf{w}}))$$

2. If $|\mathbf{x}_k^\top \mathbf{r}| \geq C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}}$, then, using eq. (133) we can upper bound term $k$ in equation 159, $\forall t > t_5$ by

$$\eta \left[\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \exp(-\nu \mathbf{x}_k^\top \tilde{\mathbf{w}}) - \left(1 - \exp\left(-\mu_- \left(\mathbf{w}(t)^\top \mathbf{x}_k\right)^\nu\right)\right) \exp\left(-\left(\mathbf{w}(t)^\top \mathbf{x}_k\right)^\nu\right)\right] |\mathbf{x}_k^\top \mathbf{r}(t)| \tag{164}$$

We used the fact that $f(t) = o(h(t))$ and therefore $\exists t_5 > \bar{t}$ such that $\forall t > t_5 :\ h(t) + f(t) \leq 0$.
We can use Taylor's theorem to show that:

$$\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_n\right)^\nu = \log(t) + \log(\nu \log^{1-\frac{1}{\nu}}(t)) + \nu\tilde{\mathbf{w}}^\top \mathbf{x}_k + f_2(t), \tag{165}$$

where $|f_2(t)| = o(\log^{-\frac{1}{\nu}}(t))$.
From Lemma 6 we know that $\mathbf{w}(t)^\top \mathbf{x}_n \to \infty$ and therefore $\exists t_6 > t_5$ so that $\forall t > t_6$:

$$\left(1 - \exp\left(-\mu_- \left(\mathbf{w}(t)^\top \mathbf{x}_k\right)^\nu\right)\right) \geq 0$$

$$\mathbf{w}(t)^\top \mathbf{x}_n = g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_n + \mathbf{r}(t)^\top \mathbf{x}_n > 0$$

Using the last equations and the fact that $\forall a > 1,\ x \in [-1, 0] :\ (1 + x)^a \leq (1 + x)$ and $-1 \leq \frac{\mathbf{r}(t)^\top \mathbf{x}_k}{g(t)+g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k} \leq 0$, we can show that $\forall t > t_6$:

$$\begin{aligned}
&\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k + \mathbf{r}(t)^\top \mathbf{x}_k\right)^\nu \\
&\leq \left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^\nu + \left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} \mathbf{r}(t)^\top \mathbf{x}_k \\
&\leq \log(t) + \log(\nu \log^{1-\frac{1}{\nu}}(t)) + \nu\tilde{\mathbf{w}}^\top \mathbf{x}_k + f_2(t) + \left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} \mathbf{r}(t)^\top \mathbf{x}_k \\
&\leq \log(t) + \log(\nu \log^{1-\frac{1}{\nu}}(t)) + \nu\tilde{\mathbf{w}}^\top \mathbf{x}_k + \log^{-\frac{1}{\nu}}(t) + \left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} \mathbf{r}(t)^\top \mathbf{x}_k
\end{aligned} \tag{166}$$

Using eq. (166), eq. (164) can be upper bounded by

$$\begin{aligned}
&\eta\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})\Big[1 - \left(1 - e^{-\mu_-\left(\mathbf{w}(t)^\top \mathbf{x}_k\right)^\nu}\right) \\
&\exp\left(-\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} \mathbf{r}(t)^\top \mathbf{x}_k - \log^{-\frac{1}{\nu}}(t)\right)\Big]|\mathbf{x}_k^\top \mathbf{r}(t)|
\end{aligned}$$

Next, we will show that $\exists t'' > t_6$ such that the last expression is strictly negative $\forall t > t''$. Let $M > 1$ be some arbitrary constant. Then, since $\exp\left(-\mu_- \left(\mathbf{w}(t)^\top \mathbf{x}_k\right)^\nu\right) \to 0$ from Lemma 6, $\exists t_M > \bar{t}$ such that $\forall t > t_M$ and if $\exp(-\mathbf{r}(t)^\top \mathbf{x}_k) \geq M > 1$ then

$$\left(1 - \exp\left(-\mu_- \left(\mathbf{w}(t)^\top \mathbf{x}_k\right)^\nu\right)\right)\exp\left(\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} |\mathbf{r}(t)^\top \mathbf{x}_k| - \log^{-\frac{1}{\nu}}(t)\right) \geq M' > 1 \tag{167}$$

Furthermore, if $\exists t > t_M$ such that $\exp(-\mathbf{r}(t)^\top \mathbf{x}_k) < M$, then

$$\begin{aligned}
&\left(1 - \exp\left(-\mu_- \left(\mathbf{w}(t)^\top \mathbf{x}_k\right)^\nu\right)\right)\exp\left(\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} |\mathbf{r}(t)^\top \mathbf{x}_k| - \log^{-\frac{1}{\nu}}(t)\right) \\
&\overset{(1)}{\geq} \left(1 - \left[t^{-1}\nu^{-1}\log^{\frac{1}{\nu}-1}(t)e^{-\nu\tilde{\mathbf{w}}^\top \mathbf{x}_k}M\right]^{\mu_-}\right)\exp\left(\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} |\mathbf{r}(t)^\top \mathbf{x}_k| - \log^{-\frac{1}{\nu}}(t)\right) \\
&\overset{(2)}{\geq} \left(1 - \left[t^{-1}\nu^{-1}\log^{\frac{1}{\nu}-1}(t)e^{-\nu\tilde{\mathbf{w}}^\top \mathbf{x}_k}M\right]^{\mu_-}\right)\exp\left(\log^{-1+\frac{5}{4}\left(1-\frac{1}{\nu}\right)}(t) - \log^{-\frac{1}{\nu}}(t)\right) \\
&\overset{(3)}{\geq} \left(1 - \left[t^{-1}\nu^{-1}\log^{\frac{1}{\nu}-1}(t)e^{-\nu\tilde{\mathbf{w}}^\top \mathbf{x}_k}M\right]^{\mu_-}\right)\left(1 + \log^{-1+\frac{5}{4}\left(1-\frac{1}{\nu}\right)}(t) - \log^{-\frac{1}{\nu}}(t)\right) \\
&\geq 1 + \log^{-1+\frac{5}{4}\left(1-\frac{1}{\nu}\right)}(t) - \log^{-\frac{1}{\nu}}(t) \\
&- \left[t^{-1}\nu^{-1}\log^{\frac{1}{\nu}-1}(t)e^{-\nu\tilde{\mathbf{w}}^\top \mathbf{x}_k}M\right]^{\mu_-}\left(1 + \log^{-1+\frac{5}{4}\left(1-\frac{1}{\nu}\right)}(t) - \log^{-\frac{1}{\nu}}(t)\right) \\
&> 1,
\end{aligned} \tag{168}$$

where in transition (1) we used eq. (142) and in transition (2) we used:

$$\begin{aligned}
&\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} |\mathbf{r}(t)^\top \mathbf{x}_k| - \log^{-\frac{1}{\nu}}(t) \geq \left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} C_0 \log^{-\frac{1}{2}-\frac{1}{2\nu}} - \log^{-\frac{1}{\nu}}(t) \\
&\geq \log^{-1+\frac{5}{4}\left(1-\frac{1}{\nu}\right)}(t) - \log^{-\frac{1}{\nu}}(t)
\end{aligned}$$

and in (3) we used $e^x \geq 1 + x$. eq. (168) is greater than 1 since $\log^{-1+\frac{5}{4}\left(1-\frac{1}{\nu}\right)}(t)$ decrease slower than the other terms. Therefore, after we substitute eqs. 167 and 168 into eq. (164), we find that $\exists t'_{-} > t''$ such that $\forall t > t'_{-}$ term $k$ in equation 213 is strictly negative.
3. If $|\mathbf{x}_k^\top \mathbf{r}| \geq \epsilon_2$, then $\exists t_7, \epsilon_3$ such that $\forall t > t_7$

$$\left(g(t) + g^{1-\nu}(t)\tilde{\mathbf{w}}^\top \mathbf{x}_k\right)^{\nu-1} |\mathbf{r}(t)^\top \mathbf{x}_k| - \log^{-\frac{1}{\nu}}(t) > \epsilon_3 \tag{169}$$

and we can find $C_5$ such that we can upper bound term $k$ in equation 159 $\forall t > t_7$ by

$$-\eta C_6 \frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \exp(-\nu \max_n (\mathbf{x}_n^\top \tilde{\mathbf{w}}))\epsilon_2 \tag{170}$$

To conclude, we choose $t_0 = \max\left[t'_1, t_3, t_4, t_6, t'_{-}, t_7\right]$:

1. If $\|\mathbf{P}_1 \mathbf{r}\,(t)\| \geq \epsilon_1$ (as in eq. (156)), we have that

$$\max_{n\in\mathcal{S}} \left|\mathbf{x}_n^\top \mathbf{r}\,(t)\right|^2 \overset{(1)}{\geq} \frac{1}{|\mathcal{S}|} \sum_{n\in\mathcal{S}} \left|\mathbf{x}_n^\top \mathbf{P}_1 \mathbf{r}\,(t)\right|^2 = \frac{1}{|\mathcal{S}|} \left\|\mathbf{X}_\mathcal{S}^\top \mathbf{P}_1 \mathbf{r}\,(t)\right\|^2 \overset{(2)}{\geq} \frac{1}{|\mathcal{S}|} \sigma^2_{\min}\left(\mathbf{X}_\mathcal{S}\right)\epsilon_1^2, \tag{171}$$

   where in (1) we used $\mathbf{P}_1^\top \mathbf{x}_n = \mathbf{x}_n \; \forall n \in \mathcal{S}$, in (2) we denoted $\mathbf{X}_\mathcal{S} \in \mathbb{R}^{d\times|\mathcal{S}|}$ as the matrix whose columns are the support vectors and by $\sigma_{\min}\left(\mathbf{X}_\mathcal{S}\right)$, the minimal non-zero singular value of $\mathbf{X}_\mathcal{S}$ and used eq. 151. Therefore, for some $k$, $\left|\mathbf{x}_k^\top \mathbf{r}\right| \geq \epsilon_2 \triangleq |\mathcal{S}|^{-1} \sigma^2_{\min}\left(\mathbf{X}_\mathcal{S}\right)\epsilon_1^2$. In this case, we denote $C''_0$ as the minimum between $C_6 \eta \nu^{-1} \exp\left(-\nu \max_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\epsilon_2$ (eq. (170)) and $C_4 \eta \nu^{-1} \exp\left(-\nu \max_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\epsilon_2$ (eq. (163)). Then we find that eq. (159) can be upper bounded by $-C''_0 t^{-1} \log^{\frac{1}{\nu}-1}(t) + o\left(t^{-1} \log^{\frac{1}{\nu}-1}(t)\right)$, $\forall t > t_0$, given eq. (151). Substituting this result, together with eqs. 156 and 158 into eq. (154), we obtain $\forall t > t_0$

$$\left(\mathbf{r}\,(t+1) - \mathbf{r}\,(t)\right)^\top \mathbf{r}\,(t) \leq -C''_0 t^{-1} \log^{\frac{1}{\nu}-1}(t) + o\left(t^{-1} \log^{\frac{1}{\nu}-1}(t)\right) .$$

   This implies that $\exists C_2 < C''_0$ and $\exists t_2 > t_0$ such that eq. (152) holds. This implies also that eq. (150) holds for $\|\mathbf{P}_1 \mathbf{r}\,(t)\| \geq \epsilon_1$.

2. Otherwise, if $\|\mathbf{P}_1 \mathbf{r}\,(t)\| < \epsilon_1$, we find that $\forall t > t_0$ , each term in eq. (159) can be upper bounded by either zero, or terms proportional to $t^{-1} \log^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)}$. Combining this together with eqs. 156, 158 into eq. (154) we obtain (for some positive constants $C_7$, $C_8$, $C_9$, and $C_6$)

$$\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq C_7 t^{-2} \log^{\frac{1}{\nu}-1}(t) + C_8 t^{-\theta^\nu} (\log(t))^{(\frac{1}{\nu}-1)\theta^\nu} + C_9 t^{-1} \log^{-1-\frac{1}{2}\left(1-\frac{1}{\nu}\right)}(t)$$

   Therefore, $\exists t_1 > t_0$ and $C_1$ such that eq. (150) holds.

### K.3 $\frac{1}{4} < \nu < 1$

In the following proofs, for any solution $\mathbf{w}(t)$, we define

$$\mathbf{r}(t) = \mathbf{w}(t) - g(t)\hat{\mathbf{w}} - g^{1-\nu}(t)\tilde{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t)\hat{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t)\log^{-1}(t)\bar{\mathbf{w}} + \frac{1}{\nu}g^{1-2\nu}(t)\bar{\mathbf{w}}_2, \tag{172}$$

where $\tilde{h}(t)$ is defined below (subsection K.3.1), $\hat{\mathbf{w}}$ follows the conditions of Theorem 11, that is $\hat{\mathbf{w}}$ is the $L_2$ max margin vector, which satisfies eq. (4):

$$\hat{\mathbf{w}} = \underset{\mathbf{w}(t)\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}(t)\|^2 \text{ s.t. } \mathbf{w}(t)^\top \mathbf{x}_n \geq 1$$

and $\tilde{\mathbf{w}}$ is a vector which satisfies the equations:

$$\forall n \in \mathcal{S} \ : \ \eta \exp(-\nu \mathbf{x}_n^\top \tilde{\mathbf{w}}) = \alpha_n, \ \bar{\mathbf{P}}_1 \tilde{\mathbf{w}} = 0 \tag{173}$$

From $\tilde{\mathbf{w}}$ definition, we know that $\exists \{\lambda_n\}_{n=1}^d$ such that

$$\forall n \in \mathcal{S} \ : \ \tilde{\mathbf{w}} = \sum_{n\in\mathcal{S}} \lambda_n \mathbf{x}_n \tag{174}$$

Using the last equation, we define $\bar{\mathbf{w}}$, a vector which satisfies the equations:

$$\forall n \in \mathcal{S} \ : \ \bar{\mathbf{w}}^\top \mathbf{x}_n = \frac{\nu \bar{C}_1 \lambda_n}{\eta} \exp(\nu \mathbf{x}_n^\top \tilde{\mathbf{w}}) \triangleq \tilde{\gamma}_n, \ \bar{\mathbf{P}}_1 \bar{\mathbf{w}} = 0, \tag{175}$$

where $\bar{C}_1 = \frac{1-\nu}{\nu}$.
$\bar{\mathbf{w}}_2$ is a vector which satisfies the equations:

$$\forall n \in \mathcal{S} \ : \ \bar{\mathbf{w}}_2^\top \mathbf{x}_n = \frac{\nu(\nu-1)}{2}(\tilde{\mathbf{w}}^\top \mathbf{x}_n)^2, \ \bar{\mathbf{P}}_1 \bar{\mathbf{w}}_2 = 0 \tag{176}$$

These equations have a unique solution for almost every dataset from Lemma 8 in Soudry et al. (2018a).

#### K.3.1 Auxiliary results

We will denote the functions:

$$\begin{aligned}
g(t) &= \log^{\frac{1}{\nu}}(t) + \frac{1}{\nu}\log(\nu \log^{1-\frac{1}{\nu}}(t)) \log^{\frac{1}{\nu}-1}(t) \\
h(t) &= \left(1 - \frac{1}{\nu}\right) \log^{-1}(t) \left(1 - \log\left(\nu \log^{1-\frac{1}{\nu}}(t)\right)\right) \\
h_2(t) &= \frac{1}{t}\left[\left(-1 + \left(\frac{1}{\nu} - 1\right)\log^{-1}(t)\right)(1 + h(t)) + \left(1 - \frac{1}{\nu}\right)\log^{-2}(t)\left(\log\left(\nu \log^{1-\frac{1}{\nu}}(t)\right) - \frac{1}{\nu}\right)\right] \\
\tilde{h}(t) &= h(t) + \frac{(\nu-1)}{2\nu}\log^{-1}(t)\log^2(\nu \log^{1-\frac{1}{\nu}}(t))
\end{aligned} \tag{177}$$

Using basic analysis, it is straightforward to show that these functions have the following properties:

$$g(t+1) - g(t) = \Theta(t^{-1}\log^{\frac{1}{\nu}-1}(t)) \tag{178}$$

$$g^{1-\nu}(t+1) - g^{1-\nu}(t) = \log^{\frac{1}{\nu}-1}(t+1) - \log^{\frac{1}{\nu}-1}(t) = \Theta(t^{-1}\log^{\frac{1}{\nu}-2}(t)) \tag{179}$$

$$\left|\frac{1}{\nu t}\log^{\frac{1}{\nu}-1}(t)\left[1 + h(t) + \frac{1}{2}h_2(t)\right] - (g(t+1) - g(t))\right| = \Theta(t^{-3}\log^{\frac{1}{\nu}-1}(t)) \tag{180}$$

$$\left|g^{1-\nu}(t)h(t) - g^{1-\nu}(t+1)h(t+1)\right| = o(t^{-1}\log^{\frac{1}{\nu}-2}) \tag{181}$$

