Journal of Machine Learning Research 19 (2018) 1-57 Submitted 4/18; Published 11/18

The Implicit Bias of Gradient Descent on Separable Data

Daniel Soudry DANIEL.SOUDRY@GMAIL.COM
Elad Hoffer ELAD.HOFFER@GMAIL.COM
Mor Shpigel Nacson MOR.SHPIGEL@GMAIL.COM
Department of Electrical Engineering,Technion
Haifa, 320003, Israel

Suriya Gunasekar SURIYA@TTIC.EDU
Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago
Chicago, Illinois 60637, USA

Editor: Leon Bottou

Abstract

We examine gradient descent on unregularized logistic regression problems, with homogeneous linear predictors on linearly separable datasets. We show the predictor converges to the direction of the max-margin (hard margin SVM) solution. The result also generalizes to other monotone decreasing loss functions with an infimum at infinity, to multi-class problems, and to training a weight layer in a deep network in a certain restricted setting. Furthermore, we show this convergence is very slow, and only logarithmic in the convergence of the loss itself. This can help explain the benefit of continuing to optimize the logistic or cross-entropy loss even after the training error is zero and the training loss is extremely small, and, as we show, even if the validation loss increases. Our methodology can also aid in understanding implicit regularization in more complex models and with other optimization methods.

Keywords: gradient descent, implicit regularization, generalization, margin, logistic regression

1. Introduction

It is becoming increasingly clear that implicit biases introduced by the optimization algorithm play a crucial role in deep learning and in the generalization ability of the learned models (Neyshabur et al., 2014, 2015; Zhang et al., 2017; Keskar et al., 2017; Neyshabur et al., 2017; Wilson et al., 2017). In particular, minimizing the training error, without explicit regularization, over models with more parameters and capacity than the number of training examples, often yields good generalization. This is despite the fact that the empirical optimization problem being highly underdetermined. That is, there are many global minima of the training objective, most of which will not generalize well, but the optimization algorithm (*e.g.* gradient descent) biases us toward a particular minimum that *does* generalize well. Unfortunately, we still do not have a good understanding of the biases introduced by different optimization algorithms in different situations.

We do have an understanding of the implicit regularization introduced by early stopping of stochastic methods or, at an extreme, of one-pass (no repetition) stochastic gradient descent (Hardt et al., 2016). However, as discussed above, in deep learning we often benefit from implicit bias even when optimizing the training error to convergence (without early stopping) using stochastic or batch methods. For loss functions with attainable, finite minimizers, such as the squared loss, we have some

understanding of this: in particular, when minimizing an underdetermined least squares problem using gradient descent starting from the origin, it can be shown that we will converge to the minimum Euclidean norm solution. However, the logistic loss, and its generalization the cross-entropy loss which is often used in deep learning, do not admit finite minimizers on separable problems. Instead, to drive the loss toward zero and thus minimize it, the norm of the predictor must diverge toward infinity.

Do we still benefit from implicit regularization when minimizing the logistic loss on separable data? Clearly the norm of the predictor itself is not minimized, since it grows to infinity. However, for prediction, only the direction of the predictor, *i.e.* the normalized $\mathbf{w}(t)/\left\|\mathbf{w}(t)\right\|$, is important. How does $\mathbf{w}(t)/\left\|\mathbf{w}(t)\right\|$ behave as $t \to \infty$ when we minimize the logistic (or similar) loss using gradient descent on separable data, *i.e.*, when it is possible to get zero misclassification error and thus drive the loss to zero?

In this paper, we show that even without any explicit regularization, for all linearly separable datasets, when minimizing logistic regression problems using gradient descent, we have that $\mathbf{w}(t)/\left\|\mathbf{w}(t)\right\|$ converges to the L_2 maximum margin separator, *i.e.* to the solution of the hard margin SVM for homogeneous linear predictors. This happens even though neither the norm $\|\mathbf{w}\|$, nor the margin constraint, are part of the objective or explicitly introduced into optimization. More generally, we show the same behavior for generalized linear problems with any smooth, monotone strictly decreasing, lower bounded loss with an exponential tail. Furthermore, we characterize the rate of this convergence, and show that it is rather slow, wherein for almost all datasets, the distance to the max-margin predictor decreasing only as $O(1/\log(t))$, and in some degenerate datasets, the rate further slows down to $O(\log\log(t)/\log(t))$. This explains why the predictor continues to improve even when the training loss is already extremely small. We emphasize that this bias is specific to gradient descent, and changing the optimization algorithm, *e.g.* using adaptive learning rate methods such as ADAM (Kingma and Ba, 2015), changes this implicit bias.

2. Main Results

Consider a dataset $\{\mathbf{x}_n, y_n\}_{n=1}^N$, with $\mathbf{x}_n \in \mathbb{R}^d$ and binary labels $y_n \in \{-1, 1\}$. We analyze learning by minimizing an empirical loss of the form

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

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector. To simplify notation, we assume that all the labels are positive: $\forall n:\ y_n = 1$ — this is true without loss of generality, since we can always re-define $y_n \mathbf{x}_n$ as $\mathbf{x}_n$.

We are particularly interested in problems that are linearly separable, and the loss is smooth strictly decreasing and non-negative:

Assumption 1 *The dataset is linearly separable:* $\exists \mathbf{w}_*$ *such that* $\forall n:\ \mathbf{w}_*^\top \mathbf{x}_n > 0$.

Assumption 2 $\ell\left(u\right)$ *is a positive, differentiable, monotonically decreasing to zero*[1]*, (so* $\forall u:\ \ell\left(u\right) > 0, \ell'\left(u\right) < 0$*,* $\lim_{u\to\infty} \ell\left(u\right) = \lim_{u\to\infty} \ell'\left(u\right) = 0$*), a* β*-smooth function,* i.e. *its derivative is* β*-Lipshitz and* $\lim_{u\to-\infty} \ell'\left(u\right) \neq 0$*.*

1. The requirement of non-negativity and that the loss asymptotes to zero is purely for convenience. It is enough to require the loss is monotone decreasing and bounded from below. Any such loss asymptotes to some constant, and is thus equivalent to one that satisfies this assumption, up to a shift by that constant.

Assumption 2 includes many common loss functions, including the logistic, exp-loss[2] and probit losses. Assumption 2 implies that $\mathcal{L}(\mathbf{w})$ is a $\beta\sigma_{\max}^2(\mathbf{X})$-smooth function, where $\sigma_{\max}(\mathbf{X})$ is the maximal singular value of the data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$.

Under these conditions, the infimum of the optimization problem is zero, but it is not attained at any finite $\mathbf{w}$. Furthermore, no finite critical point $\mathbf{w}$ exists. We consider minimizing eq. 1 using Gradient Descent (GD) with a fixed learning rate η, *i.e.,* with steps of the form:

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\nabla\mathcal{L}(\mathbf{w}(t)) = \mathbf{w}(t) - \eta\sum_{n=1}^{N}\ell'\left(\mathbf{w}(t)^\top\mathbf{x}_n\right)\mathbf{x}_n. \tag{2}$$

We do not require convexity. Under Assumptions 1 and 2, gradient descent converges to the global minimum (*i.e.* to zero loss) even without it:

Lemma 1 *Let $\mathbf{w}(t)$ be the iterates of gradient descent (eq. 2) with $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathbf{X})$ and any starting point $\mathbf{w}(0)$. Under Assumptions 1 and 2, we have: (1) $\lim_{t\to\text{ßnfty}}\mathcal{L}(\mathbf{w}(t)) = 0$, (2) $\lim_{t\to\infty}\|\mathbf{w}(t)\| = \infty$, and (3) $\forall n: \lim_{t\to\infty}\mathbf{w}(t)^\top\mathbf{x}_n = \infty$.*

Proof Since the data is linearly separable, $\exists\mathbf{w}_*$ which linearly separates the data, and therefore

$$\mathbf{w}_*^\top\nabla\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N}\ell'\left(\mathbf{w}^\top\mathbf{x}_n\right)\mathbf{w}_*^\top\mathbf{x}_n.$$

For any finite $\mathbf{w}$, this sum cannot be equal to zero, as a sum of negative terms, since $\forall n: \mathbf{w}_*^\top\mathbf{x}_n > 0$ and $\forall u: \ell'(u) < 0$. Therefore, there are no finite critical points $\mathbf{w}$, for which $\nabla\mathcal{L}(\mathbf{w}) = \mathbf{0}$. But gradient descent on a smooth loss with an appropriate stepsize is always guaranteed to converge to a critical point: $\nabla\mathcal{L}(\mathbf{w}(t)) \to \mathbf{0}$ (see, *e.g.* Lemma 10 in Appendix A.4, slightly adapted from Ganti (2015), Theorem 2). This necessarily implies that $\|\mathbf{w}(t)\| \to \infty$ while $\forall n: \mathbf{w}(t)^\top\mathbf{x}_n > 0$ for large enough t—since only then $\ell'\left(\mathbf{w}(t)^\top\mathbf{x}_n\right) \to 0$. Therefore, $\mathcal{L}(\mathbf{w}) \to 0$, so GD converges to the global minimum. ■

The main question we ask is: can we characterize the direction in which $\mathbf{w}(t)$ diverges? That is, does the limit $\lim_{t\to\infty}\mathbf{w}(t)/\|\mathbf{w}(t)\|$ always exist, and if so, what is it?

In order to analyze this limit, we will need to make a further assumption on the tail of the loss function:

Definition 2 *A function $f(u)$ has a "tight exponential tail", if there exist positive constants c, a, μ_+, μ_-, u_+ and u_- such that*

$$\forall u > u_+ : f(u) \le c\left(1+\exp\left(-\mu_+ u\right)\right)e^{-au}$$
$$\forall u > u_- : f(u) \ge c\left(1-\exp\left(-\mu_- u\right)\right)e^{-au}.$$

Assumption 3 *The negative loss derivative $-\ell'(u)$ has a tight exponential tail (Definition 2).*

For example, the exponential loss $\ell(u) = e^{-u}$ and the commonly used logistic loss $\ell(u) = \log\left(1+e^{-u}\right)$ both follow this assumption with $a = c = 1$. We will assume $a = c = 1$ — without loss of generality, since these constants can be always absorbed by re-scaling $\mathbf{x}_n$ and η.

We are now ready to state our main result:

2. The exp-loss does not have a global β smoothness parameter. However, if we initialize with $\eta < 1/\mathcal{L}(\mathbf{w}(0))$ then it is straightforward to show the gradient descent iterates maintain bounded local smoothness.

Theorem 3 *For any dataset which is linearly separable (Assumption 1), any β-smooth decreasing loss function (Assumption 2) with an exponential tail (Assumption 3), any stepsize $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathbf{X})$ and any starting point $\mathbf{w}(0)$, the gradient descent iterates (as in eq. 2) will behave as:*

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

where $\hat{\mathbf{w}}$ is the L_2 max margin vector (the solution to the hard margin SVM):

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

and the residual grows at most as $\|\boldsymbol{\rho}(t)\| = O(\log\log(t))$, and so

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

Furthermore, for almost all data sets (all except measure zero), the residual $\rho(t)$ is bounded.

Proof Sketch We first understand intuitively why an exponential tail of the loss entail asymptotic convergence to the max margin vector: Assume for simplicity that $\ell(u) = e^{-u}$ exactly, and examine the asymptotic regime of gradient descent in which $\forall n : \mathbf{w}(t)^\top \mathbf{x}_n \to \infty$, as is guaranteed by Lemma 1. If $\mathbf{w}(t)/\|\mathbf{w}(t)\|$ converges to some limit $\mathbf{w}_\infty$, then we can write $\mathbf{w}(t) = g(t)\mathbf{w}_\infty + \boldsymbol{\rho}(t)$ such that $g(t) \to \infty$, $\forall n : \mathbf{x}_n^\top \mathbf{w}_\infty > 0$, and $\lim_{t\to\infty} \boldsymbol{\rho}(t)/g(t) = 0$. The gradient can then be written as:

$$-\nabla\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} \exp\left(-\mathbf{w}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n = \sum_{n=1}^{N} \exp\left(-g(t)\,\mathbf{w}_\infty^\top \mathbf{x}_n\right) \exp\left(-\boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n . \tag{5}$$

As $g(t) \to \infty$ and the exponents become more negative, only those samples with the largest (*i.e.*, least negative) exponents will contribute to the gradient. These are precisely the samples with the smallest margin $\operatorname{argmin}_n \mathbf{w}_\infty^\top \mathbf{x}_n$, aka the "support vectors". The negative gradient (eq. 5) would then asymptotically become a non-negative linear combination of support vectors. The limit $\mathbf{w}_\infty$ will then be dominated by these gradients, since any initial conditions become negligible as $\|\mathbf{w}(t)\| \to \infty$ (from Lemma 1). Therefore, $\mathbf{w}_\infty$ will also be a non-negative linear combination of support vectors, and so will its scaling $\hat{\mathbf{w}} = \mathbf{w}_\infty / \left(\min_n \mathbf{w}_\infty^\top \mathbf{x}_n\right)$. We therefore have:

$$\hat{\mathbf{w}} = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n \qquad \forall n \ \left(\alpha_n \geq 0 \text{ and } \hat{\mathbf{w}}^\top \mathbf{x}_n = 1\right) \text{ OR } \left(\alpha_n = 0 \text{ and } \hat{\mathbf{w}}^\top \mathbf{x}_n > 1\right) \tag{6}$$

These are precisely the KKT conditions for the SVM problem (eq. 4) and we can conclude that $\hat{\mathbf{w}}$ is indeed its solution and $\mathbf{w}_\infty$ is thus proportional to it.

To prove Theorem 3 rigorously, we need to show that $\mathbf{w}(t)/\|\mathbf{w}(t)\|$ has a limit, that $g(t) = \log(t)$ and to bound the effect of various residual errors, such as gradients of non-support vectors and the fact that the loss is only approximately exponential. To do so, we substitute eq. 3 into the gradient descent dynamics (eq. 2), with $\mathbf{w}_\infty = \hat{\mathbf{w}}$ being the max margin vector and $g(t) = \log t$. We then show that, except when certain degeneracies occur, the increment in the norm of $\boldsymbol{\rho}(t)$ is bounded by $C_1 t^{-\nu}$ for some $C_1 > 0$ and $\nu > 1$, which is a converging series. This happens because the increment in the max margin term, $\hat{\mathbf{w}}\left[\log(t+1) - \log(t)\right] \approx \hat{\mathbf{w}} t^{-1}$, cancels out the dominant t^{-1} term in the gradient $-\nabla\mathcal{L}(\mathbf{w}(t))$ (eq. 5 with $g(t) = \log(t)$ and $\mathbf{w}_\infty^\top \mathbf{x}_n = 1$).

Degenerate and Non-Degenerate Data Sets An earlier conference version of this paper (Soudry et al., 2018) included a partial version of Theorem 3, which only applies to almost all data sets, in which case we can ensure the residual $\rho(t)$ is bounded. This partial statement (for almost all data sets) is restated and proved as Theorem 9 in Appendix A. It applies, *e.g.* with probability one for data sampled from any absolutely continuous distribution. It does not apply in "degenerate" cases where some of the support vectors $\mathbf{x}_n$ (for which $\hat{\mathbf{w}}^\top \mathbf{x}_n = 1$) are associated with dual variables that are zero ($\alpha_n = 0$) in the dual optimum of 4. As we show in Appendix B, this only happens on measure zero data sets. Here, we prove the more general result which applies for all data sets, including degenerate data sets. To do so, in Theorem 13 in Appendix C we provide a more complete characterization of the iterates $\mathbf{w}(t)$ that explicitly specifies all unbounded components even in the degenerate case. We then prove the Theorem by plugging in this more complete characterization and showing that the residual is bounded, thus also establishing Theorem 3.

Parallel Work on the Degenerate Case Following the publication of our initial version, and while preparing this revised version for publication, we learned of parallel work by Ziwei Ji and Matus Telgarsky that also closes this gap. Ji and Telgarsky (2018) provide an analysis of the degenerate case, establishing converges to the max margin predictor by showing that $\left\| \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} - \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|} \right\| = O\left(\sqrt{\frac{\log\log t}{\log t}}\right)$. Our analysis provides a more precise characterization of the iterates, and also shows the convergence is actually quadratically faster (see Section 3). However, Ji and Telgarsky go even further and provide a characterization also when the data is non-separable but $\mathbf{w}(t)$ still goes to infinity.

More Refined Analysis of the Residual In some non-degenerate cases, we can further characterize the asymptotic behavior of $\boldsymbol{\rho}(t)$. To do so, we need to refer to the KKT conditions (eq. 6) of the SVM problem (eq. 4) and the associated support vectors $\mathcal{S} = \operatorname{argmin}_n \hat{\mathbf{w}}^\top \mathbf{x}_n$. We then have the following Theorem, proved in Appendix A:

Theorem 4 *Under the conditions and notation of Theorem 3, for almost all datasets, if in addition the support vectors span the data (*i.e. $\operatorname{rank}(\mathbf{X}_\mathcal{S}) = \operatorname{rank}(\mathbf{X})$*, where $\mathbf{X}_\mathcal{S}$ is a matrix whose columns are only those data points $\mathbf{x}_n$ s.t. $\hat{\mathbf{w}}^\top \mathbf{x}_n = 1$), then $\lim_{t\to\infty} \boldsymbol{\rho}(t) = \tilde{\mathbf{w}}$, where $\tilde{\mathbf{w}}$ is a solution to*

$$\forall n \in \mathcal{S}: \ \eta \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) = \alpha_n \tag{7}$$

Analogies with Boosting Perhaps most similar to our study is the line of work on understanding AdaBoost in terms its implicit bias toward large L_1-margin solutions, starting with the seminal work of Schapire et al. (1998). Since AdaBoost can be viewed as coordinate descent on the exponential loss of a linear model, these results can be interpreted as analyzing the bias of coordinate descent, rather then gradient descent, on a monotone decreasing loss with an exact exponential tail. Indeed, with small enough step sizes, such a coordinate descent procedure does converge precisely to the maximum L_1-margin solution (Zhang et al., 2005; Telgarsky, 2013). In fact, Telgarsky (2013) also generalizes these results to other losses with tight exponential tails, similar to the class of losses we consider here.