We denote $\bar{C}_1 = \frac{1-\nu}{\nu}$. $\exists m_1(t) = o(\log^{\epsilon}(t)), m_2(t) = o(\log^{\epsilon}(t))$ such that $\forall \epsilon > 0$:

$$\left(g^{1-\nu}(t)\right)' = (1-\nu)g^{-\nu}(t)g'(t) = \bar{C}_1 \frac{1}{t}\log^{\frac{1}{\nu}-2}(t) + \frac{1}{t}\log^{\frac{1}{\nu}-3}(t)m_1(t) \tag{182}$$

$$\left(g^{1-\nu}(t)\right)'' = \frac{1}{t^2}\log^{\frac{1}{\nu}-2} m_2(t) \tag{183}$$

$$\left(g^{1-\nu}(t)\right)''' = o(t^{-2-\epsilon}) \tag{184}$$

There exists $m_3(t) = o(\log^{\epsilon}(t)), m_4(t) = o(\log^{\epsilon}(t))$ such that $\forall\epsilon > 0$:

$$(g^{1-\nu}(t)\tilde{h}(t))' = \frac{1}{t}\log^{\frac{1}{\nu}-3}(t)m_3(t) \tag{185}$$

$$(g^{1-\nu}(t)\tilde{h}(t))'' = \frac{1}{t^2}\log^{\frac{1}{\nu}-3}(t)m_4(t) \tag{186}$$

$$(g^{1-\nu}(t)\tilde{h}(t))''' = o(t^{-2-\epsilon}) \tag{187}$$

There exists $m_5(t) = o(\log^{\epsilon}(t)), m_6(t) = o(\log^{\epsilon}(t))$ such that $\forall\epsilon > 0$:

$$(g^{1-\nu}(t)\log^{-1}(t))' = \frac{1}{t}\log^{\frac{1}{\nu}-3}(t)m_5(t) \tag{188}$$

$$(g^{1-\nu}(t)\log^{-1}(t))'' = \frac{1}{t^2}\log^{\frac{1}{\nu}-3}(t)m_6(t) \tag{189}$$

$$(g^{1-\nu}(t)\log^{-1}(t))''' = o(t^{-2-\epsilon}) \tag{190}$$

There exists $\tilde{C}_1$ and $m_7(t) = o(\log^{\epsilon}(t)), m_8(t) = o(\log^{\epsilon}(t))$ such that $\forall\epsilon > 0$:

$$\left(g^{1-2\nu}(t)\right)' = \frac{1}{t}\log^{\frac{1}{\nu}-3}(t)m_7(t) \tag{191}$$

$$\left(g^{1-2\nu}(t)\right)'' = \frac{1}{t^2}\log^{\frac{1}{\nu}-3} m_8(t) \tag{192}$$

$$\left(g^{1-2\nu}(t)\right)''' = o(t^{-2-\epsilon}) \tag{193}$$

Combining these properties, for arbitrary constants $\alpha_1, ..., \alpha_4$ we get:

$$\begin{aligned}
&\alpha_1\left(\left(g^{1-\nu}(t)\right)' + \frac{1}{2}\left(g^{1-\nu}(t)\right)''\right) + \alpha_2\left((g^{1-\nu}(t)\tilde{h}(t))' + \frac{1}{2}(g^{1-\nu}(t)\tilde{h}(t))''\right)\\
&+\alpha_3\left((g^{1-\nu}(t)\log^{-1}(t))' + \frac{1}{2}g^{1-\nu}(t)\log^{-1}(t))''\right) + \alpha_4\left(\left(g^{1-2\nu}(t)\right)' + \frac{1}{2}\left(g^{1-2\nu}(t)\right)''\right)\\
&= \alpha_1\bar{C}_1\frac{1}{t}\log^{\frac{1}{\nu}-2}(t) + \frac{1}{t}\log^{\frac{1}{\nu}-3}\tilde{m}_1(t) + \frac{1}{t^2}\log^{\frac{1}{\nu}-2}\tilde{m}_2(t),
\end{aligned} \tag{194}$$

where $\forall\epsilon > 0 : \tilde{m}_1(t) = o(\log^{\epsilon}(t)), \tilde{m}_2(t) = o(\log^{\epsilon}(t))$.
We can use Taylor's theorem to show that:

$$\begin{aligned}
g^{\nu}(t) &= \log(t) + \log(\nu\log^{1-\frac{1}{\nu}}(t)) + \frac{(\nu-1)}{2\nu}\log^{-1}(t)\log^2(\nu\log^{1-\frac{1}{\nu}}(t))\\
&+\frac{(\nu-1)(\nu-2)}{6\nu^2}\log^{-2}(t)\log^3\left(\nu\log^{1-\frac{1}{\nu}}(t)\right) + o\left(\log^{-2}(t)\log^3\left(\nu\log^{1-\frac{1}{\nu}}(t)\right)\right)
\end{aligned}$$

and also that

$$\begin{aligned}
&\left(g(t)\hat{\mathbf{w}}^\top\mathbf{x}_n + g^{1-\nu}(t)\underbrace{\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n - \frac{1}{\nu}\tilde{h}(t)\hat{\mathbf{w}}^\top\mathbf{x}_n - \frac{1}{\nu}\log^{-1}(t)\bar{\mathbf{w}}^\top\mathbf{x}_n - \frac{1}{\nu}g^{-\nu}(t)\bar{\mathbf{w}}_2^\top\mathbf{x}_n\right)}_{\triangleq\,\tilde{f}(t)}\right)^{\nu}\\
&= g^{\nu}(t)(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu} + \nu(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-1}\tilde{f}(t) + \frac{\nu(\nu-1)}{2}g^{-\nu}(t)\left(\tilde{f}(t)\right)^2(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-2}\\
&+\frac{\nu(\nu-1)(\nu-2)}{6}g^{-2\nu}(t)\left(\tilde{f}(t)\right)^3(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-3} + o(g^{-2\nu}(t))\\
&= (\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu}\log(t) + (\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu}\log(\nu\log^{1-\frac{1}{\nu}}(t)) + \nu(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-1}\tilde{f}(t) + \tilde{f}_2(t) + f_2(t),
\end{aligned} \tag{195}$$

where we defined

$$\tilde{f}(t) \triangleq \tilde{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu}\tilde{h}(t)\hat{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu}\log^{-1}(t)\bar{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu}g^{-\nu}(t)\bar{\mathbf{w}}_2^\top \mathbf{x}_n,$$
$$\tilde{f}_2(t) \triangleq \frac{\nu(\nu-1)}{2}g^{-\nu}(t)\left(\tilde{\mathbf{w}}^\top \mathbf{x}_n\right)^2 \left(\hat{\mathbf{w}}^\top \mathbf{x}_n\right)^{\nu-2} + \frac{(\nu-1)}{2\nu}\left(\hat{\mathbf{w}}^\top \mathbf{x}_n\right)^{\nu}\log^{-1}(t)\log^2(\nu\log^{1-\frac{1}{\nu}}(t)).$$

We note that $|f_2(t)| = o(\log^{-2+\epsilon}(t)),\ \forall\epsilon > 0$ .

### K.3.2 Proof Of main Theorem ($\frac{1}{4} < \nu < 1$)

Our goal is to show that $\|\mathbf{r}(t)\| = o(g^{1-\nu}(t))$ , and therefore
$\boldsymbol{\rho}(t) = \mathbf{r}(t) + g^{1-\nu}(t)\tilde{\mathbf{w}} + o(g^{1-\nu}(t)) = O(g^{1-\nu}(t))$. To show this, we will upper bound the following equation

$$\|\mathbf{r}(t+1)\|^2 = \|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 + 2\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) + \|\mathbf{r}(t)\|^2 \tag{196}$$

First, we note that the first term in eq. (196) can be upper bounded by

$$\begin{aligned}
&\|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 \\
&\overset{(1)}{=} \Bigg\|\mathbf{w}(t+1) - \left(g(t+1) - \frac{1}{\nu}g^{1-\nu}(t+1)\tilde{h}(t+1)\right)\hat{\mathbf{w}} - g^{1-\nu}(t+1)\tilde{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t+1)\log^{-1}(t+1)\bar{\mathbf{w}} \\
&+ \frac{1}{\nu}g^{1-2\nu}(t+1)\bar{\mathbf{w}}_2 - \mathbf{w}(t) + \left(g(t) - \frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t)\right)\hat{\mathbf{w}} + g^{1-\nu}(t)\tilde{\mathbf{w}} - \frac{1}{\nu}g^{1-\nu}(t)\log^{-1}(t)\bar{\mathbf{w}} - \frac{1}{\nu}g^{1-2\nu}(t)\bar{\mathbf{w}}_2\Bigg\|^2 \\
&\overset{(2)}{=} \|-\eta\nabla L(\mathbf{w}(t)) - \hat{\mathbf{w}}\left(g(t+1) - g(t)\right) + \boldsymbol{\psi}_1(t)\|^2 \\
&\overset{(3)}{=} \eta^2\|\nabla L(\mathbf{w}(t))\|^2 + 2\eta\left(g(t+1) - g(t)\right)\hat{\mathbf{w}}^\top\nabla L(\mathbf{w}(t)) + o\left(t^{-1}\log^{\frac{1}{\nu}-2}(t)\right) \\
&\overset{(4)}{\leq} \eta^2\|\nabla L(\mathbf{w}(t))\|^2 + Ct^{-1}\log^{\frac{1}{\nu}-2}(t),
\end{aligned} \tag{197}$$

where in (1) we used

$$\mathbf{r}(t) = \mathbf{w}(t) - g(t)\hat{\mathbf{w}} - g^{1-\nu}(t)\tilde{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t)\hat{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t)\log^{-1}(t)\bar{\mathbf{w}} + \frac{1}{\nu}g^{1-2\nu}(t)\bar{\mathbf{w}}_2$$

(eq. (172)), in (2) we used $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\nabla\mathcal{L}(\mathbf{w}(t))$ (eq. (2)) and denoted

$$\begin{aligned}
\boldsymbol{\psi}_1(t) = -\hat{\mathbf{w}}\left(\frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t) - \frac{1}{\nu}g^{1-\nu}(t+1)\tilde{h}(t+1)\right) - \tilde{\mathbf{w}}\left(g^{1-\nu}(t+1) - g^{1-\nu}(t)\right) \\
- \frac{1}{\nu}\bar{\mathbf{w}}\left(g^{1-\nu}(t)\log^{-1}(t) - g^{1-\nu}(t+1)\log^{-1}(t+1)\right) - \frac{1}{\nu}\bar{\mathbf{w}}_2\left(g^{1-2\nu}(t) - g^{1-2\nu}(t+1)\right).
\end{aligned}$$

In (3) we used $g(t+1) - g(t) = \Theta\left(t^{-1}\log^{\frac{1}{\nu}-1}(t)\right)$ (from eq. (178)) and $\|\boldsymbol{\psi}_1\| = O\left(t^{-1}\log^{\frac{1}{\nu}-2}(t)\right)$ (from eqs. 179, 182, 185, 188, 191) and also $\nabla\mathcal{L}(\mathbf{w}(t)) = o(1)$ from lemma 5. In 4 we used that

$$\hat{\mathbf{w}}^\top\nabla\mathcal{L}(\mathbf{w}(t)) = \sum_{n=1}^{N}\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\hat{\mathbf{w}}^\top\mathbf{x}_n \leq 0$$

since $\hat{\mathbf{w}}^\top\mathbf{x}_n \geq 1$ from the definition of $\hat{\mathbf{w}}$ and $\ell'(u) \leq 0$ and defined $C > 0$.
Next, we will use the following the following Lemma 9 which we prove in appendix K.3.3:

**Lemma 9.** *We have*

$$\exists C_1, t_1 \;:\; \forall t > t_1 \;:\; \left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq C_1 t^{-1}\log^{\frac{1}{\nu}-2}(t) \tag{198}$$

From eq. (198) in Lemma 9, we can find $C_1, t_1$ such that $\forall t > t_1$:

$$\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq C_1 t^{-1}\log^{\frac{1}{\nu}-2}(t) \tag{199}$$

Also, from Lemma 7 we know that:

$$\|\nabla\mathcal{L}(\mathbf{w}(u))\|^2 = o(1) \text{ and } \sum_{u=0}^{\infty}\|\nabla\mathcal{L}(\mathbf{w}(u))\|^2 < \infty \tag{200}$$

Substituting eqs. 197, 200 and 199 into eq. (196), we find

$$\begin{aligned}
&\|\mathbf{r}(t)\|^2 - \|\mathbf{r}(t_1)\|^2 \\
&= \sum_{u=t_1}^{t-1}\left[\|\mathbf{r}(u+1)\|^2 - \|\mathbf{r}(u)\|^2\right] \\
&\leq \sum_{u=t_1}^{t-1}\tilde{C}t^{-1}\log^{\frac{1}{\nu}-2}(t) = O(\log^{\frac{1}{\nu}-1}(t))
\end{aligned}$$

Therefore, $\|\mathbf{r}(t)\|^2 = O(g^{1-\nu}(t))$ and $\|\mathbf{r}(t)\| = o(g^{1-\nu}(t))$.

### K.3.3 Proof Of Lemma 9

Recall that we defined

$$\mathbf{r}(t) = \mathbf{w}(t) - g(t)\hat{\mathbf{w}} - g^{1-\nu}(t)\tilde{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t)\hat{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t)\log^{-1}(t)\bar{\mathbf{w}} + \frac{1}{\nu}g^{1-\nu}(t)g^{-\nu}(t)\bar{\mathbf{w}}_2,$$

where $\hat{\mathbf{w}},\ \tilde{\mathbf{w}},\ \bar{\mathbf{w}}$ and $\bar{\mathbf{w}}_2$ were defined in section K.3.

**Lemma 9.** *We have*

$$\exists C_1, t_1\ :\ \forall t > t_1\ :\ (\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-1}\log^{\frac{1}{\nu}-2}(t) \tag{198}$$

We examine the expression we wish to bound:

$$\begin{aligned}
&(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t)\\
&\overset{(1)}{=} \Big( -\eta\nabla L(\mathbf{w}(t)) - \hat{\mathbf{w}}\left(g(t+1) - g(t) + \nu^{-1}g^{1-\nu}(t)\tilde{h}(t) - \nu^{-1}g^{1-\nu}(t+1)\tilde{h}(t+1)\right)\\
&-\tilde{\mathbf{w}}(g^{1-\nu}(t+1) - g^{1-\nu}(t)) - \frac{1}{\nu}\bar{\mathbf{w}}\left(g^{1-\nu}(t)\log^{-1}(t) - g^{1-\nu}(t+1)\log^{-1}(t+1)\right)\\
&-\frac{1}{\nu}\bar{\mathbf{w}}_2\left(g^{1-2\nu}(t) - g^{1-2\nu}(t+1)\right)\Big)^\top \mathbf{r}(t)\\
&= -\eta\sum_{n=1}^{N}\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\mathbf{x}_n^\top\mathbf{r}(t) - \hat{\mathbf{w}}^\top\mathbf{r}(t)\left(g(t+1) - g(t) + \frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t) - \frac{1}{\nu}g^{1-\nu}(t+1)\tilde{h}(t+1)\right)\\
&-\tilde{\mathbf{w}}^\top\mathbf{r}(t)(g^{1-\nu}(t+1) - g^{1-\nu}(t)) - \bar{\mathbf{w}}^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-\nu}(t)\log^{-1}(t) - g^{1-\nu}(t+1)\log^{-1}(t+1)\right)\\
&-\bar{\mathbf{w}}_2^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-2\nu}(t) - g^{1-2\nu}(t+1)\right)\\
&\overset{(2)}{=} \hat{\mathbf{w}}^\top\mathbf{r}(t)\left[\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\left(1 + h(t) + h_2(t)\right) - \left(g(t+1) - g(t)\right)\right] - \eta\sum_{n\notin\mathcal{S}}\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\mathbf{x}_n^\top\mathbf{r}(t)\\
&-\eta\sum_{n\in\mathcal{S}}\left[\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\left(1 + h(t) + h_2(t)\right)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}}) + \ell'(\mathbf{w}^\top(t)\mathbf{x}_n)\right]\mathbf{x}_n^\top\mathbf{r}(t)\\
&-\tilde{\mathbf{w}}^\top\mathbf{r}(t)(g^{1-\nu}(t+1) - g^{1-\nu}(t)) - \hat{\mathbf{w}}^\top\mathbf{r}(t)\left(\frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t) - \frac{1}{\nu}g^{1-\nu}(t+1)\tilde{h}(t+1)\right)\\
&-\bar{\mathbf{w}}^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-\nu}(t)\log^{-1}(t) - g^{1-\nu}(t+1)\log^{-1}(t+1)\right)\\
&-\bar{\mathbf{w}}_2^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-2\nu}(t) - g^{1-2\nu}(t+1)\right),
\end{aligned} \tag{201}$$

where in (1) we used eq. (172) ($\mathbf{r}(t)$ definition) and in (2) we used

$$\hat{\mathbf{w}} = \sum_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n = \sum_{n\in\mathcal{S}}\exp(-\tilde{\mathbf{w}}^\top\mathbf{x}_n)\mathbf{x}_n.$$

From $\bar{\mathbf{w}}$ and $\bar{\mathbf{w}}_2$ definitions (eqs. 175 and 176) we can find $\{\delta_n\}_{n=1}^d, \{\zeta_n\}_{n=1}^d$ such that:

$$\bar{\mathbf{w}} = \sum_{n\in\mathcal{S}}\delta_n\mathbf{x}_n \tag{202}$$

$$\bar{\mathbf{w}}_2 = \sum_{n\in\mathcal{S}}\zeta_n\mathbf{x}_n \tag{203}$$

Using Taylor's theorem, the last two equations, eq 174 and eqs. 182-193 we can find $\left\{\tilde{\lambda}_n\right\}_{n=1}^{d}$ such that:

$$
\begin{aligned}
&-\tilde{\mathbf{w}}^\top\mathbf{r}(t)(g^{1-\nu}(t+1)-g^{1-\nu}(t))-\hat{\mathbf{w}}^\top\mathbf{r}(t)\left(\frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t)-\frac{1}{\nu}g^{1-\nu}(t+1)\tilde{h}(t+1)\right)\\
&-\bar{\mathbf{w}}^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-\nu}(t)\log^{-1}(t)-g^{1-\nu}(t+1)\log^{-1}(t+1)\right)\\
&-\bar{\mathbf{w}}_2^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-2\nu}(t)-g^{1-2\nu}(t+1)\right)\\
&=-\tilde{\mathbf{w}}^\top\mathbf{r}(t)\left(g^{1-\nu}(t+1)g^{1-\nu}(t)-\left((g^{1-\nu}(t))'+\frac{1}{2}(g^{1-\nu}(t))''\right)\right)\\
&-\sum_{n\in\mathcal{S}}\lambda_n\mathbf{x}_n^\top\mathbf{r}(t)\left((g^{1-\nu}(t))'+\frac{1}{2}(g^{1-\nu}(t))''\right)\\
&+\hat{\mathbf{w}}^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-\nu}(t+1)\tilde{h}(t+1)-g^{1-\nu}(t)\tilde{h}(t)-\left((g^{1-\nu}(t)\tilde{h}(t))'+\frac{1}{2}(g^{1-\nu}(t)\tilde{h}(t))''\right)\right)\\
&+\frac{1}{\nu}\sum_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n^\top\mathbf{r}(t)\left((g^{1-\nu}(t)\tilde{h}(t))'+\frac{1}{2}(g^{1-\nu}(t)\tilde{h}(t))''\right)\\
&+\bar{\mathbf{w}}^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-\nu}(t+1)\log^{-1}(t+1)-g^{1-\nu}(t)\log^{-1}(t)-\left((g^{1-\nu}(t)\log^{-1}(t))'+\frac{1}{2}(g^{1-\nu}(t)\log^{-1}(t))''\right)\right)\\
&+\frac{1}{\nu}\sum_{n\in\mathcal{S}}\delta_n\mathbf{x}_n^\top\mathbf{r}(t)\left((g^{1-\nu}(t)\log^{-1}(t))'+\frac{1}{2}(g^{1-\nu}(t)\log^{-1}(t))''\right)\\
&+\bar{\mathbf{w}}_2^\top\mathbf{r}(t)\frac{1}{\nu}\left(g^{1-2\nu}(t+1)-g^{1-2\nu}(t)-\left((g^{1-2\nu}(t))'+\frac{1}{2}(g^{1-2\nu}(t))''\right)\right)\\
&+\frac{1}{\nu}\sum_{n\in\mathcal{S}}\zeta_n\mathbf{x}_n^\top\mathbf{r}(t)\left((g^{1-2\nu}(t))'+\frac{1}{2}(g^{1-2\nu}(t))''\right)\\
&\le-\sum_{n\in\mathcal{S}}\lambda_n\bar{C}_1\mathbf{x}_n^\top\mathbf{r}(t)\frac{1}{t}\log^{\frac{1}{\nu}-2}(t)-\frac{\eta}{\nu}\sum_{n\in\mathcal{S}}\tilde{\lambda}_n\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})\frac{1}{t}\log^{\frac{1}{\nu}-3}(t)f_3(t)\mathbf{x}_n^\top\mathbf{r}(t)+o(t^{-1-\epsilon}),
\end{aligned}
\tag{204}
$$

where $\bar{C}_1=\frac{1-\nu}{\nu}$ was defined in section K.3.1, $\epsilon>0$ and $f_3(t)=o(\log^\epsilon(t))$. In the last transition we used $\hat{\mathbf{w}}^\top\mathbf{r}(t)=o(t)$, $\tilde{\mathbf{w}}^\top\mathbf{r}(t)=o(t)$, $\bar{\mathbf{w}}^\top\mathbf{r}(t)=o(t)$, $\bar{\mathbf{w}}_2^\top\mathbf{r}(t)=o(t)$ since:

$$
\begin{aligned}
&\|\mathbf{r}(t)\|=\\
&\left\|\mathbf{w}(0)-\eta\sum_{u=0}^{t-1}\nabla\mathcal{L}(\mathbf{w}(u))-g(t)\hat{\mathbf{w}}-g^{1-\nu}(t)\tilde{\mathbf{w}}+\frac{1}{\nu}g^{1-\nu}(t)\tilde{h}(t)\hat{\mathbf{w}}+\frac{1}{\nu}g^{1-\nu}(t)\log^{-1}(t)\bar{\mathbf{w}}+\frac{1}{\nu}g^{1-\nu}(t)g^{-\nu}(t)\bar{\mathbf{w}}_2\right\|\\
&\le\eta t\min_{0\le u\le t}\|\nabla\mathcal{L}(\mathbf{w}(u))\|+O(g(t))=o(t),
\end{aligned}
$$

where in the last line we used that $\nabla\mathcal{L}(\mathbf{w}(t))=o(1)$, from Lemma 7.
Using eq. (204), eq. (201) can be upper bounded by

$$
\begin{aligned}
&\hat{\mathbf{w}}^\top\mathbf{r}(t)\left[\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\left(1+h(t)+h_2(t)\right)-\left(g(t+1)-g(t)\right)\right]-\eta\sum_{n\notin\mathcal{S}}\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\mathbf{x}_n^\top\mathbf{r}(t)\\
&-\eta\sum_{n\in\mathcal{S}}\left[\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\tilde{f}_4(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})+\ell'(\mathbf{w}(t)^\top(t)\mathbf{x}_n)\right]\mathbf{x}_n^\top\mathbf{r}(t),
\end{aligned}
\tag{205}
$$

where we recall we defined $\tilde{\gamma}_n\triangleq\frac{\nu\bar{C}_1\lambda_n}{\eta}\exp(\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})$ and we define $\tilde{f}_4(t)=1+h(t)+\tilde{\gamma}_n\log^{-1}(t)+\tilde{\lambda}_n\log^{-2}(t)f_3(t)+h_2(t)$.

We examine the three terms in eq. (205). The first term can be upper bounded $\forall t > t_2$ by

$$\begin{aligned}
&\hat{\mathbf{w}}^\top \mathbf{r}(t) \left[\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \left(1 + h(t) + h_2(t)\right) - \left(g(t+1) - g(t)\right)\right] \\
&\overset{(1)}{\leq} \left|\hat{\mathbf{w}}^\top \mathbf{r}(t)\right| C_2 t^{-3} \log^{\frac{1}{\nu}-1}(t) \\
&\overset{(2)}{\leq} C_2 t^{-2} \log^{\frac{1}{\nu}-1}(t),
\end{aligned} \tag{206}$$

where in (1) we used

$$\left|\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \left(1 + h(t) + h_2(t)\right) - \left(g(t+1) - g(t)\right)\right| = \Theta(t^{-3} \log^{\frac{1}{\nu}-1}(t))),$$

and therefore $\exists t_2, C_2$ such that $\forall t > t_2$:

$$\left|\frac{1}{\nu} t^{-1} \log^{\frac{1}{\nu}-1}(t) \left(1 + h(t) + h_2(t)\right) - \left(g(t+1) - g(t)\right)\right| \leq C_2 t^{-3} \log^{\frac{1}{\nu}-1}(t)$$

. In (2) we used $\hat{\mathbf{w}}^\top \mathbf{r}(t) = o(t)$.
Next, we wish to upper bound the second term in eq. (205). If $\mathbf{x}_n^\top \mathbf{r}(t) \geq 0$ then using eq. (172) ($\mathbf{r}(t)$ definition) we can show that

$$\left(\mathbf{w}(t)^\top \mathbf{x}_n\right) \geq \left(g(t) \hat{\mathbf{w}}^\top \mathbf{x}_n + g^{1-\nu}(t) \left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu} \tilde{h}(t) \hat{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu} \log^{-1}(t) \bar{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu} g^{-\nu}(t) \bar{\mathbf{w}}_2^\top \mathbf{x}_n\right)\right)^\nu \tag{207}$$

Using the last equation, and eq. (195) we can also show that

$$\begin{aligned}
&\exp(-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu) \\
&\leq t^{-(\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu} \left(\nu \log^{1-\frac{1}{\nu}}(t)\right)^{-(\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu} \exp\left(-\nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \tilde{f}(t) - \tilde{f}_2(t) - f_2(t)\right) \\
&\overset{(1)}{\leq} t^{-(\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu} \left(\nu \log^{1-\frac{1}{\nu}}(t)\right)^{-(\hat{\mathbf{w}}^\top \mathbf{x}_n)^\nu} \exp\left(-\nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - 1\right)\right), \ \forall t > t_3,
\end{aligned} \tag{208}$$

where in (1) we used the fact that $\exists t_3 > \bar{t}$ such that $\forall t > t_3$:

$$\tilde{f}(t) + \frac{(\hat{\mathbf{w}}^\top \mathbf{x}_n)^{1-\nu}}{\nu} (f_2(t) + \tilde{f}_2(t)) \geq -1$$