Also related is the work of Rosset et al. (2004). They considered the regularization path $\mathbf{w}_\lambda = \arg\min \mathcal{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|_p^p$ for similar loss functions as we do, and showed that $\lim_{\lambda\to 0} \mathbf{w}_\lambda / \|\mathbf{w}_\lambda\|_p$ is proportional to the maximum L_p margin solution. That is, they showed how adding infinitesimal L_p

(*e.g.* L_1 and L_2) regularization to logistic-type losses gives rise to the corresponding max-margin predictor.[3] However, Rosset et al. do not consider the effect of the optimization algorithm, and instead add explicit regularization. Here we are specifically interested in the bias implied by the algorithm *not* by adding (even infinitesimal) explicit regularization. We see that coordinate descent gives rise to the max L_1 margin predictor, while gradient descent gives rise to the max L_2 norm predictor. In Section 4.3 and in follow-up work (Gunasekar et al., 2018) we discuss also other optimization algorithms, and their implied biases.

Non-homogeneous linear predictors In this paper we focused on homogeneous linear predictors of the form $\mathbf{w}^\top \mathbf{x}$, similarly to previous works (*e.g.*, Rosset et al. (2004); Telgarsky (2013)). Specifically, we did not have the common intercept term: $\mathbf{w}^\top \mathbf{x} + b$. One may be tempted to introduce the intercept in the usual way, *i.e.*, by extending all the input vectors $\mathbf{x}_n$ with an additional $'1'$ component. In this extended input space, naturally, all our results hold. Therefore, we converge in direction to the L_2 max margin solution (eq. 4) in the extended space. However, if we translate this solution to the original $\mathbf{x}$ space we obtain

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

which is not the L_2 max margin (SVM) solution

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

where we do not have a b^2 penalty in the objective.

3. Implications: Rates of convergence

The solution in eq. 3 implies that $\mathbf{w}(t) / \|\mathbf{w}(t)\|$ converges to the normalized max margin vector $\hat{\mathbf{w}} / \|\hat{\mathbf{w}}\|$. Moreover, this convergence is very slow— logarithmic in the number of iterations. Specifically, our results imply the following tight rates of convergence:

Theorem 5 *Under the conditions and notation of Theorem 3, for any linearly separable data set, the normalized weight vector converges to the normalized max margin vector in L_2 norm*

$$\left\| \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} - \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|} \right\| = O\left(\frac{\log\log t}{\log t}\right), \tag{8}$$

with this rate improving to $O(1/\log(t))$ for almost every dataset; and in angle

$$1 - \frac{\mathbf{w}(t)^\top \hat{\mathbf{w}}}{\|\mathbf{w}(t)\| \|\hat{\mathbf{w}}\|} = O\left(\left(\frac{\log\log t}{\log t}\right)^2\right), \tag{9}$$

with this rate improving to $O(1/\log^2(t))$ for almost every dataset; and the margin converges as

$$\frac{1}{\|\hat{\mathbf{w}}\|} - \frac{\min_n \mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = O\left(\frac{1}{\log t}\right). \tag{10}$$

3. In contrast, with non-vanishing regularization (*i.e.*, $\lambda > 0$), $\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|_p^p$ is generally *not* a max margin solution.

On the other hand, the loss itself decreases as

$$\mathcal{L}\left(\mathbf{w}\left(t\right)\right) = O\left(\frac{1}{t}\right). \tag{11}$$

All the rates in the above Theorem are a direct consequence of Theorem 3, except for avoiding the $\log\log t$ factor for the degenerate cases in eq. 10 and eq. 11 (*i.e.*, establishing that the rates $1/\log t$ and $1/t$ always hold)—this additional improvement is a consequence of the more complete characterization of Theorem 13. Full details are provided in Appendix D. In this appendix, we also provide a simple construction showing all the rates in Theorem 5 are tight (except possibly for the $\log\log t$ factors).

The sharp contrast between the tight logarithmic and $1/t$ rates in Theorem 5 implies that the convergence of $\mathbf{w}(t)$ to the max-margin $\hat{\mathbf{w}}$ can be logarithmic in the loss itself, and we might need to wait until the loss is exponentially small in order to be close to the max-margin solution. This can help explain why continuing to optimize the training loss, even after the training error is zero and the training loss is extremely small, still improves generalization performance—our results suggests that the margin could still be improving significantly in this regime.

A numerical illustration of the convergence is depicted in Figure 1. As predicted by the theory, the norm $\|\mathbf{w}(t)\|$ grows logarithmically (note the semi-log scaling), and $\mathbf{w}(t)$ converges to the max-margin separator, but only logarithmically, while the loss itself decreases very rapidly (note the log-log scaling).

An important practical consequence of our theory, is that although the margin of $\mathbf{w}(t)$ keeps improving, and so we can expect the population (or test) misclassification error of $\mathbf{w}(t)$ to improve for many datasets, the same cannot be said about the expected population loss (or test loss)! At the limit, the direction of $\mathbf{w}(t)$ will converge toward the max margin predictor $\hat{\mathbf{w}}$. Although $\hat{\mathbf{w}}$ has zero training error, it will not generally have zero misclassification error on the population, or on a test or a validation set. Since the norm of $\mathbf{w}(t)$ will increase, if we use the logistic loss or any other convex loss, the loss incurred on those misclassified points will also increase. More formally, consider the logistic loss $\ell(u) = \log(1+e^{-u})$ and define also the hinge-at-zero loss $h(u) = \max(0,-u)$. Since $\hat{\mathbf{w}}$ classifies all training points correctly, we have that on the training set $\sum_{n=1}^{N} h(\hat{\mathbf{w}}^\top \mathbf{x}_n) = 0$. However, on the population we would expect some errors and so $\mathbb{E}[h(\hat{\mathbf{w}}^\top \mathbf{x})] > 0$. Since $\mathbf{w}(t) \approx \hat{\mathbf{w}} \log t$ and $\ell(\alpha u) \to \alpha h(u)$ as $\alpha \to \infty$, we have:

$$\mathbb{E}[\ell(\mathbf{w}(t)^\top \mathbf{x})] \approx \mathbb{E}[\ell((\log t)\hat{\mathbf{w}}^\top \mathbf{x})] \approx (\log t)\mathbb{E}[h(\hat{\mathbf{w}}^\top \mathbf{x})] = \Omega(\log t). \tag{12}$$

That is, the population loss increases logarithmically while the margin and the population misclassification error improve. Roughly speaking, the improvement in misclassification does not out-weight the increase in the loss of those points still misclassified.

The increase in the test loss is practically important because the loss on a validation set is frequently used to monitor progress and decide on stopping. Similar to the population loss, the validation loss will increase logarithmically with t, if there is at least one sample in the validation set which is classified incorrectly by the max margin vector (since we would not expect zero validation error). More precisely, as a direct consequence of Theorem 3 (as shown on Appendix D):

Corollary 6 *Let ℓ be the logistic loss, and $\mathcal{V}$ be an independent validation set, for which $\exists \mathbf{x} \in \mathcal{V}$ such that $\mathbf{x}^\top \hat{\mathbf{w}} < 0$. Then the validation loss increases as*

$$\mathcal{L}_{\text{val}}\left(\mathbf{w}\left(t\right)\right) = \sum_{\mathbf{x}\in\mathcal{V}} \ell\left(\mathbf{w}\left(t\right)^\top \mathbf{x}\right) = \Omega(\log(t)).$$

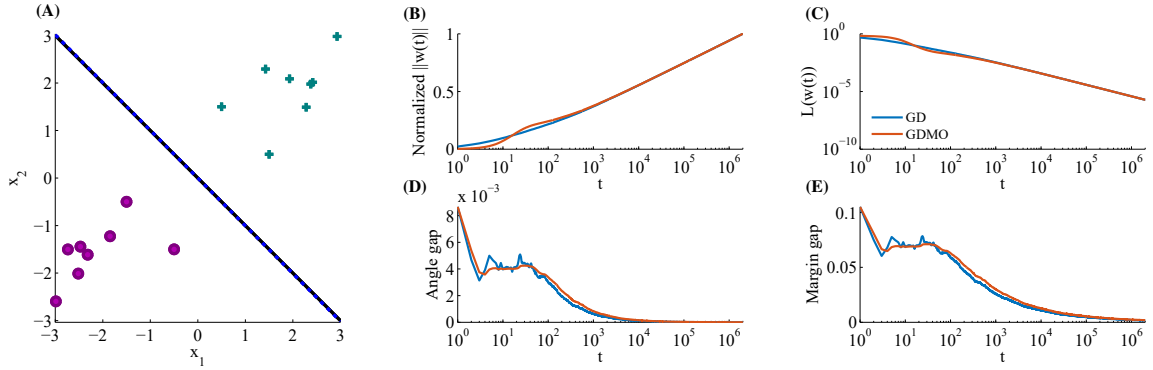

Figure 1: Visualization of or main results on a synthetic dataset in which the L_2 max margin vector $\hat{\mathbf{w}}$ is precisely known. **(A)** The dataset (positive and negatives samples ($y = \pm 1$) are respectively denoted by $'+'$ and $'\circ'$), max margin separating hyperplane (black line), and the asymptotic solution of GD (dashed blue). For both GD and GD with momentum (GDMO), we show: **(B)** The norm of $\mathbf{w}(t)$, normalized so it would equal to 1 at the last iteration, to facilitate comparison. As expected (eq. 3), the norm increases logarithmically; **(C)** the training loss. As expected, it decreases as t^{-1} (eq. 11); and **(D&E)** the angle and margin gap of $\mathbf{w}(t)$ from $\hat{\mathbf{w}}$ (eqs. 9 and 10). As expected, these are logarithmically decreasing to zero. **Implementation details:** The dataset includes 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 is then $\hat{\mathbf{w}} = (1, 1)/\sqrt{2}$ with margin equal to $\sqrt{2}$), and 12 other random datapoints (6 from each class), that are not on the margin. We used a learning rate $\eta = 1/\sigma^2_{\max}(\mathbf{X})$, where $\sigma^2_{\max}(\mathbf{X})$ is the maximal singular value of $\mathbf{X}$, momentum $\gamma = 0.9$ for GDMO, and initialized at the origin.

This behavior might cause us to think we are over-fitting or otherwise encourage us to stop the optimization. However, this increase does not actually represent the model getting worse, merely $\|\mathbf{w}(t)\|$ getting larger, and in fact the model might be getting better (increasing the margin and possibly decreasing the error rate).

4. Extensions

4.1 Multi-Class Classification with Cross-Entropy Loss

So far, we have discussed the problem of binary classification, but in many practical situations we have more then two classes. For multi-class problems, the labels are the class indices $y_n \in [K] \triangleq \{1, \ldots, K\}$ and we learn a predictor $\mathbf{w}_k$ for each class $k \in [K]$. A common loss function in multi-class classification is the following cross-entropy loss with a softmax output, which is a generalization of the logistic loss:

$$\mathcal{L}\left(\{\mathbf{w}_k\}_{k\in[K]}\right) = -\sum_{n=1}^{N} \log\left(\frac{\exp\left(\mathbf{w}_{y_n}^\top \mathbf{x}_n\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{w}_k^\top \mathbf{x}_n\right)}\right) \tag{13}$$

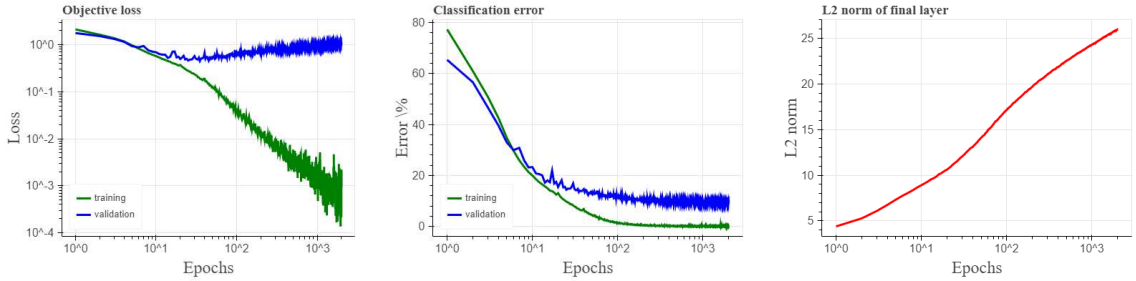

Figure 2: Training of a convolutional neural network on CIFAR10 using stochastic gradient descent with constant learning rate and momentum, softmax output and a cross entropy loss, where we achieve 8.3% final validation error. We observe that, approximately: (1) The training loss decays as a t^{-1}, (2) the L_2 norm of last weight layer increases logarithmically, (3) after a while, the validation loss starts to increase, and (4) in contrast, the validation (classification) error slowly improves.

What do the linear predictors $\mathbf{w}_k(t)$ converge to if we minimize the cross-entropy loss by gradient descent on the predictors? In Appendix E we analyze this problem for separable data, and show that again, the predictors diverge to infinity and the loss converges to zero. Furthermore, we prove the following Theorem:

Theorem 7 *For almost all multiclass datasets (*i.e.*, except for a measure zero) which are linearly separable (i.e. the constraints in eq. 15 below are feasible), any starting point* $\mathbf{w}(0)$ *and any small enough stepsize, the iterates of gradient descent on 13 will behave as:*

$$\mathbf{w}_k(t) = \hat{\mathbf{w}}_k \log(t) + \boldsymbol{\rho}_k(t), \tag{14}$$

where the residual $\boldsymbol{\rho}_k(t)$ *is bounded and* $\hat{\mathbf{w}}_k$ *is the solution of the K-class SVM:*

$$\text{argmin}_{\mathbf{w}_1,\ldots,\mathbf{w}_k} \sum_{k=1}^{K} ||\mathbf{w}_k||^2 \; s.t. \; \forall n, \forall k \neq y_n : \mathbf{w}_{y_n}^\top \mathbf{x}_n \geq \mathbf{w}_k^\top \mathbf{x}_n + 1. \tag{15}$$

4.2 Deep networks

So far we have only considered linear prediction. Naturally, it is desirable to generalize our results also to non-linear models and especially multi-layer neural networks.

Even without a formal extension and description of the precise bias, our results already shed light on how minimizing the cross-entropy loss with gradient descent can have a margin maximizing effect, how the margin might improve only logarithmically slow, and why it might continue to improve even as the validation loss increases. These effects are demonstrated in Figure 2 and Table 1 which portray typical training of a convolutional neural network using unregularized gradient descent[4]. As can be seen, the norm of the weight increases, but the validation error continues decreasing, albeit very slowly (as predicted by the theory), even after the training error is zero and the training loss is extremely small. We can now understand how even though the loss is already

4. Code available here: `https://github.com/paper-submissions/MaxMargin`

Epoch	50	100	200	400	2000	4000
L_2 norm	13.6	16.5	19.6	20.3	25.9	27.54
Train loss	0.1	0.03	0.02	0.002	10^{-4}	$3 \cdot 10^{-5}$
Train error	4%	1.2%	0.6%	0.07%	0%	0%
Validation loss	0.52	0.55	0.77	0.77	1.01	1.18
Validation error	12.4%	10.4%	11.1%	9.1%	8.92%	8.9%

Table 1: Sample values from various epochs in the experiment depicted in Fig. 2.

extremely small, some sort of margin might be gradually improving as we continue optimizing. We can also observe how the validation loss increases despite the validation error decreasing, as discussed in Section 3.

As an initial advance toward tackling deep network, we can point out that for several special cases, our results may be directly applied to multi-layered networks. First, somewhat trivially, our results may be applied directly to the last weight layer of a neural network if the last hidden layer becomes fixed and linearly separable after a certain number of iterations. This can become true, either approximately, if the input to the last hidden layer is normalized (*e.g.*, using batch norm), or exactly, if the last hidden layer is quantized (Hubara et al., 2018).

Second, as we show next, our results may be applied exactly on deep networks if only a single weight layer is being optimized, and, furthermore, after a sufficient number of iterations, the activation units stop switching and the training error goes to zero.

Corollary 8 *We examine a multilayer neural network with component-wise ReLU functions* $f(z) = \max[z, 0]$*, and weights* $\{\mathbf{W}_l\}_{l=1}^{L}$*. Given input* $\mathbf{x}_n$ *and target* $y_n \in \{-1, 1\}$*, the DNN produces a scalar output*

$$u_n = \mathbf{W}_L f\left(\mathbf{W}_{L-1} f\left(\cdots \mathbf{W}_2 f\left(\mathbf{W}_1 \mathbf{x}_n\right)\right)\right)$$

and has loss $\ell(y_n u_n)$*, where* ℓ *obeys assumptions 2 and 3.*

If we optimize a single weight layer $\mathbf{w}_l = \mathrm{vec}\left(\mathbf{W}_l^\top\right)$ *using gradient descent, so that* $\mathcal{L}(\mathbf{w}_l) = \sum_{n=1}^{N} \ell(y_n u_n(\mathbf{w}_l))$ *converges to zero, and* $\exists t_0$ *such that* $\forall t > t_0$ *the ReLU inputs do not switch signs, then* $\mathbf{w}_l(t) / \|\mathbf{w}_l(t)\|$ *converges to*

$$\hat{\mathbf{w}}_l = \underset{\mathbf{w}_l}{\operatorname{argmin}} \|\mathbf{w}_l\|^2 \ \text{s.t.}\ y_n u_n(\mathbf{w}_l) \geq 1.$$

Proof We examine the output of the network given a single input $\mathbf{x}_n$, for $t > t_0$**.** Since the ReLU inputs do not switch signs, we can write $\mathbf{v}_l$, the output of layer l, as

$$\mathbf{v}_{l,n} = \prod_{m=1}^{l} \mathbf{A}_{m,n} \mathbf{W}_m \mathbf{x}_n \,,$$

where we defined $\mathbf{A}_{l,n}$ for $l < L$ as a diagonal 0-1 matrix, which diagonal is the ReLU slopes at layer l, sample n, and $\mathbf{A}_{L,n} = 1$. Additionally, we define

$$\boldsymbol{\delta}_{l,n} = \mathbf{A}_{l,n} \prod_{m=L}^{l+1} \mathbf{W}_m^\top \mathbf{A}_{m,n} \,;\, \tilde{\mathbf{x}}_{l,n} = \boldsymbol{\delta}_{l,n} \otimes \mathbf{u}_{l-1,n} \,.$$

Using this notation we can write

$$u_n(\mathbf{w}_l) = v_{L,n} = \prod_{m=1}^{L} \mathbf{A}_{m,n}\mathbf{W}_m\mathbf{x}_n = \boldsymbol{\delta}_{l,n}^\top \mathbf{W}_l \mathbf{u}_{l-1,n} = \tilde{\mathbf{x}}_{l,n}^\top \mathbf{w}_l \,. \tag{16}$$

This implies that

$$\mathcal{L}(\mathbf{w}_l) = \sum_{n=1}^{N} \ell\left(y_n u_n(\mathbf{w}_l)\right) = \sum_{n=1}^{N} \ell\left(y_n \tilde{\mathbf{x}}_{l,n}^\top \mathbf{w}_l\right),$$

which is the same as the original linear problem. Since the loss converges to zero, the dataset $\{\tilde{\mathbf{x}}_{l,n}, y_n\}_{n=1}^N$ must be linearly separable. Applying Theorem 3, and recalling that $u(\mathbf{w}_l) = \tilde{\mathbf{x}}_l^\top \mathbf{w}_l$ from eq. 16, we prove this corollary. ■

Importantly, this case is non-convex, unless we are optimizing the last layer. Note we assumed ReLU functions for simplicity, but this proof can be easily generalized for any other piecewise linear constant activation functions (*e.g.*, leaky ReLU, max-pooling).

Lastly, in a follow-up work (Gunasekar et al., 2018b), given a few additional assumptions, extended our results to linear predictors which can be written as a homogeneous polynomial in the parameters. These results seem to indicate that, in many cases, GD operating on exp-tailed loss with positively homogeneous predictors aims to a specific direction. This is the direction of the max margin predictor minimizing the L_2 norm in the parameter space. It is not yet clear how to generally translate such an implicit bias in the parameter space to the implicit bias in the predictor space — except in special cases, such as deep linear neural nets, as we have shown in (Gunasekar et al., 2018b). Moreover, in non-linear neural nets, there are many equivalent max-margin solutions which minimize the L_2 norm of the parameters. Therefore, it is natural to expect that GD would have additional implicit biases, which select a specific subset of these solutions.

4.3 Other optimization methods

In this paper we examined the implicit bias of gradient descent. Different optimization algorithms exhibit different biases, and understanding these biases and how they differ is crucial to understanding and constructing learning methods attuned to the inductive biases we expect. Can we characterize the implicit bias and convergence rate in other optimization methods?

In Figure 1 we see that adding momentum does not qualitatively affect the bias induced by gradient descent. In Figure 4 in Appendix F we also repeat the experiment using stochastic gradient descent, and observe a similar asymptotic bias (this was later proved in Nacson et al. (2018)). This is consistent with the fact that momentum, acceleration and stochasticity do not change the bias when using gradient descent to optimize an under determined least squares problem. It would be beneficial, though, to rigorously understand how much we can generalize our result to gradient descent variants, and how the convergence rates might change in these cases.

On the other hand, as an example of how changing the optimization algorithm does change the bias, consider adaptive methods, such as AdaGrad (Duchi et al., 2011) and ADAM (Kingma and Ba, 2015). In Figure 3 we show the predictors obtained by ADAM and by gradient descent on a simple data set. Both methods converge to zero training error solutions. But although gradient descent converges to the L_2 max margin predictor, as predicted by our theory, ADAM does not. The implicit bias of adaptive methods has in fact been a recent topic of interest, with Hoffer et al. (2017) and

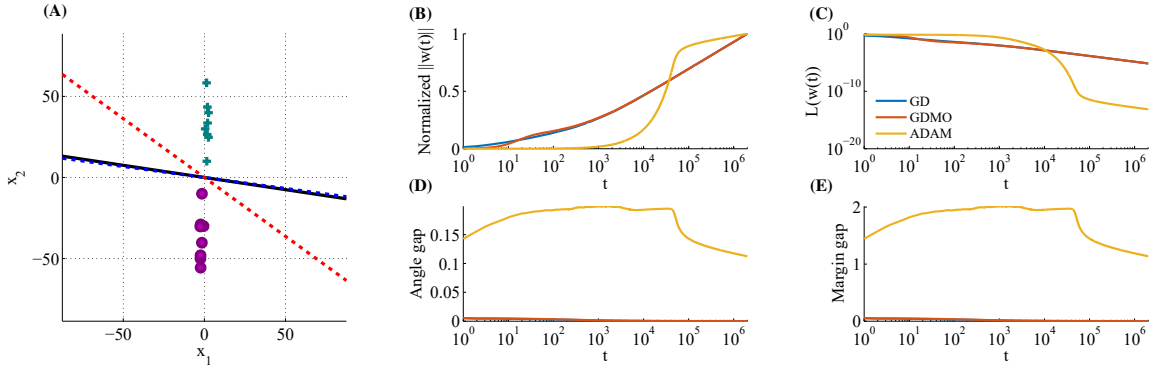

Figure 3: Same as Fig. 1, except we multiplied all x_2 values in the dastaset by 20, and also train using ADAM. The final weight vector produced after $2 \cdot 10^6$ epochs of optimization using ADAM (red dashed line) does not converge to L2 max margin solution (black line), in contrast to GD (blue dashed line), or GDMO.

Wilson et al. (2017) suggesting they lead to worse generalization, and Wilson et al. (2017) providing examples of the differences in the bias for linear regression problems with the squared loss. Can we characterize the bias of adaptive methods for logistic regression problems? Can we characterize the bias of other optimization methods, providing a general understanding linking optimization algorithms with their biases?

In a follow-up paper (Gunasekar et al., 2018) provided initial answers to these questions. Gunasekar et al. (2018) derived a precise characterization of the limit direction of steepest descent for general norms when optimizing the exp-loss, and show that for adaptive methods such as Adagrad the limit direction can depend on the initial point and step size and is thus not as predictable and robust as with non-adaptive methods.

4.4 Other loss functions

In this work we focused on loss functions with exponential tail and observed a very slow, logarithmic convergence of the normalized weight vector to the L_2 max margin direction. A natural question that follows is how does this behavior change with types of loss function tails. Specifically, does the normalized weight vector always converge to the L_2 max margin solution? How is the convergence rate affected? Can we improve the convergence rate beyond the logarithmic rate found in this work?

In a follow-up work Nacson et al. (2018) provided partial answers to these questions. They proved that the exponential tail has the optimal convergence rate, for tails for which $\ell'(u)$ is of the form $\exp(-u^\nu)$ with $\nu > 0.25$. They then conjectured, based on heuristic analysis, that the exponential tail is optimal among all possible tails. Furthermore, they demonstrated that polynomial or heavier tails do not converge to the max margin solution. Lastly, for the exponential loss they proposed a normalized gradient scheme which can significantly improve convergence rate, achieving $O(\log(t)/\sqrt{t})$.

4.5 Matrix Factorization

With multi-layered neural networks in mind, Gunasekar et al. (2017) recently embarked on a study of the implicit bias of under-determined matrix factorization problems, where the *squared loss* of the linear observation of a matrix is minimized by gradient descent on its factorization. Since a matrix factorization can be viewed as a two-layer network with linear activations, this is perhaps the simplest deep model one can study in full, and can thus provide insight and direction to studying more complex neural networks. Gunasekar et al. conjectured, and provided theoretical and empirical evidence, that gradient descent on the factorization for an under-determined problem converges to the minimum nuclear norm solution, but only if the initialization is infinitesimally close to zero and the step-sizes are infinitesimally small. With finite step-sizes or finite initialization, Gunasekar et al. could not characterize the bias.

The follow-up paper (Gunasekar et al., 2018) studied this same problem with exponential loss instead of squared loss. Under additional assumptions on the asymptotic convergence of update directions and gradient directions, they were able to relate the direction of gradient descent iterates on the factorized parameterization asymptotically to the maximum margin solution with unit nuclear norm. Unlike the case of squared loss, the result for exponential loss are independent of initialization and with only mild conditions on the step size. Here again, we see the asymptotic nature of exponential loss on separable data nullifying the initialization effects thereby making the analysis simpler compared to squared loss.

5. Summary

We characterized the implicit bias induced by gradient descent on homogeneous linear predictors when minimizing smooth monotone loss functions with an exponential tail. This is the type of loss commonly being minimized in deep learning. We can now rigorously understand:

1. How gradient descent, without early stopping, induces implicit L_2 regularization and converges to the maximum L_2 margin solution, when minimizing for binary classification with logistic loss, exp-loss, or other exponential tailed monotone decreasing loss, as well as for multi-class classification with cross-entropy loss. Notably, even though the logistic loss and the exp-loss behave very different on non-separable problems, they exhibit the same behaviour for separable problems. This implies that the non-tail part does not affect the bias. The bias is also independent of the step-size used (as long as it is small enough to ensure convergence) and is also independent on the initialization (unlike for least square problems).

2. The convergence of the direction of gradient descent updates to the maximum L_2 margin solution, however is very slow compared to the convergence of training loss, which explains why it is worthwhile continuing to optimize long after we have zero training error, and even when the loss itself is already extremely small.

3. We should not rely on plateauing of the training loss or on the loss (logistic or exp or cross-entropy) evaluated on a validation data, as measures to decide when to stop. Instead, we should look at the 0–1 error on the validation dataset. We might improve the validation and test errors even when when the decrease in the training loss is tiny and even when the validation loss itself increases.

Perhaps that gradient descent leads to a max L_2 margin solution is not a big surprise to those for whom the connection between L_2 regularization and gradient descent is natural. Nevertheless, we are not familiar with any prior study or mention of this fact, let alone a rigorous analysis and study of how this bias is exact and independent of the initial point and the step-size. Furthermore, we also analyze the rate at which this happens, leading to the novel observations discussed above. Even more importantly, we hope that our analysis can open the door to further analysis of different optimization methods or in different models, including deep networks, where implicit regularization is not well understood even for least square problems, or where we do not have such a natural guess as for gradient descent on linear problems. Analyzing gradient descent on logistic/cross-entropy loss is not only arguably more relevant than the least square loss, but might also be technically easier.

Acknowledgments

The authors are grateful to J. Lee, and C. Zeno for helpful comments on the manuscript. The research of DS was supported by the Israel Science Foundation (grant No. 31/1031), by the Taub foundation and of NS by the National Science Foundation.

Appendix

Appendix A. Proof of Theorems 3 and 4 for almost every dataset

In the following sub-sections we first prove Theorem 9 below, which is a version of Theorem 3, specialized for almost every dataset. We then prove Theorem 4 (which is already stated for almost every dataset).

Theorem 9 *For almost every dataset which is linearly separable (Assumption 1), any β-smooth decreasing loss function (Assumption 2) with an exponential tail (Assumption 3), any stepsize $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathbf{X})$ and any starting point $\mathbf{w}(0)$, the gradient descent iterates (as in eq. 2) will behave as:*

$$\mathbf{w}(t) = \hat{\mathbf{w}} \log t + \boldsymbol{\rho}(t) \,, \tag{17}$$

where $\hat{\mathbf{w}}$ is the L_2 max margin vector

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

the residual $\rho(t)$ is bounded, and so

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

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

$$\mathbf{r}(t) = \mathbf{w}(t) - \hat{\mathbf{w}} \log t - \tilde{\mathbf{w}},$$

where $\hat{\mathbf{w}}$ and $\tilde{\mathbf{w}}$ follow the conditions of Theorems 3 and 4, *i.e.* $\hat{\mathbf{w}}$ is the L_2 is the max margin vector defined above, and $\tilde{\mathbf{w}}$ is a vector which satisfies eq. 7:

$$\forall n \in \mathcal{S}: \; \eta \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) = \alpha_n \,, \tag{18}$$

where we recall that we denoted $\mathbf{X}_\mathcal{S} \in \mathbb{R}^{d\times|\mathcal{S}|}$ as the matrix whose columns are the support vectors, a subset $\mathcal{S} \subset \{1,\ldots,N\}$ of the columns of $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_N] \in \mathbb{R}^{d\times N}$.

In Lemma 12 (Appendix B) we prove that for almost every dataset $\boldsymbol{\alpha}$ is uniquely defined, there are no more then d support vectors and $\alpha_n \neq 0, \forall n \in \mathcal{S}$. Therefore, eq. 18 is well-defined in those cases. If the support vectors do not span the data, then the solution $\tilde{\mathbf{w}}$ to eq. 18 might not be unique. In this case, we can use any such solution in the proof.

We furthermore denote the minimum margin to a non-support vector as:

$$\theta = \min_{n\notin\mathcal{S}} \mathbf{x}_n^\top \hat{\mathbf{w}} > 1 \,, \tag{19}$$

and by C_i,ϵ_i,t_i ($i \in \mathbb{N}$) various positive constants which are independent of t. Lastly, we define $\mathbf{P}_1 \in \mathbb{R}^{d\times d}$ as the orthogonal projection matrix[5] to the subspace spanned by the support vectors (the columns of $\mathbf{X}_\mathcal{S}$), and $\bar{\mathbf{P}}_1 = \mathbf{I} - \mathbf{P}_1$ as the complementary projection (to the left nullspace of $\mathbf{X}_\mathcal{S}$).

5. This matrix can be written as $\mathbf{P}_1 = \mathbf{X}_\mathcal{S}\mathbf{X}_\mathcal{S}^+$, where $\mathbf{M}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{M}$.

A.1 Simple proof of Theorem 9

In this section we first examine the special case that $\ell(u) = e^{-u}$ and take the continuous time limit of gradient descent: $\eta \to 0$, so

$$\dot{\mathbf{w}}(t) = -\nabla \mathcal{L}(\mathbf{w}(t)) .$$

The proof in this case is rather short and self-contained (*i.e.*, does not rely on any previous results), and so it helps to clarify the main ideas of the general (more complicated) proof which we will give in the next sections.

Recall we defined

$$\mathbf{r}(t) = \mathbf{w}(t) - \log(t)\,\hat{\mathbf{w}} - \tilde{\mathbf{w}} . \tag{20}$$

Our goal is to show that $\|\mathbf{r}(t)\|$ is bounded, and therefore $\boldsymbol{\rho}(t) = \mathbf{r}(t) + \tilde{\mathbf{w}}$ is bounded. Eq. 20 implies that

$$\dot{\mathbf{r}}(t) = \dot{\mathbf{w}}(t) - \frac{1}{t}\hat{\mathbf{w}} = -\nabla \mathcal{L}(\mathbf{w}(t)) - \frac{1}{t}\hat{\mathbf{w}} \tag{21}$$

and therefore

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\mathbf{r}(t)\|^2 = \dot{\mathbf{r}}^\top(t)\,\mathbf{r}(t) \\
&= \sum_{n=1}^{N} \exp\left(-\mathbf{x}_n^\top \mathbf{w}(t)\right) \mathbf{x}_n^\top \mathbf{r}(t) - \frac{1}{t}\hat{\mathbf{w}}^\top \mathbf{r}(t) \\
&= \left[\sum_{n\in\mathcal{S}} \exp\left(-\log(t)\,\hat{\mathbf{w}}^\top \mathbf{x}_n - \tilde{\mathbf{w}}^\top \mathbf{x}_n - \mathbf{x}_n^\top \mathbf{r}(t)\right) \mathbf{x}_n^\top \mathbf{r}(t) - \frac{1}{t}\hat{\mathbf{w}}^\top \mathbf{r}(t)\right] \\
&+ \left[\sum_{n\notin\mathcal{S}} \exp\left(-\log(t)\,\hat{\mathbf{w}}^\top \mathbf{x}_n - \tilde{\mathbf{w}}^\top \mathbf{x}_n - \mathbf{x}_n^\top \mathbf{r}(t)\right) \mathbf{x}_n^\top \mathbf{r}(t)\right],
\end{aligned} \tag{22}$$

where in the last equality we used eq. 20 and decomposed the sum over support vectors $\mathcal{S}$ and non-support vectors. We examine both bracketed terms. Recall that $\hat{\mathbf{w}}^\top \mathbf{x}_n = 1$ for $n \in \mathcal{S}$, and that we defined (in eq. 18) $\tilde{\mathbf{w}}$ so that $\sum_{n\in\mathcal{S}} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{x}_n = \hat{\mathbf{w}}$. Thus, the first bracketed term in eq. 22 can be written as

$$\begin{aligned}
&\frac{1}{t}\sum_{n\in\mathcal{S}} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n - \mathbf{x}_n^\top \mathbf{r}(t)\right) \mathbf{x}_n^\top \mathbf{r}(t) - \frac{1}{t}\sum_{n\in\mathcal{S}} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{x}_n^\top \mathbf{r}(t) \\
=&\frac{1}{t}\sum_{n\in\mathcal{S}} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \left(\exp\left(-\mathbf{x}_n^\top \mathbf{r}(t)\right) - 1\right) \mathbf{x}_n^\top \mathbf{r}(t) \le 0,
\end{aligned} \tag{23}$$

since $\forall z,\ z(e^{-z} - 1) \le 0$. Furthermore, since $\forall z\ e^{-z} z \le 1$ and $\theta = \operatorname{argmin}_{n\notin\mathcal{S}} \mathbf{x}_n^\top \hat{\mathbf{w}} > 1$ (eq. 19), the second bracketed term in eq. 22 can be upper bounded by

$$\sum_{n\notin\mathcal{S}} \exp\left(-\log(t)\,\hat{\mathbf{w}}^\top \mathbf{x}_n - \tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \exp\left(-\mathbf{x}_n^\top \mathbf{r}(t)\right) \mathbf{x}_n^\top \mathbf{r}(t) \le \frac{1}{t^\theta}\sum_{n\notin\mathcal{S}} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) . \tag{24}$$

Substituting eq. 23 and 24 into eq. 22 and integrating, we obtain, that $\exists C, C'$ such that

$$\forall t_1, \forall t > t_1 : \|\mathbf{r}(t)\|^2 - ||\mathbf{r}(t_1)||^2 \le C\int_{t_1}^{t} \frac{dt}{t^\theta} \le C' < \infty ,$$

since $\theta > 1$ (eq. 19). Thus, we showed that $\mathbf{r}(t)$ is bounded, which completes the proof for the special case. ■

A.2 Complete proof of Theorem 9

Next, we give the proof for the general case (non-infinitesimal step size, and exponentially-tailed functions). Though it is based on a similar analysis as in the special case we examined in the previous section, it is somewhat more involved since we have to bound additional terms.