We divide into two cases.
First if $\mathbf{x}_n^\top \mathbf{r}(t) \leq C_3 \log^{\frac{1}{\nu}+1}(t)$:

$$\begin{aligned}
&-\eta \sum_{n \notin \mathcal{S}} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(1)}{\leq} \eta \sum_{\substack{n \notin \mathcal{S} \\ \mathbf{x}_n^\top \mathbf{r}(t) \geq 0}} \left(1 + \exp(-\mu_+ (\mathbf{w}(t)^\top \mathbf{x}_n)^\nu)\right) \exp(-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(2)}{\leq} \eta \sum_{\substack{n \notin \mathcal{S} \\ \mathbf{x}_n^\top \mathbf{r}(t) \geq 0}} \left(1 + \left[t^{-\theta^\nu} \left(\nu \log^{1-\frac{1}{\nu}}(t)\right)^{-\theta^\nu} \exp\left(-\nu (\hat{\mathbf{w}}^\top \mathbf{x}_n)^{\nu-1} \left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - 1\right)\right)\right]^{\mu_+}\right) \\
&t^{-\theta^\nu} \left(\nu \log^{1-\frac{1}{\nu}}(t)\right)^{-\theta^\nu} \exp\left(-\nu \theta^{\nu-1} \left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - 1\right)\right) C_1 \log^{\frac{1}{\nu}+1}(t) \\
&\overset{(3)}{\leq} 2\eta C_3 N t^{-\theta^\nu} \nu^{-\theta^\nu} \log^{\theta^\nu \left(\frac{1}{\nu}-1\right)} \exp\left(-\nu \min_{n \notin \mathcal{S}} (\hat{\mathbf{w}}^T \mathbf{x}_n)^{\nu-1} (\tilde{\mathbf{w}}^\top \mathbf{x}_n - 1)\right) \log^{\frac{1}{\nu}+1}(t) \\
&\leq \tilde{C}_2 t^{-\theta^\nu} \log^{\theta^\nu \left(\frac{1}{\nu}-1\right) + \frac{1}{\nu}+1}(t), \ \forall t > t_2',
\end{aligned} \tag{209}$$

where in (1) we used eq. (132), in (2) we used 208 and $\theta = \min_{n\notin\mathcal{S}} \hat{\mathbf{w}}^T\mathbf{x}_n > 1$, in (3) we used the fact that $\exists t'$ such that $\forall t > t'$ : $\left[t^{-\theta^\nu}\left(\nu\log^{1-\frac{1}{\nu}}(t)\right)^{-\theta^\nu}\exp\left(-\nu(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-1}\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1\right)\right)\right]^{\mu_+} \leq 1$. We defined $t_2' = \max(t_3, t')$ and $\tilde{C}_2 = 2\eta C_3 N\nu^{-\theta^\nu}\exp\left(-\nu\min_{n\notin\mathcal{S}}(\hat{\mathbf{w}}^T\mathbf{x}_n)^{\nu-1}(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1)\right)$.
Second, if $\mathbf{x}_n^\top\mathbf{r}(t) > C_3\log^{\frac{1}{\nu}+1}(t)$:

$$\begin{aligned}
&(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu\\
&\overset{(1)}{=}\left(g(t)\hat{\mathbf{w}}^\top\mathbf{x}_n + g^{1-\nu}(t)\tilde{f}(t)+\mathbf{x}_n^\top\mathbf{r}(t)\right)^\nu\\
&=(\mathbf{x}_n^\top\mathbf{r}(t))^\nu\left(1+\frac{g(t)\hat{\mathbf{w}}^\top\mathbf{x}_n+g^{1-\nu}(t)\tilde{f}(t)}{\mathbf{x}_n^\top\mathbf{r}(t)}\right)^\nu\\
&\overset{(2)}{\geq}(\mathbf{x}_n^\top\mathbf{r}(t))^\nu\left(\nu+\left(\frac{g(t)\hat{\mathbf{w}}^\top\mathbf{x}_n+g^{1-\nu}(t)\tilde{f}(t)}{\mathbf{x}_n^\top\mathbf{r}(t)}\right)^\nu\right)\\
&=\nu(\mathbf{x}_n^\top\mathbf{r}(t))^\nu+\left(g(t)\hat{\mathbf{w}}^\top\mathbf{x}_n+g^{1-\nu}(t)\tilde{f}(t)\right)^\nu\\
&\overset{(3)}{=}\nu(\mathbf{x}_n^\top\mathbf{r}(t))^\nu+(\hat{\mathbf{w}}^\top\mathbf{x}_n)^\nu\log(t)+(\hat{\mathbf{w}}^\top\mathbf{x}_n)^\nu\log(\nu\log^{1-\frac{1}{\nu}}(t))+\nu(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-1}\tilde{f}(t)\\
&+\tilde{f}_2(t)+f_2(t),
\end{aligned}\tag{210}$$

where in (1) we used eq. (172), in (2) we used that $\forall x < 0.5, a < 1:\ (1+x)^a \geq a + x^a$ and that $\exists t_4$ such that

$$\forall t > t_4:\ \frac{g(t)\hat{\mathbf{w}}^\top\mathbf{x}_n+g^{1-\nu}(t)\tilde{f}(t)}{\mathbf{x}_n^\top\mathbf{r}(t)} < 0.5$$

and in (3) we used eq. (195).
Using 210 we can upper bound $\exp(-(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu),\ \forall t > t_4$:

$$\begin{aligned}
&\exp(-(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu)\\
&\overset{(1)}{\leq}\exp\left(-\nu\left(\mathbf{x}_n^\top\mathbf{r}(t)\right)^\nu\right)t^{-(\hat{\mathbf{w}}^\top\mathbf{x}_n)^\nu}\left(\nu\log^{1-\frac{1}{\nu}}(t)\right)^{-(\hat{\mathbf{w}}^\top\mathbf{x}_n)^\nu}\exp\left(-\nu(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-1}\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1\right)\right)
\end{aligned}\tag{211}$$

Using the last equation we can upper bound the second term in eq. (205) $\forall t > t_5$:

$$\begin{aligned}
&-\eta\sum_{n\notin\mathcal{S}}\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\mathbf{x}_n^\top\mathbf{r}(t)\\
&\overset{(1)}{\leq}\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\mathbf{r}(t)\geq0}}\left(1+\exp(-\mu_+(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu)\right)\exp(-(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu)\mathbf{x}_n^\top\mathbf{r}(t)\\
&\overset{(2)}{\leq}\eta\sum_{\substack{n\notin\mathcal{S}\\ \mathbf{x}_n^\top\mathbf{r}(t)\geq0}}\left(1+\left[t^{-\theta^\nu}\left(\nu\log^{1-\frac{1}{\nu}}(t)\right)^{-\theta^\nu}\exp\left(-\nu(\hat{\mathbf{w}}^\top\mathbf{x}_n)^{\nu-1}\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1\right)\right)\right]^{\mu_+}\right)\\
&t^{-\theta^\nu}\left(\nu\log^{1-\frac{1}{\nu}}(t)\right)^{-\theta^\nu}\exp\left(-\nu\theta^{\nu-1}\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1\right)\right)\exp\left(-\nu\left(\mathbf{x}_n^\top\mathbf{r}\right)^\nu\right)\mathbf{x}_n^\top\mathbf{r}(t)\\
&\overset{(3)}{\leq}2\eta Nt^{-\theta^\nu}\nu^{-\theta^\nu}\log^{\theta^\nu\left(\frac{1}{\nu}-1\right)}\exp\left(-\nu\min_{n\notin\mathcal{S}}(\hat{\mathbf{w}}^T\mathbf{x}_n)^{\nu-1}(\tilde{\mathbf{w}}^\top\mathbf{x}_k-1)\right)\\
&\leq\tilde{C}_2' t^{-\theta^\nu}\log^{\theta^\nu\left(\frac{1}{\nu}-1\right)},
\end{aligned}\tag{212}$$

where in (1) we used 132, in (2) we used eqs. 208 and 211 and $\theta = \min_{n\notin\mathcal{S}}\hat{\mathbf{w}}^T\mathbf{x}_n > 1$ (eq. (130)). In (3) we used the fact that $\lim_{t\to\infty}\exp\left(-\nu(\mathbf{x}_n^\top\mathbf{r}(t))^\nu\right)\mathbf{x}_n^\top\mathbf{r}(t) = 0$, therefore exists $t''$ such that $\forall t > t'$ : $\exp\left(-\nu(\mathbf{x}_n^\top\mathbf{r}(t))^\nu\right)\mathbf{x}_n^\top\mathbf{r}(t) \leq 1$. We define $t_5 = \max(t_4, t'')$ and

$$\tilde{C}_2' = 2\eta N\nu^{-\theta^\nu}\exp\left(-\nu\min_{n\notin\mathcal{S}}(\hat{\mathbf{w}}^T\mathbf{x}_n)^{\nu-1}(\tilde{\mathbf{w}}^\top\mathbf{x}_k-1)\right).$$