First, we state two auxiliary lemmata, that are proven below in appendix sections A.4 and A.5:

Lemma 10 *Let $\mathcal{L}(\mathbf{w})$ be a β-smooth non-negative objective. If $\eta < 2\beta^{-1}$, then, for any $\mathbf{w}(0)$, with the GD sequence*

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla \mathcal{L}(\mathbf{w}(t)) \tag{25}$$

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

Lemma 11 *We have*

$$\exists C_1, t_1 : \forall t > t_1 : (\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-\min(\theta, 1+1.5\mu_+, 1+0.5\mu_-)}. \tag{26}$$

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

$$\|\mathbf{P}_1 \mathbf{r}(t)\| \geq \epsilon_1, \tag{27}$$

then the following improved bound holds

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq -C_2 t^{-1} < 0. \tag{28}$$

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

$$\|\mathbf{r}(t+1)\|^2 = \|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 + 2(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) + \|\mathbf{r}(t)\|^2 \tag{29}$$

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

$$\begin{aligned}
&\|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 \\
&\overset{(1)}{=} \|\mathbf{w}(t+1) - \hat{\mathbf{w}} \log(t+1) - \tilde{\mathbf{w}} - \mathbf{w}(t) + \hat{\mathbf{w}} \log(t) + \tilde{\mathbf{w}}\|^2 \\
&\overset{(2)}{=} \|-\eta \nabla \mathcal{L}(\mathbf{w}(t)) - \hat{\mathbf{w}} [\log(t+1) - \log(t)]\|^2 \\
&= \eta^2 \|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 + \|\hat{\mathbf{w}}\|^2 \log^2\left(1 + t^{-1}\right) + 2\eta \hat{\mathbf{w}}^\top \nabla \mathcal{L}(\mathbf{w}(t)) \log\left(1 + t^{-1}\right) \\
&\overset{(3)}{\leq} \eta^2 \|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 + \|\hat{\mathbf{w}}\|^2 t^{-2}
\end{aligned} \tag{30}$$

where in (1) we used eq. 20, in (2) we used eq. 2, and in (3) we used $\forall x > 0 : x \geq \log(1+x) > 0$, and also that

$$\hat{\mathbf{w}}^\top \nabla \mathcal{L}(\mathbf{w}(t)) = \sum_{n=1}^{N} \ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n\right) \hat{\mathbf{w}}^\top \mathbf{x}_n \leq 0, \tag{31}$$

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 10 we know that

$$\|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 = o(1) \text{ and } \sum_{t=0}^{\infty} \|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 < \infty . \tag{32}$$

Substituting eq. 32 into eq. 30, and recalling that a $t^{-\nu}$ power series converges for any $\nu > 1$, we can find C_0 such that

$$\|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 = o(1) \text{ and } \sum_{t=0}^{\infty} \|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 = C_0 < \infty . \tag{33}$$

Note that this equation also implies that $\forall \epsilon_0$

$$\exists t_0 : \forall t > t_0 : |\|\mathbf{r}(t+1)\| - \|\mathbf{r}(t)\|| < \epsilon_0 . \tag{34}$$

Next, we would like to bound the second term in eq. 29. From eq. 26 in Lemma 11, we can find t_1, C_1 such that $\forall t > t_1$:

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-\min(\theta, 1+1.5\mu_+, 1+0.5\mu_-)} . \tag{35}$$

Thus, by combining eqs. 35 and 33 into eq. 29, we find

$$\begin{aligned}
&\|\mathbf{r}(t)\|^2 - \|\mathbf{r}(t_1)\|^2 \\
&= \sum_{u=t_1}^{t-1} \left[\|\mathbf{r}(u+1)\|^2 - \|\mathbf{r}(u)\|^2 \right] \\
&\leq C_0 + 2 \sum_{u=t_1}^{t-1} C_1 u^{-\min(\theta, 1+1.5\mu_+, 1+0.5\mu_-)}
\end{aligned}$$

which is a bounded, since $\theta > 1$ (eq. 19) and $\mu_-, \mu_+ > 0$ (Definition 2). Therefore, $\|\mathbf{r}(t)\|$ is bounded. ■

A.3 Proof of Theorem 4

All that remains now is to show that $\|\mathbf{r}(t)\| \to 0$ if $\operatorname{rank}(\mathbf{X}_\mathcal{S}) = \operatorname{rank}(\mathbf{X})$, and that $\tilde{\mathbf{w}}$ is unique given $\mathbf{w}(0)$. To do so, this proof will continue where the proof of Theorem 3 stopped, using notations and equations from that proof.

Since $\mathbf{r}(t)$ has a bounded norm, its two orthogonal components $\mathbf{r}(t) = \mathbf{P}_1 \mathbf{r}(t) + \bar{\mathbf{P}}_1 \mathbf{r}(t)$ also have bounded norms (recall that $\mathbf{P}_1, \bar{\mathbf{P}}_1$ were defined in the beginning of appendix section A). From eq. 2, $\nabla \mathcal{L}(\mathbf{w})$ is spanned by the columns of $\mathbf{X}$. If $\operatorname{rank}(\mathbf{X}_\mathcal{S}) = \operatorname{rank}(\mathbf{X})$, then it is also spanned by the columns of $\mathbf{X}_\mathcal{S}$, and so $\bar{\mathbf{P}}_1 \nabla \mathcal{L}(\mathbf{w}) = 0$. Therefore, $\bar{\mathbf{P}}_1 \mathbf{r}(t)$ is not updated during GD, and remains constant. Since $\tilde{\mathbf{w}}$ in eq. 20 is also bounded, we can absorb this constant $\bar{\mathbf{P}}_1 \mathbf{r}(t)$ into $\tilde{\mathbf{w}}$ without affecting eq. 7 (since $\forall n \in \mathcal{S} : \mathbf{x}_n^\top \bar{\mathbf{P}}_1 \mathbf{r}(t) = 0$). Thus, without loss of generality, we can assume that $\mathbf{r}(t) = \mathbf{P}_1 \mathbf{r}(t)$.

We define the set

$$\mathcal{T} = \{t > \max[t_2, t_0] : \|\mathbf{r}(t)\| < \epsilon_1\} .$$

By contradiction, we assume that the complementary set is not finite,

$$\bar{\mathcal{T}} = \{t > \max[t_2, t_0] : \|\mathbf{r}(t)\| \geq \epsilon_1\} .$$

Additionally, the set $\mathcal{T}$ is not finite: if it were finite, it would have had a finite maximal point $t_{\max} \in \mathcal{T}$, and then, combining eqs. 28, 29, and 33, we would find that $\forall t > t_{\max}$

$$\|\mathbf{r}(t)\|^2 - \|\mathbf{r}(t_{\max})\|^2 = \sum_{u=t_{\max}}^{t-1} \left[\|\mathbf{r}(u+1)\|^2 - \|\mathbf{r}(u)\|^2\right] \leq C_0 - 2C_2 \sum_{u=t_{\max}}^{t-1} u^{-1} \to -\infty ,$$

which is impossible since $\|\mathbf{r}(t)\|^2 \geq 0$. Furthermore, eq. 33 implies that

$$\sum_{u=0}^{t} \|\mathbf{r}(u+1) - \mathbf{r}(t)\|^2 = C_0 - h(t)$$

where $h(t)$ is a positive monotone function decreasing to zero. Let t_3, t be any two points such that $t_3 < t$, $\{t_3, t_3+1, \dots t\} \subset \bar{\mathcal{T}}$, and $(t_3 - 1) \in \mathcal{T}$. For all such t_3 and t, we have

$$\begin{aligned}
\|\mathbf{r}(t)\|^2 &\leq \|\mathbf{r}(t_3)\|^2 + \sum_{u=t_3}^{t-1} \left[\|\mathbf{r}(u+1)\|^2 - \|\mathbf{r}(u)\|^2\right] \\
&= \|\mathbf{r}(t_3)\|^2 + \sum_{u=t_3}^{t-1} \left[\|\mathbf{r}(u+1) - \mathbf{r}(u)\|^2 + 2(\mathbf{r}(u+1) - \mathbf{r}(u))^\top \mathbf{r}(u)\right] \\
&\leq \|\mathbf{r}(t_3)\|^2 + h(t_3) - h(t-1) - 2C_2 \sum_{u=t_3}^{t-1} u^{-1} \\
&\leq \|\mathbf{r}(t_3)\|^2 + h(t_3) . \qquad (36)
\end{aligned}$$

Also, recall that $t_3 > t_0$, so from eq. 34, we have that $|\|\mathbf{r}(t_3)\| - \|\mathbf{r}(t_3 - 1)\|| < \epsilon_0$. Since $\|\mathbf{r}(t_3 - 1)\| < \epsilon_1$ (from $\mathcal{T}$ definition), we conclude that $\|\mathbf{r}(t_3)\| \leq \epsilon_1 + \epsilon_0$. Moreover, since $\bar{\mathcal{T}}$ is an infinite set, we can choose t_3 as large as we want. This implies that $\forall \epsilon_2 > 0$ we can find t_3 such that $\epsilon_2 > h(t_3)$, since $h(t)$ is a monotonically decreasing function. Therefore, from eq. 36, $\forall \epsilon_1, \epsilon_0, \epsilon_2$, $\exists t_3 \in \bar{\mathcal{T}}$ such that

$$\forall t > t_3 : \|\mathbf{r}(t)\|^2 \leq \epsilon_1 + \epsilon_0 + \epsilon_2 .$$

This implies that $\|\mathbf{r}(t)\| \to 0$.

Lastly, we note that since $\bar{\mathbf{P}}_1 \mathbf{r}(t)$ is not updated during GD, we have that $\bar{\mathbf{P}}_1(\tilde{\mathbf{w}} - \mathbf{w}(0)) = 0$. This sets $\tilde{\mathbf{w}}$ uniquely, together with eq. 7. ■

A.4 Proof of Lemma 10

Lemma 10 *Let $\mathcal{L}(\mathbf{w})$ be a β-smooth non-negative objective. If $\eta < 2\beta^{-1}$, then, for any $\mathbf{w}(0)$, with the GD sequence*

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla \mathcal{L}(\mathbf{w}(t)) \qquad (25)$$

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

This proof is a slightly modified version of the proof of Theorem 2 in (Ganti, 2015). Recall a well-known property of β-smooth functions:

$$\left| f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x}-\mathbf{y}) \right| \leq \frac{\beta}{2} \|\mathbf{x}-\mathbf{y}\|^2 . \tag{37}$$

From the β-smoothness of $\mathcal{L}(\mathbf{w})$

$$\begin{aligned}
\mathcal{L}(\mathbf{w}(t+1)) &\leq \mathcal{L}(\mathbf{w}(t)) + \nabla\mathcal{L}(\mathbf{w}(t))^\top (\mathbf{w}(t+1)-\mathbf{w}(t)) + \frac{\beta}{2}\|\mathbf{w}(t+1)-\mathbf{w}(t)\|^2 \\
&= \mathcal{L}(\mathbf{w}(t)) - \eta\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \frac{\beta\eta^2}{2}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \\
&= \mathcal{L}(\mathbf{w}(t)) - \eta\left(1-\frac{\beta\eta}{2}\right)\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2
\end{aligned}$$

Thus, we have

$$\frac{\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}(t+1))}{\eta\left(1-\frac{\beta\eta}{2}\right)} \geq \|\nabla\mathcal{L}(\mathbf{w}(t))\|^2$$

which implies

$$\sum_{u=0}^{t}\|\nabla\mathcal{L}(\mathbf{w}(u))\|^2 \leq \sum_{u=0}^{t}\frac{\mathcal{L}(\mathbf{w}(u)) - \mathcal{L}(\mathbf{w}(u+1))}{\eta\left(1-\frac{\beta\eta}{2}\right)} = \frac{\mathcal{L}(\mathbf{w}(0)) - \mathcal{L}(\mathbf{w}(t+1))}{\eta\left(1-\frac{\beta\eta}{2}\right)} .$$

The right hand side is upper bounded by a finite constant, since $L(\mathbf{w}(0)) < \infty$ and $0 \leq \mathcal{L}(\mathbf{w}(t+1))$. This implies

$$\sum_{u=0}^{\infty}\|\nabla\mathcal{L}(\mathbf{w}(u))\|^2 < \infty ,$$

and therefore $\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \to 0$. ■

A.5 Proof of Lemma 11

Recall that we defined $\mathbf{r}(t) = \mathbf{w}(t) - \hat{\mathbf{w}}\log t - \tilde{\mathbf{w}}$, with $\hat{\mathbf{w}}$ and $\tilde{\mathbf{w}}$ follow the conditions of the Theorems 3 and 4, *i.e*, $\hat{\mathbf{w}}$ is the L_2 max margin vector and (eq. 4), and eq. 7 holds

$$\forall n \in \mathcal{S} : \eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right) = \alpha_n .$$

Lemma 11 *We have*

$$\exists C_1, t_1 : \forall t > t_1 : (\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-\min(\theta, 1+1.5\mu_+, 1+0.5\mu_-)} . \tag{26}$$

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

$$\|\mathbf{P}_1\mathbf{r}(t)\| \geq \epsilon_1, \tag{27}$$

then the following improved bound holds

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq -C_2 t^{-1} < 0 . \tag{28}$$

From Lemma 1, $\forall n:\ \lim_{t\to\infty}\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n=\infty$. In addition, from assumption 3 the negative loss derivative $-\ell'\left(u\right)$ has an exponential tail e^{-u} (recall we assume $a=c=1$ without loss of generality). Combining both facts, we have positive constants μ_-,μ_+,t_- and t_+ such that $\forall n$

$$\forall t>t_+:-\ell'\left(\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right)\leq\left(1+\exp\left(-\mu_+\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right)\right)\exp\left(-\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right) \tag{38}$$

$$\forall t>t_-:-\ell'\left(\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right)\geq\left(1-\exp\left(-\mu_-\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right)\right)\exp\left(-\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right) \tag{39}$$

Next, we examine the expression we wish to bound, recalling that $\mathbf{r}\left(t\right)=\mathbf{w}\left(t\right)-\hat{\mathbf{w}}\log t-\tilde{\mathbf{w}}$:

$$\begin{aligned}
&\left(\mathbf{r}\left(t+1\right)-\mathbf{r}\left(t\right)\right)^{\top}\mathbf{r}\left(t\right)\\
&=\left(-\eta\nabla\mathcal{L}\left(\mathbf{w}\left(t\right)\right)-\hat{\mathbf{w}}\left[\log\left(t+1\right)-\log\left(t\right)\right]\right)^{\top}\mathbf{r}\left(t\right)\\
&=-\eta\sum_{n=1}^{N}\ell'\left(\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right)\mathbf{x}_n^{\top}\mathbf{r}\left(t\right)-\hat{\mathbf{w}}^{\top}\mathbf{r}\left(t\right)\log\left(1+t^{-1}\right)\\
&=\hat{\mathbf{w}}^{\top}\mathbf{r}\left(t\right)\left[t^{-1}-\log\left(1+t^{-1}\right)\right]-\eta\sum_{n\notin\mathcal{S}}\ell'\left(\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right)\mathbf{x}_n^{\top}\mathbf{r}\left(t\right)\\
&-\eta\sum_{n\in\mathcal{S}}\left[t^{-1}\exp\left(-\tilde{\mathbf{w}}^{\top}\mathbf{x}_n\right)+\ell'\left(\mathbf{w}\left(t\right)^{\top}\mathbf{x}_n\right)\right]\mathbf{x}_n^{\top}\mathbf{r}\left(t\right)
\end{aligned} \tag{40}$$

where in last line we used eqs. 6 and 7 to obtain

$$\hat{\mathbf{w}}=\sum_{n\in\mathcal{S}}\alpha_n\mathbf{x}_n=\eta\sum_{n\in\mathcal{S}}\exp\left(-\tilde{\mathbf{w}}^{\top}\mathbf{x}_n\right)\mathbf{x}_n\,.$$

We examine the three terms in eq. 40. The first term can be upper bounded by

$$\begin{aligned}
&\hat{\mathbf{w}}^{\top}\mathbf{r}\left(t\right)\left[t^{-1}-\log\left(1+t^{-1}\right)\right]\\
&\leq\max\left[\hat{\mathbf{w}}^{\top}\mathbf{r}\left(t\right),0\right]\left[t^{-1}-\log\left(1+t^{-1}\right)\right]\\
&\overset{(1)}{\leq}\max\left[\hat{\mathbf{w}}^{\top}\mathbf{P}_1\mathbf{r}\left(t\right),0\right]t^{-2}\\
&\overset{(2)}{\leq}\begin{cases}\|\hat{\mathbf{w}}\|\epsilon_1 t^{-2} & ,\text{ if }\|\mathbf{P}_1\mathbf{r}\left(t\right)\|\leq\epsilon_1\\ o\left(t^{-1}\right) & ,\text{ if }\|\mathbf{P}_1\mathbf{r}\left(t\right)\|>\epsilon_1\end{cases}
\end{aligned} \tag{41}$$

where in (1) we used that $\bar{\mathbf{P}}_1\hat{\mathbf{w}}=\bar{\mathbf{P}}_1\mathbf{X}_{\mathcal{S}}\boldsymbol{\alpha}=0$ from eq. 6, and in (2) we used that $\hat{\mathbf{w}}^{\top}\mathbf{r}\left(t\right)=o\left(t\right)$, since

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

where in the last line we used that $\nabla\mathcal{L}\left(\mathbf{w}\left(t\right)\right)=o\left(1\right)$, from Lemma 10.

Next, we upper bound the second term in eq. 40. From eq. 38 $\exists t'_+$, such that $\forall>t_0>t'_+$,

$$\ell'(\mathbf{w}(t)^{\top}\mathbf{x}_n)\leq 2\exp(-\mathbf{w}(t)^{\top}\mathbf{x}_n). \tag{42}$$

Therefore, $\forall t > t'_+$:

$$\begin{aligned}
&-\eta \sum_{n\notin\mathcal{S}} \ell'\left(\mathbf{w}\left(t\right)^\top \mathbf{x}_n\right) \mathbf{x}_n^\top \mathbf{r}\left(t\right) \\
\leq &-\eta \sum_{n\notin\mathcal{S}:\, \mathbf{x}_n^\top \mathbf{r}(t)\geq 0} \ell'\left(\mathbf{w}\left(t\right)^\top \mathbf{x}_n\right) \mathbf{x}_n^\top \mathbf{r}\left(t\right) \\
\overset{(1)}{\leq} &\eta \sum_{n\notin\mathcal{S}:\, \mathbf{x}_n^\top \mathbf{r}(t)\geq 0} 2\exp\left(-\mathbf{w}\left(t\right)^\top \mathbf{x}_n\right) \mathbf{x}_n^\top \mathbf{r}\left(t\right) \\
\overset{(2)}{\leq} &\eta \sum_{n\notin\mathcal{S}:\, \mathbf{x}_n^\top \mathbf{r}(t)\geq 0} 2t^{-\mathbf{x}_n^\top \hat{\mathbf{w}}} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n - \mathbf{x}_n^\top \mathbf{r}\left(t\right)\right) \mathbf{x}_n^\top \mathbf{r}\left(t\right) \\
\overset{(3)}{\leq} &\eta \sum_{n\notin\mathcal{S}:\, \mathbf{x}_n^\top \mathbf{r}(t)\geq 0} 2t^{-\mathbf{x}_n^\top \hat{\mathbf{w}}} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \\
\overset{(4)}{\leq} &\eta N \exp\left(-\min_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right) t^{-\theta}
\end{aligned} \tag{43}$$

where in (1) we used eq. 42, in (2) we used $\mathbf{w}\left(t\right) = \hat{\mathbf{w}}\log t + \tilde{\mathbf{w}} + \mathbf{r}\left(t\right)$, in (3) we used $xe^{-x} \leq 1$ and $\mathbf{x}_n^\top \mathbf{r}\left(t\right) \geq 0$,and in (4) we used $\theta > 1$, from eq. 19.

Lastly, we will bound the sum in the third term in eq. 40

$$-\eta \sum_{n\in\mathcal{S}} \left[t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) + \ell'\left(\mathbf{w}\left(t\right)^\top \mathbf{x}_n\right)\right] \mathbf{x}_n^\top \mathbf{r}\left(t\right) . \tag{44}$$

We examine each term n in this sum, and divide into two cases, depending on the sign of $\mathbf{x}_n^\top \mathbf{r}\left(t\right)$.

First, if $\mathbf{x}_n^\top \mathbf{r}\left(t\right) \geq 0$, then term n in eq. 44 can be upper bounded $\forall t > t_+$, using eq. 38, by

$$\eta t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\left[\left(1 + t^{-\mu_+}\exp\left(-\mu_+ \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\right)\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right) - 1\right] \mathbf{x}_n^\top \mathbf{r}\left(t\right) \tag{45}$$

We further divide into cases:

1. If $\left|\mathbf{x}_n^\top \mathbf{r}(t)\right| \leq C_0 t^{-0.5\mu_+}$, then we can upper bound eq. 45 with

$$\eta \exp\left(-\left(1+\mu_+\right)\min_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right) C_0 t^{-1-1.5\mu_+} . \tag{46}$$

2. If $\left|\mathbf{x}_n^\top \mathbf{r}(t)\right| > C_0 t^{-0.5\mu_+}$, then we can find $t''_+ > t'_+$ to upper bound eq. 45 $\forall t > t''_+$:

$$\begin{aligned}
&\eta t^{-1} e^{-\tilde{\mathbf{w}}^\top \mathbf{x}_n}\left[\left(1 + t^{-\mu_+} e^{-\mu_+ \tilde{\mathbf{w}}^\top \mathbf{x}_n}\right)\exp\left(-C_0 t^{-0.5\mu_+}\right) - 1\right] \mathbf{x}_n^\top \mathbf{r}\left(t\right) \\
\overset{(1)}{\leq} &\eta t^{-1} e^{-\tilde{\mathbf{w}}^\top \mathbf{x}_n}\left[\left(1 + t^{-\mu_+} e^{-\mu_+ \tilde{\mathbf{w}}^\top \mathbf{x}_n}\right)\left(1 - C_0 t^{-0.5\mu_+} + C_0^2 t^{-\mu_+}\right) - 1\right] \mathbf{x}_n^\top \mathbf{r}\left(t\right) \\
\leq &\eta t^{-1} e^{-\tilde{\mathbf{w}}^\top \mathbf{x}_n}\left[\left(1 - C_0 t^{-0.5\mu_+} + C_0^2 t^{-\mu_+}\right) e^{-\mu_+ \min_n \tilde{\mathbf{w}}^\top \mathbf{x}_n} t^{-\mu_+} - C_0 t^{-0.5\mu_+} + C_0^2 t^{-\mu_+}\right] \mathbf{x}_n^\top \mathbf{r}\left(t\right) \\
\overset{(2)}{\leq} &0,\ \forall t > t''_+
\end{aligned} \tag{47}$$

where in (1) we used the fact that $e^{-x} \leq 1 - x + x^2$ for $x \geq 0$ and in (2) we defined t''_+ so that the previous expression is negative — since $t^{-0.5\mu_+}$ decreases slower than $t^{-\mu_+}$.

3. If $\left|\mathbf{x}_n^\top \mathbf{r}(t)\right| \geq \epsilon_2$, then we define $t_+''' > t_+''$ such that $t_+''' > \exp\left(\min_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\left[e^{0.5\epsilon_2}-1\right]^{-1/\mu_+}$, and therefore $\forall t > t_+'''$, we have $\left(1+t^{-\mu_+}\exp\left(-\mu_+\tilde{\mathbf{w}}^\top\mathbf{x}_n\right)\right)e^{-\epsilon_2} < e^{-0.5\epsilon_2}$.

   This implies that $\forall t > t_+'''$ we can upper bound eq. 45 by

$$-\eta \exp\left(-\max_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\left(1-e^{-0.5\epsilon_2}\right)\epsilon_2 t^{-1}. \tag{48}$$

Second, if $\mathbf{x}_n^\top \mathbf{r}(t) < 0$, we again further divide into cases:

1. If $\left|\mathbf{x}_n^\top \mathbf{r}(t)\right| \leq C_0 t^{-0.5\mu_-}$, then, since $-\ell'\left(\mathbf{w}\left(t\right)^\top \mathbf{x}_n\right) > 0$, we can upper bound term n in eq. 44 with

$$\eta t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\mathbf{x}_n\right)\left|\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right| \leq \eta \exp\left(-\min_n \tilde{\mathbf{w}}^\top\mathbf{x}_n\right) C_0 t^{-1-0.5\mu_-} \tag{49}$$

2. If $\left|\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right| > C_0 t^{-0.5\mu_-}$, then, using eq. 39 we upper bound term n in eq. 44 with

$$\begin{aligned}
&\eta\left[-t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n} - \ell'\left(\mathbf{w}\left(t\right)^\top\mathbf{x}_n\right)\right]\mathbf{x}_n^\top\mathbf{r}\left(t\right)\\
\leq&\eta\left[-t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n} + \left(1-\exp\left(-\mu_-\mathbf{w}\left(t\right)^\top\mathbf{x}_n\right)\right)\exp\left(-\mathbf{w}\left(t\right)^\top\mathbf{x}_n\right)\right]\mathbf{x}_n^\top\mathbf{r}\left(t\right)\\
=&\eta t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}\left[1-\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\left(1-\left[t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\right]^{\mu_-}\right)\right]\left|\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right|
\end{aligned} \tag{50}$$

   Next, we will show that $\exists t'_- > t_-$ such that the last expression is strictly negative $\forall t > t'_-$. Let $M > 1$ be some arbitrary constant. Then, since $\left[t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\right]^{\mu_-} = \exp\left(-\mu_-\mathbf{w}\left(t\right)^\top\mathbf{x}_n\right) \to 0$ from Lemma 1, $\exists t_M > \max(t_-, Me^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n})$ such that $\forall t > t_M$, if $\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right) \geq M > 1$ then

$$\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\left(1-\left[t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\right]^{\mu_-}\right) \geq M' > 1\,. \tag{51}$$

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

$$\begin{aligned}
&\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\left(1-\left[t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\right]^{\mu_-}\right)\\
>&\exp\left(-\mathbf{r}\left(t\right)^\top\mathbf{x}_n\right)\left(1-\left[t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}M\right]^{\mu_-}\right).
\end{aligned} \tag{52}$$

   which is lower bounded by

$$\begin{aligned}
&\left(1+C_0t^{-0.5\mu_-}\right)\left(1-t^{-\mu_-}\left[e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}M\right]^{\mu_-}\right)\\
\geq&1+C_0t^{-0.5\mu_-}-t^{-\mu_-}\left[e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}M\right]^{\mu_-}-t^{-1.5\mu_-}\left[e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_n}M\right]^{\mu_-}C_0
\end{aligned}$$

   since $\left|\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right| > C_0t^{-0.5\mu_-}$, $\mathbf{x}_n^\top\mathbf{r}\left(t\right) < 0$ and $e^x \geq 1+x$. In this case last line is strictly larger than 1 for sufficiently large t. Therefore, after we substitute eqs. 51 and 52 into 50, we find that $\exists t'_- > t_M > t_-$ such that $\forall t > t'_-$, term k in eq. 44 is strictly negative

$$\eta\left[-t^{-1}e^{-\tilde{\mathbf{w}}^\top\mathbf{x}_k} - \ell'\left(\mathbf{w}\left(t\right)^\top\mathbf{x}_k\right)\right]\mathbf{x}_k^\top\mathbf{r}\left(t\right) < 0 \tag{53}$$

3. If $\left|\mathbf{x}_k^\top \mathbf{r}(t)\right| \geq \epsilon_2$, which is a special case of the previous case ($\left|\mathbf{x}_k^\top \mathbf{r}\left(t\right)\right| > C_0 t^{-0.5\mu_-}$) then $\forall t > t'_-$, either eq. 51 or 52 holds. Furthermore, in this case, $\exists t''_- > t'_-$ and $M'' > 1$ such that $\forall t > t''_-$ eq. 52 can be lower bounded by

$$\exp\left(\epsilon_2\right)\left(1 - \left[t^{-1} e^{-\tilde{\mathbf{w}}^\top \mathbf{x}_k} M\right]^{\mu_-}\right) > M'' > 1\,.$$

Substituting this, together with eq. 51, into eq. 50, we can find $C'_0 > 0$ such we can upper bound term k in eq. 44 with

$$-C'_0 t^{-1}\,,\ \forall t > t''_-\,. \tag{54}$$

To conclude, we choose $t_0 = \max\left[t'''_+, t''_-\right]$:

1. If $\|\mathbf{P}_1 \mathbf{r}\left(t\right)\| \geq \epsilon_1$ (as in Eq. 27), we have that

$$\max_{n\in\mathcal{S}} \left|\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right|^2 \overset{(1)}{\geq} \frac{1}{|\mathcal{S}|}\sum_{n\in\mathcal{S}} \left|\mathbf{x}_n^\top \mathbf{P}_1 \mathbf{r}\left(t\right)\right|^2 = \frac{1}{|\mathcal{S}|}\left\|\mathbf{X}_\mathcal{S}^\top \mathbf{P}_1 \mathbf{r}\left(t\right)\right\|^2 \overset{(2)}{\geq} \frac{1}{|\mathcal{S}|}\sigma^2_{\min}\left(\mathbf{X}_\mathcal{S}\right)\epsilon_1^2 \tag{55}$$

where in (1) we used $\mathbf{P}_1^\top \mathbf{x}_n = \mathbf{x}_n\ \forall n \in \mathcal{S}$, in (2) we denoted by $\sigma_{\min}\left(\mathbf{X}_\mathcal{S}\right)$, the minimal non-zero singular value of $\mathbf{X}_\mathcal{S}$ and used eq. 27. Therefore, for some k, $\left|\mathbf{x}_k^\top \mathbf{r}\right| \geq \epsilon_2 \triangleq \sqrt{|\mathcal{S}|^{-1}\sigma^2_{\min}\left(\mathbf{X}_\mathcal{S}\right)\epsilon_1^2}$. In this case, we denote C''_0 as the minimum between C'_0 (eq. 54) and $\eta \exp\left(-\max_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)\left(1 - e^{-0.5\epsilon_2}\right)\epsilon_2$ (eq. 48). Then we find that eq. 44 can be upper bounded by $-C''_0 t^{-1} + o\left(t^{-1}\right)$, $\forall t > t_0$, given eq. 27. Substituting this result, together with eqs. 41 and 43 into eq. 40, we obtain $\forall t > t_0$

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

This implies that $\exists C_2 < C''_0$ and $\exists t_2 > t_0$ such that eq. 28 holds. This implies also that eq. 26 holds for $\|\mathbf{P}_1 \mathbf{r}\left(t\right)\| \geq \epsilon_1$.

2. Otherwise, if $\|\mathbf{P}_1 \mathbf{r}\left(t\right)\| < \epsilon_1$, we find that $\forall t > t_0$, each term in eq. 44 can be upper bounded by either zero (eqs. 47 and 53), or terms proportional to $t^{-1-1.5\mu_+}$ (eq. 46) or $t^{-1-0.5\mu_-}$, (eq. 49). Combining this together with eqs. 41, 43 into eq. 40 we obtain (for some positive constants C_3, C_4, C_5, and C_6)

$$\left(\mathbf{r}\left(t+1\right) - \mathbf{r}\left(t\right)\right)^\top \mathbf{r}\left(t\right) \leq C_3 t^{-1-1.5\mu_+} + C_4 t^{-1-0.5\mu_-} + C_5 t^{-2} + C_6 t^{-\theta}\,.$$

Therefore, $\exists t_1 > t_0$ and C_1 such that eq. 26 holds. ■

Appendix B. Generic solutions of the KKT conditions in eq. 6

Lemma 12 *For almost all datasets there is a unique $\boldsymbol{\alpha}$ which satisfies the KKT conditions (eq. 6):*

$$\hat{\mathbf{w}} = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n \qquad \forall n\ \left(\alpha_n \geq 0 \textit{ and } \hat{\mathbf{w}}^\top \mathbf{x}_n = 1\right)\ \textit{OR}\ \left(\alpha_n = 0 \textit{ and } \hat{\mathbf{w}}^\top \mathbf{x}_n > 1\right)$$

Furthermore, in this solution $\alpha_n \neq 0$ if $\hat{\mathbf{w}}^\top \mathbf{x}_n = 1$, i.e.*, $\mathbf{x}_n$ is a support vector ($n \in \mathcal{S}$), and there are at most d such support vectors.*

For almost every set $\mathbf{X}$, no more than d points $\mathbf{x}_n$ can be on the same hyperplane. Therefore, since all support vectors must lie on the same hyperplane, there can be at most d support vectors, for almost every $\mathbf{X}$.

Given the set of support vectors, $\mathcal{S}$, the KKT conditions of eq. 6 entail that $\alpha_n = 0$ if $n \notin \mathcal{S}$ and

$$\mathbf{1} = \mathbf{X}_{\mathcal{S}}^\top \hat{\mathbf{w}} = \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} \boldsymbol{\alpha}_{\mathcal{S}} \,, \tag{56}$$

where we denoted $\boldsymbol{\alpha}_{\mathcal{S}}$ as $\boldsymbol{\alpha}$ restricted to the support vector components. For almost every set $\mathbf{X}$, since $d \geq |\mathcal{S}|$, $\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is invertible. Therefore, $\boldsymbol{\alpha}_{\mathcal{S}}$ has the unique solution

$$\left(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}\right)^{-1} \mathbf{1} = \boldsymbol{\alpha}_{\mathcal{S}} \,. \tag{57}$$

This implies that $\forall n \in \mathcal{S}$, α_n is equal to a rational function in the components of $\mathbf{X}_S$, *i.e.*, $\alpha_n = p_n\left(\mathbf{X}_{\mathcal{S}}\right) / q_n\left(\mathbf{X}_{\mathcal{S}}\right)$, where p_n and q_n are polynomials in the components of $\mathbf{X}_S$. Therefore, if $\alpha_n = 0$, then $p_n\left(\mathbf{X}_{\mathcal{S}}\right) = 0$, so the components of $\mathbf{X}_{\mathcal{S}}$ must be at a root of the polynomial p_n. The roots of the polynomial p_n have measure zero, unless $\forall \mathbf{X}_{\mathcal{S}} :\ p_n\left(\mathbf{X}_{\mathcal{S}}\right) = 0$. However, p_n cannot be identically equal to zero, since, for example, if $\mathbf{X}_{\mathcal{S}}^\top = \left[\mathbf{I}_{|\mathcal{S}|\times|\mathcal{S}|}, \mathbf{0}_{|\mathcal{S}|\times(d-|\mathcal{S}|)}\right]$, then $\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} = \mathbf{I}_{|\mathcal{S}|\times|\mathcal{S}|}$, and so in this case $\forall n \in \mathcal{S}$, $\alpha_n = 1 \neq 0$, from eq. 57.

Therefore, for a given $\mathcal{S}$, the event that "eq. 56 has a solution with a zero component" has a zero measure. Moreover, the union of these events, for all possible $\mathcal{S}$, also has zero measure, as a finite union of zero measures sets (there are only finitely many possible sets $\mathcal{S} \subset \{1, \ldots, N\}$). This implies that, for almost all datasets $\mathbf{X}$, $\alpha_n = 0$ only if $n \notin \mathcal{S}$. Furthermore, for almost all datasets the solution $\boldsymbol{\alpha}$ is unique: for each dataset, $\mathcal{S}$ is uniquely determined, and given $\mathcal{S}$, the solution eq. 56 is uniquely given by eq. 57. ■

Appendix C. Completing the proof of Theorem 3 for zero measure cases

In the preceding Appendices, we established Theorem 4, which only applied when all support vectors are associated with non-zero coefficients. This characterizes almost all data sets, *i.e.* all except for measure zero. We now turn to presenting and proving a more complete characterization of the limit behaviour of gradient descent, which covers all data sets, including those degenerate data sets not covered by Theorem 4, thus establishing Theorem 3.