Next, we wish to upper bound the third term in eq. (205):

$$-\eta\sum_{n\in\mathcal{S}}\left[\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\tilde{f}_4(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})+\ell'(\mathbf{w}(t)^\top(t)\mathbf{x}_n)\right]\mathbf{x}_n^\top\mathbf{r}(t), \tag{213}$$

where we recall we denoted $\tilde{f}_4(t)=1+h(t)+\tilde{\gamma}_n\log^{-1}(t)+\tilde{\lambda}_n\log^{-2}(t)f_3(t)+h_2(t)$.
We can use eq. (195) to show that for $n\in\mathcal{S}$:

$$\begin{aligned}&\left(g(t)+g^{1-\nu}(t)\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-\frac{1}{\nu}\tilde{h}(t)-\frac{1}{\nu}\log^{-1}(t)\bar{\mathbf{w}}^\top\mathbf{x}_n-\frac{1}{\nu}g^{-\nu}(t)\bar{\mathbf{w}}_2^\top\mathbf{x}_n\right)\right)^\nu\\&=\log(t)+\log(\nu\log^{1-\frac{1}{\nu}}(t))+\nu\tilde{\mathbf{w}}^\top\mathbf{x}_n-\tilde{h}(t)-\log^{-1}(t)(\bar{\mathbf{w}}^\top\mathbf{x}_n)-g^{-\nu}(t)(\bar{\mathbf{w}}_2^\top\mathbf{x}_n)\\&+\frac{\nu(\nu-1)}{2}g^{-\nu}(t)\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n\right)^2+\frac{(\nu-1)}{2\nu}\log^{-1}(t)\log^2(\nu\log^{1-\frac{1}{\nu}}(t))+f_2(t)\\&=\log(t)+\log(\nu\log^{1-\frac{1}{\nu}}(t))+\nu\tilde{\mathbf{w}}^\top\mathbf{x}_n-h(t)-\log^{-1}(t)(\bar{\mathbf{w}}^\top\mathbf{x}_n)+f_2(t),\end{aligned} \tag{214}$$

where in the last transition we used that $\bar{\mathbf{w}}_2^\top\mathbf{x}_n=\frac{\nu(\nu-1)}{2}(\tilde{\mathbf{w}}^\top\mathbf{x}_n)^2$ (eq. (176)) and $\tilde{h}(t)$ definition (eq. (177)).
We examine each term $k$ in the sum, and divide into two cases, depending on the sign of $\mathbf{x}_k^\top\mathbf{r}(t)$.
From this point we assume $\nu>\frac{1}{4}$.
First, if $\mathbf{x}_k^\top\mathbf{r}(t)\geq 0$, then using eqs. 132 and 214 term $k$ in equation 213 can be upper bounded by:

$$\begin{aligned}&-\eta\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})\Big[\Big(1+h(t)+\tilde{\gamma}_n\log^{-1}(t)+\tilde{\lambda}_n\log^{-2}(t)f_3(t)+h_2(t)\Big)\\&-\left(1+\exp(-\mu_+(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu)\right)\exp\left(h(t)+\log^{-1}(t)\bar{\mathbf{w}}^\top\mathbf{x}_n-f_2(t)\right)\Big]\mathbf{x}_k^\top\mathbf{r}(t)\\&\overset{(1)}{\leq}-\eta\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})\Bigg[\tilde{\lambda}_n\log^{-2}(t)f_3(t)+h_2(t)+f_2(t)-\tilde{f}_5^2(t)\\&-\left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1\right)\right)\right]^{\mu_+}\left(1+\tilde{f}_5(t)+\tilde{f}_5^2(t)\right)\Bigg]\mathbf{x}_k^\top\mathbf{r}(t)\\&\overset{(2)}{\leq}\eta\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})\log^{-\frac{1}{2\nu}}(t)\mathbf{x}_k^\top\mathbf{r}(t),\end{aligned} \tag{215}$$

where in (1) we defined $\tilde{f}_5(t)=h(t)+\tilde{\gamma}_n\log^{-1}(t)-f_2(t)$ and used that:

$$\begin{aligned}&\left(1+\exp(-\mu_+(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu)\right)\exp\left(h(t)+\log^{-1}(t)\bar{\mathbf{w}}^\top\mathbf{x}_n-f_2(t)\right)\\&\overset{(a)}{\leq}\left(1+\left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1\right)\right)\right]^{\mu_+}\right)\left(1+\tilde{f}_5(t)+\tilde{f}_5^2(t)\right),\end{aligned}$$

where in (a) we used $\forall x\geq 0:\ e^{-x}\leq 1-x+x^2$, $h(t)\leq 0$ and $\bar{\mathbf{w}}^\top\mathbf{x}_n=\tilde{\gamma}_n$ (eq. (175)). In transition (2) we used the fact that all the terms in the square brackets are $o(\log^{-2+\epsilon}(t)),\ \forall\epsilon>0$ and $-\frac{1}{2\nu}>-2$.
1. If $|\mathbf{x}_k^\top\mathbf{r}(t)|\leq C_4\log^{-1+\frac{1}{2\nu}}(t)$ then term $k$ in equation 213 can be upper bounded by:

$$\eta C_4\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-2}(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}}) \tag{216}$$

2. If $|\mathbf{x}_k^\top\mathbf{r}(t)|>C_4\log^{-1+\frac{1}{2\nu}}(t)$ term $k$ in equation 213 can be upper bounded by:

$$\begin{aligned}&-\eta\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-1}(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}})\Bigg[\tilde{f}_4(t)-\left(1+\left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1\right)\right)\right]^{\mu_+}\right)\\&\exp\left(\tilde{f}_5(t)-C_5\log^{-\frac{1}{2\nu}}(t)-f_4(t)\right)\Bigg]\mathbf{x}_n^\top\mathbf{r}(t),\end{aligned} \tag{217}$$

where $f_4(t) = o(\log^{-\frac{1}{2\nu}}(t))$, since:

$$\begin{aligned}
&(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu = \left(g(t)\hat{\mathbf{w}}^\top \mathbf{x}_n + g^{1-\nu}(t)\left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu}h(t)\hat{\mathbf{w}}^\top \mathbf{x}_n - \frac{1}{\nu}\log^{-1}(t)\bar{\mathbf{w}}^\top \mathbf{x}_n\right) + \mathbf{x}_n^\top \mathbf{r}(t)\right)^\nu \\
&\geq \left(g(t) + g^{1-\nu}(t)\tilde{f}(t) + C_4 \log^{-1+\frac{1}{2\nu}}(t)\right)^\nu \\
&= \left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^\nu \left(1 + \frac{C_4 \log^{-1+\frac{1}{2\nu}}(t)}{g(t) + g^{1-\nu}(t)\tilde{f}(t)}\right)^\nu \\
&= \left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^\nu + \nu\left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1} C_4 \log^{-1+\frac{1}{2\nu}}(t) + f_4(t) \\
&\geq \log(t) + \log(\nu \log^{1-\frac{1}{\nu}}(t)) + \nu\tilde{\mathbf{w}}^\top \mathbf{x}_n - h(t) - \tilde{\gamma}_n \log^{-1}(t) + f_2(t) + C_5 \log^{-\frac{1}{2\nu}}(t) + f_4(t),
\end{aligned}$$

in the last two transitions we used Taylor's theorem and eq. (214).
eq. (217) can be upper bounded by zero since:

$$\begin{aligned}
&\left[\tilde{f}_4(t) - \left(1 + \left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - 1\right)\right)\right]^{\mu_+}\right)\exp\left(\tilde{f}_5(t) - C_5\log^{-\frac{1}{2\nu}}(t) - f_4(t)\right)\right] \\
&\overset{(1)}{\geq} \left[\tilde{f}_4(t) - \left(1 + \left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - 1\right)\right)\right]^{\mu_+}\right)\right. \\
&\left.\left(1 + \tilde{f}_5(t) - C_5\log^{-\frac{1}{2\nu}}(t) - f_4(t) + \left(\tilde{f}_5(t) - C_5\log^{-\frac{1}{2\nu}}(t) - f_4(t)\right)^2\right)\right] \\
&\overset{(2)}{\geq} \left[\tilde{\lambda}_n\log^{-2}(t)f_3(t) + h_2(t) + f_2(t) + C_5\log^{-\frac{1}{2\nu}}(t) + f_4(t) - \left(\tilde{f}_5(t) - C_5\log^{-\frac{1}{2\nu}}(t) - f_4(t)\right)^2\right. \\
&- \left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - 1\right)\right)\right]^{\mu_+}\left(1 + \tilde{f}_5(t) - C_5\log^{-\frac{1}{2\nu}}(t) - f_4(t)\right. \\
&\left.\left.+ \left(\tilde{f}_5(t) - C_5\log^{-\frac{1}{2\nu}}(t) - f_4(t)\right)^2\right)\right] \overset{(3)}{\geq} 0,
\end{aligned}$$

where in (1) we used $e^{-x} \leq 1 - x + x^2, \forall x \geq 0$ and $h(t) < 0$, in (2) we used $\tilde{f}_4(t) = 1 + h(t) + \tilde{\gamma}_n \log^{-1}(t) + \tilde{\lambda}_n \log^{-2}(t) f_3(t) + h_2(t)$, $\tilde{f}_5(t) = h(t) + \tilde{\gamma}_n \log^{-1}(t) - f_2(t)$ and in (3) we used the fact that $\log^{-\frac{1}{2\nu}}(t)$ decreases slower than the other terms. Second, if $\mathbf{x}_k^\top \mathbf{r}(t) < 0$ , then using eq. (133), term $k$ in equation 213 can be upper bounded by:

$$\begin{aligned}
&\eta\left[\nu^{-1}t^{-1}\log^{\frac{1}{\nu}-1}(t)\tilde{f}_4(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}}) - (1 - \exp(-\mu_-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu))\exp(-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu)\right]|\mathbf{x}_k^\top \mathbf{r}(t)| \\
&\leq \eta\nu^{-1}t^{-1}\log^{\frac{1}{\nu}-1}(t)\exp(-\nu\mathbf{x}_k^\top\tilde{\mathbf{w}})\left[\tilde{f}_4(t) - (1 - \exp(-\mu_-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu))\right. \\
&\left.\exp\left(\tilde{f}_5(t) - \nu\left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_n^\top \mathbf{r}(t)\right)\right]|\mathbf{x}_k^\top \mathbf{r}(t)|,
\end{aligned} \tag{218}$$

where in the last transition we used:

$$\begin{aligned}
&(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu \\
&\overset{(1)}{=} \left(g(t) + g^{1-\nu}(t)\tilde{f}(t) + \mathbf{x}_n^\top \mathbf{r}(t)\right)^\nu \\
&\overset{(2)}{\leq} \left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^\nu + \nu\left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(3)}{\leq} \log(t) + \log(\nu\log^{1-\frac{1}{\nu}}(t)) + \nu\tilde{\mathbf{w}}^\top \mathbf{x}_n - h(t) - \log^{-1}(t)(\bar{\mathbf{w}}^\top \mathbf{x}_n) + f_2(t) + \nu\left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_n^\top \mathbf{r}(t) \\
&\leq \log(t) + \log(\nu\log^{1-\frac{1}{\nu}}(t)) + \nu\tilde{\mathbf{w}}^\top \mathbf{x}_n - h(t) - \log^{-1}(t)(\bar{\mathbf{w}}^\top \mathbf{x}_n) + f_2(t),
\end{aligned}$$

where in (1) we used eq. (172) ($\mathbf{r}(t)$ definition), in (2) we used Bernoulli's inequality:

$$\forall 0 < r < 1, x \geq -1 : (1+x)^r \leq 1 + rx \tag{219}$$

and the fact that from Lemma 6 $\lim_{t\to\infty}\mathbf{w}(t)^\top\mathbf{x}_n=\infty$ and therefore for sufficiently large t

$$\frac{\mathbf{x}_n^\top\mathbf{r}(t)}{g(t)+g^{1-\nu}(t)\tilde{f}(t)}\geq -1 \tag{220}$$

In (3) we used eq. (214).
We further divide into cases.
1. If $|\mathbf{x}_k^\top\mathbf{r}(t)|\leq C_4\log^{-1+\frac{1}{2\nu}}(t)$:
We can lower bound $(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu$ as follows

$$\begin{aligned}
&(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu\\
&\overset{(1)}{=}\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)+\mathbf{x}_n^\top\mathbf{r}(t)\right)^\nu\\
&\geq\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)-C_4\log^{-1+\frac{1}{2\nu}}(t)\right)^\nu\\
&=\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^\nu\left(1+\frac{-C_4\log^{-1+\frac{1}{2\nu}}(t)}{g(t)+g^{1-\nu}(t)\tilde{f}(t)}\right)^\nu\\
&=\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^\nu-\nu\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}C_4\log^{-1+\frac{1}{2\nu}}(t)+f_5(t)\\
&\geq\log(t)+\log(\nu\log^{1-\frac{1}{\nu}}(t))+\nu\tilde{\mathbf{w}}^\top\mathbf{x}_n-h(t)-\log^{-1}(t)(\bar{\mathbf{w}}^\top\mathbf{x}_n)+f_2(t)-C_6\log^{-\frac{1}{2\nu}}(t)+f_5(t)\\
&\geq\log(t)+\log(\nu\log^{1-\frac{1}{\nu}}(t))+\nu\tilde{\mathbf{w}}^\top\mathbf{x}_n-1,
\end{aligned}$$

where $f_5(t)=o(\log^{-\frac{1}{2\nu}}(t))$. In the last transition we used Taylor's theorem and eq. (214).
Using this bound and the fact that $e^x>1+x$ we can find $C_7$ such that eq. (218) can be upper bounded by

$$\eta C_7\frac{1}{\nu}t^{-1}\log^{\frac{1}{\nu}-2}(t)\exp(-\nu\mathbf{x}_n^\top\tilde{\mathbf{w}}) \tag{221}$$

2. If $|\mathbf{x}_k^\top\mathbf{r}(t)|>C_4\log^{-1+\frac{1}{2\nu}}(t)$ then we will show that $\exists t'_-$ such that eq. (218) is strictly negative $\forall t>t'_-$.
Let $M>1$ be some arbitrary constant. Then, since $\exp\left(-\mu_-\left(\mathbf{w}(t)^\top\mathbf{x}_k\right)^\nu\right)\to 0$ from Lemma 6, $\exists t_M>\bar{t}$ such that $\forall t>t_M$ and if

$$\exp\left(-\nu\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_n^\top\mathbf{r}(t)\right)\geq M>1 \tag{222}$$

then

$$\left(1-\exp(-\mu_-(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu)\right)\exp\left(\tilde{f}_5(t)\right)\exp\left(-\nu\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_n^\top\mathbf{r}(t)\right)\geq M'>1$$

Furthermore, if $\exists t>t_M$ such that

$$\exp\left(-\nu\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_n^\top\mathbf{r}(t)\right)<M$$

then $\exists M'',t_6$ such that $\forall t>t_6$: $\left|\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_k^\top\mathbf{r}(t)\right|\leq M''$. We can use this to show that

$$\begin{aligned}
&\left(1-\exp(-\mu_-(\mathbf{w}(t)^\top\mathbf{x}_n)^\nu)\right)\exp\left(\tilde{f}_5(t)-\nu\left(g(t)+g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1}\mathbf{x}_n^\top\mathbf{r}(t)\right)\\
&\geq\left(1-\left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1-M''\right)\right)\right]^{\mu_-}\right)\exp\left(\tilde{f}_5(t)+C_6\log^{-\frac{1}{2\nu}}(t)\right)\\
&\overset{(1)}{\geq}\left(1-\left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top\mathbf{x}_n-1-M''\right)\right)\right]^{\mu_-}\right)\left(1+\tilde{f}_5(t)+C_6\log^{-\frac{1}{2\nu}}(t)\right),
\end{aligned}$$

where in (1) we used $e^x \geq 1 + x$.
Using the last equation we can show that eq. (218) is negative since:

$$\begin{aligned}
&\left[\left(1 + h(t) + \tilde{\gamma}_n \log^{-1}(t) + \tilde{\lambda}_n \log^{-2}(t) f_3(t) + h_2(t)\right) - (1 - \exp(-\mu_-(\mathbf{w}(t)^\top \mathbf{x}_n)^\nu))\right.\\
&\left.\exp\left(h(t) + \tilde{\gamma}_n \log^{-1}(t) - f_2(t) - \nu\left(g(t) + g^{1-\nu}(t)\tilde{f}(t)\right)^{\nu-1} \mathbf{x}_n^\top \mathbf{r}(t)\right)\right]\\
&\leq \left(\tilde{\lambda}_n \log^{-2}(t) f_3(t) + h_2(t) + f_2(t) - C_6 \log^{-\frac{1}{2\nu}}(t)\right.\\
&\left.+ \left[\frac{1}{t}\log^{\frac{1}{\nu}-1}(t)\frac{1}{\nu}\exp\left(-\nu\left(\tilde{\mathbf{w}}^\top \mathbf{x}_n - \tilde{M}\right)\right)\right]^{\mu_-}\left(1 + h(t) + \tilde{\gamma}_n \log^{-1}(t) - f_2(t) + C_6 \log^{-\frac{1}{2\nu}}(t)\right)\right)\\
&\leq 0
\end{aligned} \tag{223}$$

In the last transition we used the fact that $\log^{-\frac{1}{2\nu}}(t)$ decreases slower than the other terms.

To conclude, we choose $t_0 = \max\left[t_2, t_2', t_5, t_-'\right]$. We find that $\forall t > t_0$ , each term in eq. (213) can be upper bounded by either zero, or terms proportional to $t^{-1}\log^{-\frac{1}{\nu}-2}$. Combining this together with eqs. 206, 209, 212 into eq. (205) we obtain (for some positive constants $C_8$, $C_9$, $C_{10}$)

$$\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) \leq C_8 t^{-2} \log^{\frac{1}{\nu}-1}(t) + C_9 t^{-\theta^\nu} (\log(t))^{(\frac{1}{\nu}-1)\theta^\nu} + C_1 0 t^{-1} \log^{\frac{1}{\nu}-2}(t)$$

Therefore, $\exists t_1 > t_0$ and $C_1$ such that eq. (150) holds.