In order to do so, we first have to introduce additional notation and a recursive treatment of the data set. We will define a sequence of data sets $\bar{\mathbf{P}}_m \mathbf{X}_{\bar{\mathcal{S}}_m}$ obtained by considering only a subset $\bar{\mathcal{S}}_m$ of the points, and projecting them using the projection matrix $\bar{\mathbf{P}}_m$. We start, for $m = 0$, with the full original data set, *i.e.* $\bar{\mathcal{S}}_0 = \{1, \ldots, N\}$ and $\bar{\mathbf{P}}_0 = \mathbf{I}_{d\times d}$. We then define $\hat{\mathbf{w}}_m$ as the max margin predictor for $\bar{\mathbf{P}}_{m-1} \mathbf{X}_{\bar{\mathcal{S}}_{m-1}}$, *i.e.*:

$$\hat{\mathbf{w}}_m = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \ \text{ s.t. } \mathbf{w}^\top \bar{\mathbf{P}}_{m-1} \mathbf{x}_n \geq 1 \,\forall n \in \bar{\mathcal{S}}_{m-1} \,. \tag{58}$$

In particular, $\hat{\mathbf{w}}_1$ is the max margin predictor for the original data set. We then denote $\mathcal{S}_m^+$ the indices of non-support vectors for 58, $\mathcal{S}_m$ the indices of support vector of 58 with non-zero coefficients for the dual variables corresponding to the margin constraints (for some dual solution), and $\bar{\mathcal{S}}_m$ the set

of support vector with zero coefficients. That is:

$$\begin{aligned}
\mathcal{S}_m^+ &= \left\{ n \in \bar{\mathcal{S}}_{m-1} | \hat{\mathbf{w}}_m^\top \bar{\mathbf{P}}_{m-1} \mathbf{x}_n > 1 \right\} \\
\mathcal{S}_m^= &= \left\{ n \in \bar{\mathcal{S}}_{m-1} | \hat{\mathbf{w}}_m^\top \bar{\mathbf{P}}_{m-1} \mathbf{x}_n = 1 \right\} = \bar{\mathcal{S}}_m \setminus \mathcal{S}_m^+ \\
\mathcal{S}_m &= \left\{ n \in \mathcal{S}_m^= | \exists \boldsymbol{\alpha} \in \mathbb{R}_{\geq 0}^N : \hat{\mathbf{w}}_m = \sum_{k=1}^N \alpha_k \bar{\mathbf{P}}_{m-1} \mathbf{x}_k, \alpha_n > 0, \forall i \notin \mathcal{S}_m^= : \alpha_i = 0 \right\} \\
\bar{\mathcal{S}}_m &= \mathcal{S}_m^= \setminus \mathcal{S}_m \,.
\end{aligned} \tag{59}$$

The problematic degenerate case, not covered by the analysis of Theorem 4, is when there are support vectors with zero coefficients, *i.e.*, when $\bar{\mathcal{S}}_m \neq \emptyset$. In this case we recurse on these zero-coefficient support vectors (*i.e.*, on $\bar{\mathcal{S}}_m$), but only consider their components orthogonal to the non-zero-coefficient support vectors (*i.e.*, not spanned by points in $\mathcal{S}_m$). That is, we project using:

$$\bar{\mathbf{P}}_m = \bar{\mathbf{P}}_{m-1} \left(\mathbf{I}_d - \mathbf{X}_{\mathcal{S}_m} \mathbf{X}_{\mathcal{S}_m}^\dagger \right) \tag{60}$$

where we denoted $\mathbf{A}^\dagger$ as the Moore-Penrose pseudo-inverse of $\mathbf{A}$. We also denote $\mathbf{P}_m = \mathbf{I}_d - \bar{\mathbf{P}}_m$.

This recursive treatment continues as long as $\bar{\mathcal{S}}_m \neq \emptyset$, defining a sequence $\hat{\mathbf{w}}_m$ of max margin predictors, for smaller and lower dimensional data sets $\bar{\mathbf{P}}_{m-1} \mathbf{X}_{\bar{\mathcal{S}}_{m-1}}$. We stop when $\bar{\mathcal{S}}_m = \emptyset$ and denote the stopping stage M—that is, M is the minimal m such that $\bar{\mathcal{S}}_m = \emptyset$. Our characterization will be in terms of the sequence $\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_M$. As established in Lemma 12 of Appendix B, for almost all data sets we will not have support vectors with non-zero coefficients, and so we will have $M = 1$, and so the characterization only depends on the max margin predictor $\hat{\mathbf{w}}_1$ of the original data set. But, even for the measure zero of data sets in which $M > 1$, we provide the following more complete characterization:

Theorem 13 *For all datasets which are linearly separable (Assumption 1) and given a β-smooth loss function (Assumption 2) with an exponential tail (Assumption 3), gradient descent (as in eq. 2) with step size $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathbf{X})$ and any starting point $\mathbf{w}(0)$, the iterates of gradient descent can be written as:*

$$\mathbf{w}(t) = \sum_{m=1}^{M} \hat{\mathbf{w}}_m \log^{\circ m}(t) + \boldsymbol{\rho}(t) \,, \tag{61}$$

where $\log^{\circ m}(t) = \overbrace{\log\log\cdots\log}^{m \text{ times}}(t)$, $\hat{\mathbf{w}}_m$ is the L_2 max margin vector defined in eq. 58, and the residual $\boldsymbol{\rho}(t)$ is bounded.

C.1 Auxiliary notation

We say that a function $f : \mathbb{N} \to \mathbb{R}$ is absolutely summable if $\sum_{t=1}^{\infty} |f(t)| < \infty$, and then we denote $f(t) \in L_1$. Furthermore, we define

$$\mathbf{r}(t) = \mathbf{w}(t) - \sum_{m=1}^{M} \left[\hat{\mathbf{w}}_m \log^{\circ m}(t) + \tilde{\mathbf{w}}_m + \sum_{k=1}^{m-1} \frac{\check{\mathbf{w}}_{k,m}}{\prod_{r=k}^{m-1} \log^{\circ r}(t)} \right]$$

where $\tilde{\mathbf{w}}_m$ and $\check{\mathbf{w}}_{k,m}$ are defined next, and additionally, we denote

$$\tilde{\mathbf{w}} = \sum_{m=1}^{M} \tilde{\mathbf{w}}_m \,.$$

We define, $\forall m \geq 1$, $\tilde{\mathbf{w}}_m$ as the solution of

$$\forall m \geq 1 : \forall n \in \mathcal{S}_m : \eta \sum_{n \in \mathcal{S}_m} \exp\left(-\sum_{k=1}^{m} \tilde{\mathbf{w}}_k^\top \mathbf{x}_n\right) \bar{\mathbf{P}}_{m-1} \mathbf{x}_n = \hat{\mathbf{w}}_m \,, \tag{62}$$

such that

$$\mathbf{P}_{m-1} \tilde{\mathbf{w}}_m = 0 \,\text{and}\, \bar{\mathbf{P}}_m \tilde{\mathbf{w}}_m = 0. \tag{63}$$

The existence and uniqueness of the solution, $\tilde{\mathbf{w}}_m$ are proved in appendix section C.4.

Lastly, we define, $\forall m > k \geq 1$, $\check{\mathbf{w}}_{k,m}$ as the solution of

$$\sum_{n \in \mathcal{S}_m} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{P}_{m-1} \mathbf{x}_n = \sum_{k=1}^{m-1} \left[\sum_{n \in \mathcal{S}_k} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{x}_n \mathbf{x}_n^\top\right] \check{\mathbf{w}}_{k,m} \tag{64}$$

such that

$$\mathbf{P}_{k-1} \check{\mathbf{w}}_{k,m} = 0 \,\text{and}\, \bar{\mathbf{P}}_k \check{\mathbf{w}}_{k,m} = 0 \,. \tag{65}$$

The existence and uniqueness of the solution $\check{\mathbf{w}}_{k,m}$ are proved in appendix section C.5.

Together, eqs. 62-65 entail the existence of a unique decomposition, $\forall m \geq 1$:

$$\hat{\mathbf{w}}_m = \eta \sum_{n \in \mathcal{S}_m} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{x}_n - \eta \sum_{k=1}^{m-1} \left[\sum_{n \in \mathcal{S}_k} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{x}_n \mathbf{x}_n^\top\right] \check{\mathbf{w}}_{k,m} \tag{66}$$

given the constraints in eqs. 63 and 65 hold.

C.2 Proof of Theorem 13

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

$$\boldsymbol{\tau}(t) = \sum_{m=2}^{M} \hat{\mathbf{w}}_m \log^{\circ m}(t) + \sum_{m=1}^{M} \sum_{k=1}^{m-1} \frac{\check{\mathbf{w}}_{k,m}}{\prod_{r=k}^{m-1} \log^{\circ r}(t)}$$

noting that

$$\|\boldsymbol{\tau}(t+1) - \boldsymbol{\tau}(t)\| \leq \frac{C_\tau}{t \log(t)}$$

and

$$\mathbf{r}(t) = \mathbf{w}(t) - \hat{\mathbf{w}}_1 \log(t) - \tilde{\mathbf{w}} - \boldsymbol{\tau}(t) \tag{67}$$

where $\tilde{\mathbf{w}}$ follow the conditions of Theorem 13. Our goal is to show that $\|\mathbf{r}(t)\|$ is bounded. To show this, we will upper bound the following equation

$$\|\mathbf{r}(t+1)\|^2 = \|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 + 2\left(\mathbf{r}(t+1) - \mathbf{r}(t)\right)^\top \mathbf{r}(t) + \|\mathbf{r}(t)\|^2 \tag{68}$$

First, we note that $\exists t_0$ such that $\forall t > t_0$ the first term in this equation can be upper bounded by

$$\begin{aligned}
&||\mathbf{r}(t+1) - \mathbf{r}(t)||^2 \\
&\overset{(1)}{=} ||\mathbf{w}(t+1) - \hat{\mathbf{w}}_1 \log(t+1) - \boldsymbol{\tau}(t+1) - \mathbf{w}(t) + \hat{\mathbf{w}}_1 \log(t) + \boldsymbol{\tau}(t)||^2 \\
&\overset{(2)}{=} ||-\eta \nabla L(\mathbf{w}(t)) - \hat{\mathbf{w}}_1(\log(t+1) - \log(t)) - (\boldsymbol{\tau}(t+1) - \boldsymbol{\tau}(t))||^2 \\
&= \eta^2 ||\nabla L(\mathbf{w}(t))||^2 + \|\hat{\mathbf{w}}_1\|^2 \log^2\left(1 + t^{-1}\right) + \|\boldsymbol{\tau}(t+1) - \boldsymbol{\tau}(t)\|^2 \\
&+ 2\eta \nabla L(\mathbf{w}(t))^\top \left(\hat{\mathbf{w}}_1 \log\left(1 + t^{-1}\right) + \boldsymbol{\tau}(t+1) - \boldsymbol{\tau}(t)\right) \\
&+ 2\hat{\mathbf{w}}_1^\top (\boldsymbol{\tau}(t+1) - \boldsymbol{\tau}(t)) \log\left(1 + t^{-1}\right) \\
&\overset{(3)}{\leq} \eta^2 ||\nabla L(\mathbf{w}(t))||^2 + \|\hat{\mathbf{w}}_1\|^2 t^{-2} + C_\tau^2 t^{-2} \log^{-2}(t) + 2C_\tau \|\hat{\mathbf{w}}_1\| t^{-2} \log^{-1}(t) \ \ , \forall t > t_0
\end{aligned} \tag{69}$$

where in (1) we used eq. 67, in (2) we used eq. 2 and in (3) we used $\forall x > 0 : \ x \geq \log(1+x) > 0$, and also using $\ell'(\mathbf{w}(t)^\top \mathbf{x}_n) < 0$ for large enough t, we have that

$$\left(\hat{\mathbf{w}}_1 \log\left(1 + t^{-1}\right) + \boldsymbol{\tau}(t+1) - \boldsymbol{\tau}(t)\right)^\top \nabla \mathcal{L}(\mathbf{w}(t)) \leq \sum_{n=1}^{N} \ell'(\mathbf{w}(t)^\top \mathbf{x}_n) \left(\hat{\mathbf{w}}_1^\top \mathbf{x}_n \log\left(1 + t^{-1}\right) - \frac{\|\mathbf{x}_n\| C'_\tau}{t \log(t)}\right) \tag{70}$$

which is negative for sufficiently large t_0 (since $\log\left(1 + t^{-1}\right)$ decreases as t^{-1}, which is slower then $1/(t \log(t))$), $\forall n : \ \hat{\mathbf{w}}_1^\top \mathbf{x}_n \geq 1$ and $\ell'(u) \leq 0$.

Also, from Lemma 10 we know that:

$$\|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 = o(1) \text{ and } \sum_{u=0}^{\infty} \|\nabla \mathcal{L}(\mathbf{w}(u))\|^2 < \infty \tag{71}$$

Substituting eq. 71 into eq. 69, and recalling that $t^{-\nu_1} \log^{-\nu_2}(t)$ converges for any $\nu_1 > 1$ and any ν_2, and so

$$\kappa_0(t) \triangleq ||\mathbf{r}(t+1) - \mathbf{r}(t)||^2 \in L_1 . \tag{72}$$

Also, in the next subsection we will prove that

Lemma 14 *Let $\kappa_1(t)$ and $\kappa_2(t)$ be functions in L_1, then*

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq \kappa_1(t) \|\mathbf{r}(t)\| + \kappa_2(t) \tag{73}$$

Thus, by combining eqs. 73 and 72 into eq. 68, we find

$$\|\mathbf{r}(t+1)\|^2 \leq \kappa_0(t) + 2\kappa_1(t) \|\mathbf{r}(t)\| + 2\kappa_2(t) + \|\mathbf{r}(t)\|^2$$

On this result we apply the following lemma (with $\phi(t) = \|\mathbf{r}(t)\|$, $h(t) = 2\kappa_1(t)$, and $z(t) = \kappa_0(t) + 2\kappa_2(t)$), which we prove in appendix C.6:

Lemma 15 *Let $\phi(t), h(t), z(t)$ be three functions from $\mathbb{N}$ to $\mathbb{R}_{\geq 0}$, and C_1, C_2, C_3 be three positive constants. Then, if $\sum_{t=1}^{\infty} h(t) \leq C_1 < \infty$, and*

$$\phi^2(t+1) \leq z(t) + h(t)\phi(t) + \phi^2(t) \tag{74}$$

we have

$$\phi^2(t+1) \leq C_2 + C_3 \sum_{u=1}^{t} z(u) \tag{75}$$

and obtain that

$$\|\mathbf{r}(t+1)\|^2 \leq C_2 + C_3 \sum_{u=1}^{t} \left(\kappa_0(u) + 2\kappa_2(u)\right) \leq C_4 < \infty\,,$$

since we assumed that $\forall i = 0,1,2:\ \kappa_i(t) \in L_1$. This completes our proof. ■

C.3 Proof of Lemma 14

Before we prove Lemma 14, we prove the following auxilary Lemma:

Lemma 16 *Consider the function* $f(t) = t^{-\nu_1}(\log(t))^{-\nu_2}(\log\log(t))^{-\nu_3}\ldots(\log^{\circ M}(t))^{-\nu_{M+1}}$. *If* $\exists m_0 \leq M+1$ *such that* $\nu_{m_0} > 1$ *and for all* $m' < m_0$,$\nu_{m'} = 1$, *then* $f(t) \in L_1$.

Proof To prove Lemma 16, we will show that the improper integeral $\int_{t_1}^{\infty} f(t)dt$ for any $t_1 > 0$ is bounded, *i.e.*, $\forall t_1 > 0, \int_{t_1}^{\infty} f(t)dt < C$. Using the integeral test for convergence (or Maclaurin–Cauchy test) this in turn implies that $\forall t_1 > 0, \sum_{t_1}^{\infty} f(t) < C$, and thus $f(t) \in L_1$.

First, if $m_0 > 1$, then $\nu_1 = \nu_2 \ldots = \nu_{m_0-1} = 1$ and $\nu_{m_0} = 1+\epsilon$ for some $\epsilon > 0$. Using change of variables $y = \log^{\circ(m_0-1)}(t)$, we have

$$\mathrm{d}y = \left(t\prod_{r=1}^{m_0-2}\log^{\circ r}(t)\right)^{-1}\mathrm{d}t = t^{-\nu_1}\prod_{r=1}^{m_0-2}\left(\log^{\circ r}(t)\right)^{-\nu_{r+1}}\mathrm{d}t$$

and for all $m > m_0$, $\left(\log^{\circ(m-1)}(t)\right)^{-\nu_m} = \left(\log^{\circ(m-m_0)}(y)\right)^{-\nu_m} \leq (\log(y))^{|\nu_m|}$. Thus, denoting $\tilde{\nu} = \sum_{m=m_0+1}^{M+1}|\nu_m|$ and $\log^{\circ(m_0-1)}(t_1) = y_1$, we have

$$\int_{t_1}^{\infty} f(t)\mathrm{d}t = \int_{y_1}^{\infty} y^{-\nu_{m_0}}\prod_{m=m_0+1}^{M+1}\left(\log^{\circ m-m_0}(y)\right)^{-\nu_m}\mathrm{d}(y) \leq \int_{y_1}^{\infty}\frac{(\log(y))^{\tilde{\nu}}}{y^{1+\epsilon}}\mathrm{d}y. \tag{76}$$

For $m_0 = 1$, we have $\nu_1 = 1+\epsilon$ for some $\epsilon > 0$, and for $m > 1$, $\left(\log^{\circ(m-1)}(t)\right)^{-\nu_m} \leq (\log(t))^{|\nu_m|}$. Thus, denoting, $\tilde{\nu} = \sum_{m=2}^{M+1}|\nu_m|$, we have $\int_{t_1}^{\infty} f(t)\mathrm{d}t \leq \int_{t_1}^{\infty}\frac{(\log(t))^{\tilde{\nu}}}{t^{1+\epsilon}}\mathrm{d}t$.

Thus, for any m_0, we only need to show that for all $t_1 > 0, \epsilon > 0$ and $\tilde{\nu} > 0$, $\int_{t_1}^{\infty}\frac{(\log(t))^{\tilde{\nu}}}{t^{1+\epsilon}}\mathrm{d}t < \infty$.

Let us now look at $\int_{t_1}^{\infty}\frac{(\log(t))^{\tilde{\nu}}}{t^{1+\epsilon}}\mathrm{d}t$. using $u = (\log(t))^{\tilde{\nu}}$ and $\mathrm{d}v = \frac{1}{t^{1+\epsilon}}$, we have $\mathrm{d}u = \tilde{\nu}t^{-1}(\log(t))^{\tilde{\nu}-1}$ and $v = -\frac{1}{\epsilon t^{\epsilon}}$. Using integration by parts, $\int u\mathrm{d}v = uv - \int v\mathrm{d}u$, we have

$$\int\frac{(\log(t))^{\tilde{\nu}}}{t^{1+\epsilon}}\mathrm{d}t = -\frac{(\log(t))^{\tilde{\nu}}}{\epsilon t^{\epsilon}} + \frac{\bar{\nu}}{\epsilon}\int\frac{(\log(t))^{\tilde{\nu}-1}}{t^{1+\epsilon}}\mathrm{d}t$$

Recursing the above equation K times such that $\tilde{\nu} - K < 0$, we have positive constants $c_0, c_1, \dots c_K > 0$ independent of t, such that

$$\begin{aligned}\int_{t_1}^{\infty} \frac{(\log(t))^{\tilde{\nu}}}{t^{1+\epsilon}}\mathrm{d}t &= \left[-\sum_{k=0}^{K-1} \frac{c_k (\log(t))^{\tilde{\nu}-k}}{\epsilon t^{\epsilon}}\right]_{t=t_1}^{\infty} + c_K \int_{t=t_1}^{\infty} \frac{(\log(t))^{\tilde{\nu}-K}}{t^{1+\epsilon}}\mathrm{d}t \\ &\overset{(1)}{=} \sum_{k=0}^{K-1} \frac{c_k (\log(t_1))^{\tilde{\nu}-k}}{\epsilon t_1^{\epsilon}} + c_K \int_{t=t_1}^{\infty} \frac{(\log(t))^{\tilde{\nu}-K}}{t^{1+\epsilon}}\mathrm{d}t \\ &\overset{(2)}{\leq} \sum_{k=0}^{K-1} \frac{c_k (\log(t_1))^{\tilde{\nu}-k}}{\epsilon t_1^{\epsilon}} + c_K \int_{t=t_1}^{\infty} \frac{1}{t^{1+\epsilon}} \overset{(3)}{} \\ &= \sum_{k=0}^{K-1} \frac{c_k (\log(t_1))^{\tilde{\nu}-k}}{\epsilon t_1^{\epsilon}} y + \frac{c_K}{\epsilon t_1^{\epsilon}} < \infty \end{aligned} \tag{77}$$

where (1) follows as $\sum_{k=0}^{K-1} \frac{c_k(\log(t))^{\tilde{\nu}-k}}{\epsilon t^{\epsilon}} \overset{t\to\infty}{\to} 0$, (2) follows as K is chosen such that $\tilde{\nu} - K < 0$ and hence for all $t > 0$, $(\log(t))^{\tilde{\nu}-K} < 1$. This completes the proof of the lemma. ■

Lemma 14 *Let $\kappa_1(t)$ and $\kappa_2(t)$ be functions in L_1, then*

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^{\top} \mathbf{r}(t) \leq \kappa_1(t) \|\mathbf{r}(t)\| + \kappa_2(t) \tag{73}$$

Proof Recall that we defined

$$\mathbf{r}(t) = \mathbf{w}(t) - \mathbf{q}(t) \tag{78}$$

where

$$\mathbf{q}(t) = \sum_{m=1}^{M} \left[\hat{\mathbf{w}}_m \log^{\circ m}(t) + \mathbf{h}_m(t)\right] . \tag{79}$$

$$\mathbf{h}_m(t) = \tilde{\mathbf{w}}_m + \sum_{k=1}^{m-1} \frac{\check{\mathbf{w}}_{k,m}}{\prod_{r=k}^{m-1} \log^{\circ r}(t)} \tag{80}$$

with $\hat{\mathbf{w}}_m$, $\tilde{\mathbf{w}}_m$ and $\check{\mathbf{w}}_{k,m}$ defined in eqs. 58, 62 and 64, respectively. We note that

$$\|\mathbf{q}(t+1) - \mathbf{q}(t) - \dot{\mathbf{q}}(t)\| \leq C_q t^{-2} \in L_1 \tag{81}$$

where

$$\dot{\mathbf{q}}(t) = \sum_{m=1}^{M} \hat{\mathbf{w}}_m \frac{1}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} + \dot{\mathbf{h}}_m(t) . \tag{82}$$

Additionally, we define C_h, C_h' so that

$$\|\mathbf{h}_m(t)\| \leq \|\tilde{\mathbf{w}}_m\| + \sum_{k=1}^{m} \|\check{\mathbf{w}}_{k,m}\| \leq C_h \tag{83}$$

and

$$\left\|\dot{\mathbf{h}}_m(t)\right\| \leq \frac{C_h'}{t\left(\prod_{r=1}^{m-2}\log^{\circ r}(t)\right)\left(\log^{\circ(m-1)}(t)\right)^2} \in L_1\,. \tag{84}$$

We wish to calculate

$$\begin{aligned}
&(\mathbf{r}(t+1)-\mathbf{r}(t))^\top \mathbf{r}(t)\\
&\overset{(1)}{=}[\mathbf{w}(t+1)-\mathbf{w}(t)-[\mathbf{q}(t+1)-\mathbf{q}(t)]]^\top \mathbf{r}(t)\\
&\overset{(2)}{=}[-\eta\nabla\mathcal{L}(\mathbf{w}(t))-\dot{\mathbf{q}}(t)]^\top \mathbf{r}(t)-[\mathbf{q}(t+1)-\mathbf{q}(t)-\dot{\mathbf{q}}(t)]^\top \mathbf{r}(t)
\end{aligned} \tag{85}$$

where in (1) we used eq. 78 and in (2) we used the definition of GD in eq. 2. We can bound the second term using Cauchy-Shwartz inequality and eq. 81:

$$[\mathbf{q}(t+1)-\mathbf{q}(t)-\dot{\mathbf{q}}(t)]^\top \mathbf{r}(t) \leq \|\mathbf{q}(t+1)-\mathbf{q}(t)-\dot{\mathbf{q}}(t)\|\,\|\mathbf{r}(t)\| \leq C_q t^{-2}\|\mathbf{r}(t)\|\,.$$

Next, we examine the second term in eq. 85

$$\begin{aligned}
&[-\eta\nabla\mathcal{L}(\mathbf{w}(t))-\dot{\mathbf{q}}(t)]^\top \mathbf{r}(t)\\
&=\left[-\eta\sum_{n=1}^{N}\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\,\mathbf{x}_n-\dot{\mathbf{q}}(t)\right]^\top \mathbf{r}(t)\\
&\overset{(1)}{=}-\sum_{m=1}^{M}\dot{\mathbf{h}}_m(t)^\top \mathbf{r}(t)-\eta\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m^+}\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\,\mathbf{x}_n^\top\mathbf{r}(t)\\
&+\left[\eta\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}-\ell'(\mathbf{w}(t)^\top\mathbf{x}_n)\,\mathbf{x}_n-\sum_{m=1}^{M}\hat{\mathbf{w}}_m\frac{1}{t\prod_{r=1}^{m-1}\log^{\circ r}(t)}\right]^\top \mathbf{r}(t)\,,
\end{aligned} \tag{86}$$

where in (1) recall from eq. 59 that $\mathcal{S}_m, \mathcal{S}_m^+$ are mutually exclusive and $\cup_{m=1}^{M}\mathcal{S}_m \cup \mathcal{S}_m^+ = [N]$.

Next we upper bound the three terms in eq. 86.

To bound the first term in eq. 86 we use Cauchy-Shartz, and eq. 84.

$$\sum_{m=1}^{M}\dot{\mathbf{h}}_m(t)^\top \mathbf{r}(t) \leq \sum_{m=1}^{M}\left\|\dot{\mathbf{h}}_m(t)\right\|\|\mathbf{r}(t)\| \leq \frac{MC_h'}{t\left(\prod_{r=1}^{m-2}\log^{\circ r}(t)\right)\left(\log^{\circ(m-1)}(t)\right)^2}\|\mathbf{r}(t)\|$$

In bounding the second term in eq. 86, note that for tight exponential tail loss, since $\mathbf{w}(t)^\top\mathbf{x}_n \to \infty$, for large enough t_0, we have $-\ell'(\mathbf{w}(t)^\top\mathbf{x}_n) \leq (1+\exp(-\mu_+\mathbf{w}(t)^\top\mathbf{x}_n))\exp(-\mathbf{w}(t)^\top\mathbf{x}_n) \leq 2\exp(-\mathbf{w}(t)^\top\mathbf{x}_n)$ for all $t > t_0$. The first term in eq. 86 can be bounded by the following set of

inequalities, for $t > t_0$,

$$\begin{aligned}
&\eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m^+} -\ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\, \mathbf{x}_n^\top \mathbf{r}(t) \leq \eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m^+ : \mathbf{x}_n^\top \mathbf{r}(t) \geq 0} -\ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\, \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(1)}{\leq} 2\eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m^+ : \mathbf{x}_n^\top \mathbf{r}(t) \geq 0} \exp\left(-\sum_{l=1}^{M} \left[\hat{\mathbf{w}}_l^\top \mathbf{x}_n \log^{\circ l}(t) + \mathbf{x}_n^\top \mathbf{h}_l(t) \right] - \mathbf{x}_n^\top \mathbf{r}(t) \right) \mathbf{x}_n^\top \mathbf{r}(t) \\
&\overset{(2)}{\leq} 2\eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m^+ : \mathbf{x}_n^\top \mathbf{r}(t) \geq 0} \exp\left(-\sum_{l=1}^{M} \left[\hat{\mathbf{w}}_l^\top \mathbf{x}_n \log^{\circ l}(t) + \mathbf{x}_n^\top \mathbf{h}_l(t) \right] \right) \\
&\overset{(3)}{\leq} 2\eta \sum_{m=1}^{M} \left| \mathcal{S}_m^+ \max_{n \in \mathcal{S}_m^+} \right| \exp\left(M \left\| \mathbf{x}_n \right\| C_h \right) \exp\left(-\sum_{l=1}^{M} \hat{\mathbf{w}}_l^\top \mathbf{x}_n \log^{\circ l}(t) \right) \\
&\overset{(4)}{\leq} \begin{cases} \sum_{m=1}^{M} \frac{2\eta \left| \mathcal{S}_m^+ \right| \exp\left(M \max_{n \in \mathcal{S}_m^+} \|\mathbf{x}_n\| C_h \right)}{t \left(\prod_{k=1}^{m-1} \log^{\circ k}(t) \right) \left(\log^{\circ m-1}(t) \right)^{\theta_m} \left(\prod_{k=m}^{M-1} (\log^{\circ m}(t))^{\hat{\mathbf{w}}_k^\top \mathbf{x}_n} \right)} & \text{if } M > 1 \\ \frac{2\eta \left| \mathcal{S}_1^+ \right| \exp(\max_n \|\mathbf{x}_n\| C_h)}{t^{\theta_1}} & \text{if } M = 1 \end{cases} \in L_1. \qquad (87)
\end{aligned}$$

where in (1) we used eqs. 78 and 79, in (2) we used that $\forall x : xe^{-x} \leq 1$ and $\mathbf{x}_n^\top \mathbf{r}(t) \geq 0$, (3) we used eq. 83 and in (4) we denoted $\theta_m = \min_{n \in \mathcal{S}_m^+} \hat{\mathbf{w}}_m^\top \mathbf{x}_n > 1$ and the last line is integrable based on Lemma 16.

Next, we bound the last term in eq. 86. For exponential tailed losses (Assumption 3), since $\mathbf{w}(t)^\top \mathbf{x}_n \to \infty$, we have positive constants $\mu_-, \mu_+ > 0$, t_- and t_+ such that $\forall n$

$$\begin{aligned}
\forall t > t_+ : -\ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n \right) &\leq \left(1 + \exp\left(-\mu_+ \mathbf{w}(t)^\top \mathbf{x}_n \right) \right) \exp\left(-\mathbf{w}(t)^\top \mathbf{x}_n \right) \\
\forall t > t_- : -\ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n \right) &\geq \left(1 - \exp\left(-\mu_- \mathbf{w}(t)^\top \mathbf{x}_n \right) \right) \exp\left(-\mathbf{w}(t)^\top \mathbf{x}_n \right)
\end{aligned}$$

We define $\gamma_n(t)$ as

$$\gamma_n(t) = \begin{cases} (1 + \exp(-\mu_+ \mathbf{w}(t)^\top \mathbf{x}_n) & \text{if } \mathbf{r}(t)^\top \mathbf{x}_n \geq 0 \\ (1 - \exp(-\mu_- \mathbf{w}(t)^\top \mathbf{x}_n) & \text{if } \mathbf{r}(t)^\top \mathbf{x}_n < 0 \end{cases}. \qquad (88)$$

This implies $t > \max(t_+, t_-)$, $-\ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\, \mathbf{x}_n^\top \mathbf{r}(t) \leq \gamma_n(t) \exp\left(-\mathbf{w}^\top(t) \mathbf{x}_n \right) \mathbf{x}_n$.

From this result, we have the following set of inequalities:

$$
\begin{aligned}
&\eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} -\ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\, \mathbf{x}_n^\top \mathbf{r}\,(t) \leq \eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} \gamma_n(t) \exp\left(-\mathbf{w}\,(t)^\top \mathbf{x}_n\right) \mathbf{x}_n^\top \mathbf{r}\,(t) \\
&\overset{(1)}{=} \eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} \gamma_n(t) \exp\left(- \sum_{l=1}^{M} \left[\hat{\mathbf{w}}_l^\top \mathbf{x}_n \log^{\circ l}(t) + \mathbf{x}_n^\top \tilde{\mathbf{w}}_l + \sum_{k=1}^{l-1} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{k,l}}{\prod_{r=k}^{l-1} \log^{\circ r}(t)} \right] - \mathbf{x}_n^\top \mathbf{r}\,(t) \right) \mathbf{x}_n^\top \mathbf{r}\,(t) \\
&\overset{(2)}{=} \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} \frac{\eta \gamma_n(t) \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) \exp\left(-\mathbf{x}_n^\top \mathbf{r}\,(t)\right) \mathbf{x}_n^\top \mathbf{r}\,(t)}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} \exp\left(- \sum_{k=1}^{m} \sum_{l=k+1}^{M} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{k,l}}{\prod_{r=k}^{l-1} \log^{\circ r}(t)} \right) \\
&\overset{(3)}{=} \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} \frac{\eta \gamma_n(t) \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) \exp\left(-\mathbf{x}_n^\top \mathbf{r}\,(t)\right) \mathbf{x}_n^\top \mathbf{r}\,(t)}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} \exp\left(- \sum_{l=m+1}^{M} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{m,l}}{\prod_{r=m}^{l-1} \log^{\circ r}(t)} \right) \psi_m(t) \\
&\leq \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} \frac{\eta \gamma_n(t) \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) \exp\left(-\mathbf{x}_n^\top \mathbf{r}\,(t)\right) \mathbf{x}_n^\top \mathbf{r}\,(t)}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} \psi_m(t) \left[\left(1 - \sum_{l=m}^{M-1} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l} \log^{\circ r}(t)} \right) \right. \\
&\qquad \left. + \exp\left(- \sum_{l=m}^{M-1} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l} \log^{\circ r}(t)} \right) - \left(1 - \sum_{l=m}^{M-1} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l} \log^{\circ r}(t)} \right) \right] \qquad (89)
\end{aligned}
$$

where in (1) we used eqs. 78 and 79, and in (2) we used $\mathbf{P}_{k-1} \check{\mathbf{w}}_{k,m} = 0$ from eq. 65 (so $\mathbf{x}_n^\top \check{\mathbf{w}}_{k,l} = 0$ if $m < k$) and in (3) defined

$$
\psi_m(t) = \exp\left(- \sum_{k=1}^{m-1} \sum_{l=k+1}^{M} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{k,l}}{\prod_{r=k}^{l-1} \log^{\circ r}(t)} \right). \qquad (90)
$$

Note $\exists t_\psi$ such that $\forall t > t_\psi$, we can bound $\psi_m(t)$ by

$$
\exp\left(\frac{-M \max_n \|\mathbf{x}_n\| C_h}{\log^{\circ (m-1)}(t)} \right) \leq \psi_m(t) \leq 1\,. \qquad (91)
$$

Thus, the third term in 86 is given by

$$
\begin{aligned}
&\eta \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} -\ell'(\mathbf{w}(t)^\top \mathbf{x}_n)\, \mathbf{x}_n^\top \mathbf{r}\,(t) - \sum_{m=1}^{M} \frac{\hat{\mathbf{w}}_m^\top \mathbf{r}\,(t)}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} \\
&\overset{(1)}{\leq} \sum_{m=1}^{M} \sum_{n \in \mathcal{S}_m} \frac{\eta \gamma_n(t) \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) \exp\left(-\mathbf{x}_n^\top \mathbf{r}\,(t)\right) \mathbf{x}_n^\top \mathbf{r}\,(t)}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} \psi_m(t) \left[\exp\left(- \sum_{l=m}^{M-1} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l} \log^{\circ r}(t)} \right) \right. \\
&\qquad \left. - \left(1 - \sum_{l=m}^{M-1} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l} \log^{\circ r}(t)} \right) \right] \\
&\quad + \sum_{m=1}^{M} \left[\sum_{n \in \mathcal{S}_m} \frac{\eta \gamma_n(t) \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) \exp\left(-\mathbf{x}_n^\top \mathbf{r}\,(t)\right) \mathbf{x}_n^\top \mathbf{r}\,(t)}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} \psi_m(t) \left(1 - \sum_{l=m}^{M-1} \frac{\mathbf{x}_n^\top \check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l} \log^{\circ r}(t)} \right) \right. \\
&\quad \left. - \frac{\mathbf{r}(t)^\top \hat{\mathbf{w}}_m}{t \prod_{r=1}^{m-1} \log^{\circ r}(t)} \right], \qquad (92)
\end{aligned}
$$

where (1) follows from the bound in eq. 89.

We examine the first term in eq. 92

$$\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}\frac{\eta\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\psi_m\left(t\right)$$
$$\cdot\left[\exp\left(-\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)-\left(1-\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)\right]$$

$\forall t>t_1>t_\psi$, where we will determine t_1 later. We have the following for all $m\in[M]$

$$\sum_{n\in\mathcal{S}_m}\frac{\eta\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\psi_m\left(t\right)$$
$$\cdot\left[\exp\left(-\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)-\left(1-\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)\right]$$
$$\overset{(1)}{\leq}\sum_{\substack{n\in\mathcal{S}_m:\\ \mathbf{x}_n^\top\mathbf{r}(t)\geq 0}}\frac{\eta\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\psi_m\left(t\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\left[\exp\left(-\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)-\left(1-\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)\right]$$
$$\overset{(2)}{\leq}\sum_{\substack{n\in\mathcal{S}_m:\\ \mathbf{x}_n^\top\mathbf{r}(t)\geq 0}}\frac{\eta\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\left(t\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\psi_m\left(t\right)}\left(\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)^2\in L_1\,, \tag{93}$$

where we set $t_1>0$ such that $\forall t>t_1$ the term in the square bracket is positive and

$$\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}>-1\,,$$

in (1) we used that since $e^{-x}\geq 1-x$, and also from using $e^{-x}x\leq 1$ and in (2) we use that $\forall x\geq -1$ we have that $e^{-x}\leq 1-x+x^2$ and $\psi_m\left(t\right)\leq 1$ from eq. 91.

We examine the second term in eq. 92 using the decomposition of $\hat{\mathbf{w}}_m$ from eq. 66

$$
\begin{aligned}
&\sum_{m=1}^{M}\left[\sum_{n\in\mathcal{S}_m}\frac{\eta\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\psi_m\left(t\right)\left(1-\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\right)-\frac{\mathbf{x}_n^\top\hat{\mathbf{w}}_m}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\right]\\
&\overset{(1)}{=}\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}\frac{\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\left(\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\psi_m\left(t\right)-1\right)\\
&\quad-\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}\frac{\eta\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)\psi_m(t)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\sum_{l=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{\prod_{r=m}^{l}\log^{\circ r}\left(t\right)}\\
&\quad+\sum_{m=1}^{M}\sum_{k=1}^{m-1}\sum_{n\in\mathcal{S}_k}\frac{\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)\mathbf{x}_n^\top\check{\mathbf{w}}_{k,m}}{\prod_{r=1}^{m-1}t\log^{\circ r}\left(t\right)}\\
&\overset{(2)}{=}\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}\frac{\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\left(\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\psi_m\left(t\right)-1\right)\\
&\quad-\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}\sum_{l=m}^{M-1}\frac{\eta\gamma_n(t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)\psi_m(t)\mathbf{x}_n^\top\check{\mathbf{w}}_{m,l+1}}{t\prod_{r=1}^{l}\log^{\circ r}\left(t\right)}\\
&\quad+\sum_{k=1}^{M}\sum_{n\in\mathcal{S}_k}\sum_{m=k}^{M-1}\frac{\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)\mathbf{x}_n^\top\check{\mathbf{w}}_{k,m+1}}{\prod_{r=1}^{m}t\log^{\circ r}\left(t\right)}\\
&\overset{(3)}{=}\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}\left[\frac{1}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}-\sum_{k=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,k+1}}{t\prod_{r=1}^{k}\log^{\circ r}\left(t\right)}\right]\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\left(\gamma_n(t)\psi_m\left(t\right)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)-1\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)\\
&:=\sum_{m=1}^{M}\sum_{n\in\mathcal{S}_m}\Gamma_{m,n}(t),
\end{aligned}
\tag{94}
$$

where in (1) we used eq. 66, in (2) we re-arranged the order of summation in the last term, and in (3) we just use a change of variables.

Next, we examine $\Gamma_{m,n}(t)$ for each m and $n\in\mathcal{S}_m$ in eq. 94. Note that, $\exists t_2 > t_\psi$ such that $\forall t > t_2$ we have

$$
\left|\sum_{k=m}^{M-1}\frac{\mathbf{x}_n^\top\check{\mathbf{w}}_{m,k+1}}{t\prod_{r=1}^{k}\log^{\circ r}\left(t\right)}\right|\leq\frac{0.5}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}.
$$

In this case, $\forall t > t_2$

$$
\Gamma_{m,n}(t)\overset{(1)}{\leq}\eta\left[\frac{\kappa(n,t)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\right]\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)\left(\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)-1\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right),
\tag{95}
$$

where in (1) follows from the definition of t_2, wherein

$$
\kappa_n(t)=\begin{cases}1.5 & \text{if }\left(\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)-1\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)>0\\ 0.5 & \text{if }\left(\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)-1\right)\mathbf{x}_n^\top\mathbf{r}\left(t\right)<0\end{cases}.
$$

1. First, if $\mathbf{x}_n^\top\mathbf{r}\left(t\right)>0$, then $\gamma_n(t)=(1+\exp(-\mu_+\mathbf{w}(t)^\top\mathbf{x}_n))>0$.

We further divide into two cases. In the following C_0, C_1 are some constants independent of t.

(a) If $\left|\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right| > C_0 t^{-0.5\mu_+}$, then we have the following

$$\begin{aligned}
&\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right)\\
&\overset{(1)}{\leq}\left(1+\exp\left(-\mu_+\sum_{l=1}^{M}\left[\hat{\mathbf{w}}_l^\top \mathbf{x}_n \log^{\circ l}(t)+\mathbf{h}_l^\top \mathbf{x}_n\right]\right)\right)\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right)\\
&\overset{(2)}{\leq}\left(1+\frac{\exp(\mu_+ C_h\left\|\mathbf{x}_n\right\|)}{\left(t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)\right)^{\mu_+}}\right)\exp(-C_0 t^{-0.5\mu_+})\\
&\overset{(3)}{\leq}\left(1+C_1 t^{-\mu_+}\right)\left(1-C_0 t^{-0.5\mu_+}+0.5C_0^2 t^{-\mu_+}\right),\forall t>t'_+\\
&\leq 1-C_0 t^{-0.5\mu_+}\left(1+C_1 t^{-\mu_+}\right)+0.5C_0^2 t^{-\mu_+}\left(1+C_1 t^{-\mu_+}\right)\overset{(4)}{\leq}1,\forall t>t''_+, \qquad (96)
\end{aligned}$$

where in (1), we use $\psi_m(t)\leq 1$ from eq. 91 and using eq. 78, in (2) we used bound on $\mathbf{h}_m$ from eq. 83, in (3) for some large enough $t'_+ > t_+$, we have $\frac{\exp(\mu_+ C_h\|\mathbf{x}_n\|)}{\left(\prod_{r=1}^{m-1}\log^{\circ r}(t)\right)^{\mu_+}}\leq C_1$, and for the second term we used the inequality $e^{-x}\leq 1-x+0.5x^2$ for $x>0$, and (4) holds asymptotically for $t > t''_+$ for large enough $t''_+ > t'_+$ as $C_0 t^{-0.5\mu_+}$ converges slower than $0.5C_0^2 t^{-\mu_+}$ to 0.

Thus, using eq. 96 in eq. 95, $\forall t > \max\left(t_2, t''_+\right)$, we have

$$\Gamma_{m,n}(t)\leq\left[\frac{\eta\kappa(n,t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\right]\left(\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right)-1\right)\mathbf{x}_n^\top \mathbf{r}\left(t\right)\leq 0$$

(b) If $0<\mathbf{x}_n^\top \mathbf{r}\left(t\right)<C_0 t^{-0.5\mu_+}$, then we have the following: $\psi_m(t)\leq 1$ from eq. 91, $\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right)\leq 1$ as $\mathbf{x}_n^\top \mathbf{r}\left(t\right)>0$, and since $\mathbf{w}(t)^\top\mathbf{x}_n\to\infty$, for large enough $t > t'''_+$, $\gamma_n(t)=\left(1+\exp\left(-\mu_+\mathbf{w}(t)^\top\mathbf{x}_n\right)\right)\leq 2$

This gives us, $\left(\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right)-1\right)\mathbf{x}_n^\top \mathbf{r}\left(t\right)\leq \mathbf{x}_n^\top \mathbf{r}\left(t\right)\leq C_0 t^{-0.5\mu_+}$, and using this in eq. 95, $\forall t > \max\left(t_2, t'_+\right)$

$$\Gamma_{m,n}(t)\leq\left[\frac{\eta\kappa(n,t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\right]C_0 t^{-0.5\mu_+}\in L_1.$$

2. Second, if $\mathbf{x}_n^\top \mathbf{r}\left(t\right)\leq 0$, then $\gamma_n(t)=\left(1-\exp(-\mu_-\mathbf{w}(t)^\top\mathbf{x}_n)\right)\in(0,1)$. We again divide into following special cases.

(a) If $\left|\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right|\leq C_0\left(\log^{\circ(m-1)}(t)\right)^{-0.5\tilde{\mu}_-}$, where $\tilde{\mu}_-=\min\left(\mu_-,1\right)$, then we have

$$\begin{aligned}
\Gamma_{m,n}(t)&\leq\left[\frac{1.5\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\right]\left(1-\gamma_n(t)\psi_m\left(t\right)\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right)\right)\left|\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right|\\
&\overset{(1)}{\leq}\left[\frac{1.5\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)}{t\prod_{r=1}^{m-2}\log^{\circ r}\left(t\right)}\right]C_0\left(\log^{\circ(m-1)}(t)\right)^{-1-0.5\tilde{\mu}_-}\in L_1.
\end{aligned}$$

where in (1) we used that $\left(1-\gamma_n(t)\psi_m\left(t\right)\exp\left(-\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right)\right)<1$ and $\left|\mathbf{x}_n^\top \mathbf{r}\left(t\right)\right|\leq C_0\left(\log^{\circ(m-1)}(t)\right)^{-0.5\tilde{\mu}_-}$.

(b) If $\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)<1$, then, from eq. 91

$$\frac{-M\max_n\|\mathbf{x}_n\|C_h}{\log^{\circ(m-1)}(t)}\leq\log\psi_m(t)<\mathbf{x}_n^\top\mathbf{r}(t). \tag{97}$$

In this case, since $\gamma_n(t)=1-\exp(-\mathbf{w}(t)^\top\mathbf{x}_n)<1$, we also have $\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)<1$, and hence $\left(\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)-1\right)\mathbf{x}_n^\top\mathbf{r}(t)>0$. Thus, $\forall t>t_2$, in 95, $\kappa_n(t)=1.5$, and we have

$$\begin{aligned}\Gamma_{m,n}(t)&\leq\left[\frac{1.5\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}(t)}\right]\left(1-\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)\right)\left|\mathbf{x}_n^\top\mathbf{r}(t)\right|\\&\overset{(1)}{\leq}\left[\frac{1.5\eta\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}(t)}\right]\frac{M\max_n\|\mathbf{x}_n\|C_h}{\log^{\circ(m-1)}(t)}\leq\frac{C_2}{t\prod_{r=1}^{m-2}\log^{\circ r}(t)\left(\log^{\circ(m-1)}(t)\right)^2}\in L_1,\end{aligned}$$

where (1) follows from $\left(1-\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)\right)<1$ and the bound on $|\mathbf{x}_n^\top\mathbf{r}(t)|=-\mathbf{x}_n^\top\mathbf{r}(t)$ from eq. 97.

(c) If $\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)>1$, and $\left|\mathbf{x}_n^\top\mathbf{r}(t)\right|>C_0\left(\log^{\circ(m-1)}(t)\right)^{-0.5\tilde{\mu}_-}$, where $\tilde{\mu}_=\min(1,\mu_-)$. Since, $\mathbf{x}_n^\top\mathbf{w}(t)\to\infty$ and $\psi_m(t)\to1$ from eq. 90, for large enough $t'_->t_-$, we have $\forall t>t'_-$, $\psi_m(t)>0.5$ and $\gamma_n(t)=(1-\exp(-\mu_-\mathbf{x}_n^\top\mathbf{w}(t)))>0.5$. Let $\tau>\max(4,t'_-)$ be an arbitrarily large constant. For all $t>\tau$, if $\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)>\tau\geq4$, then $\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)>0.25\tau\geq1$.
On the other hand, if there exists $t>\tau\geq4$, such that $\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)<\tau$, then for some constants C_1,C_2 we have the following

(i) $\exp(-\mathbf{x}_n^\top\mathbf{r}(t))=\exp(|\mathbf{x}_n^\top\mathbf{r}(t)|)\geq\left(1+C_0\left(\log^{\circ(m-1)}(t)\right)^{-0.5\tilde{\mu}_-}\right)$, since $e^x>1+x$ for all x,

(ii) $\psi_m(t)\geq\exp\left(-C_1\left(\log^{\circ(m-1)}(t)\right)^{-1}\right)\geq\left(1-C_1\left(\log^{\circ(m-1)}(t)\right)^{-1}\right)$ from eq. 91 and again using $e^x>1+x$ for all x,

(iii)

$$\begin{aligned}\gamma_n(t)&=\left(1-\left[\frac{\exp(-\mathbf{h}_l(t)^\top\mathbf{x}_n)\exp\left(-\mathbf{x}_n^\top\mathbf{r}(t)\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}(t)}\right]^{\mu_-}\right)\\&\geq\left(1-\left[\frac{\exp(-C_h\|x_n\|)\tau}{t\prod_{r=1}^{m-1}\log^{\circ r}(t)}\right]^{\mu_-}\right)\geq\left(1-\left(C_2\log^{\circ(m-1)}(t)\right)^{-\mu_-}\right),\forall t>t''_-\end{aligned}$$

where the last inequality follows as for large enough $t''_->t'_-$, we have $\frac{\exp(-C_h\|x_n\|)\tau}{t\prod_{r=1}^{m-2}\log^{\circ r}(t)}\leq C_2$.

Using the above inequalities, we have

$$\begin{aligned}
&\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\\
&\geq\left(1+C_0\left(\log^{\circ(m-1)}(t)\right)^{-0.5\tilde{\mu}_-}\right)\left(1-C_1\left(\log^{\circ(m-1)}(t)\right)^{-1}\right)\left(1-C_2\left(\log^{\circ(m-1)}(t)\right)^{-\mu_-}\right)\\
&\overset{(1)}{\geq}1+C_0\left(\log^{\circ(m-1)}(t)\right)^{-0.5\tilde{\mu}_-}-C_1\left(\log^{\circ(m-1)}(t)\right)^{-1}-C_2\left(\log^{\circ(m-1)}(t)\right)^{-\mu_-}\\
&\quad-C_0C_2\left(\log^{\circ(m-1)}(t)\right)^{-\mu_-0.5\tilde{\mu}_-}-C_0C_1\left(\log^{\circ(m-1)}(t)\right)^{-1-0.5\tilde{\mu}_-}\overset{(2)}{\geq}1,\forall t>t_-'''\,,
\end{aligned}\tag{98}$$

where in (1) we dropped the other positive terms, and (2) follows for large enough $t_-'''>t_-''$ as the $C_0\log\left(\log^{\circ(m-1)}(t)\right)^{-0.5\tilde{\mu}_-}$ converges to 0 more slowly than the other negative terms. Finally, using eq. 98 in eq. 95, we have for all $t>\max\left(t_2,\tau,t_\psi,t_-'''\right)$

$$\Gamma_{m,n}(t)\leq\left[\frac{\eta\kappa(n,t)\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}\right)}{t\prod_{r=1}^{m-1}\log^{\circ r}\left(t\right)}\right]\left(1-\gamma_n(t)\psi_m(t)\exp\left(-\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right)\right)\left|\mathbf{x}_n^\top\mathbf{r}\left(t\right)\right|\leq 0\tag{99}$$

Collecting all the terms from the above special cases, and substituting back into eq. 85, we note that all terms are either negative, in L_1, or of the form $f\left(t\right)\left\|\mathbf{r}\left(t\right)\right\|$, where $f\left(t\right)\in L_1$, thus proving the lemma. ■

C.4 Proof of the existence and uniqueness of the solution to eqs. 62-63

We wish to prove that $\forall m\geq 1$:

$$\sum_{n\in\mathcal{S}_m}\exp\left(-\sum_{k=1}^{m}\tilde{\mathbf{w}}_k^\top\mathbf{x}_n\right)\bar{\mathbf{P}}_{m-1}\mathbf{x}_n=\hat{\mathbf{w}}_m\,,\tag{100}$$

such that

$$\mathbf{P}_{m-1}\tilde{\mathbf{w}}_m=0\,\text{and}\,\bar{\mathbf{P}}_m\tilde{\mathbf{w}}_m=0,\tag{101}$$

we have a unique solution. From eq. 101, we can modify eq. 100 to

$$\sum_{n\in\mathcal{S}_m}\exp\left(-\sum_{k=1}^{m}\tilde{\mathbf{w}}_k^\top\bar{\mathbf{P}}_{k-1}\mathbf{x}_n\right)\bar{\mathbf{P}}_{m-1}\mathbf{x}_n=\hat{\mathbf{w}}_m\,,.$$

To prove this, without loss of generality, and with a slight abuse of notation, we will denote $\mathcal{S}_m$ as $\mathcal{S}_1$, $\bar{\mathbf{P}}_{m-1}\mathbf{x}_n$ as $\mathbf{x}_n$ and $\beta_n=\exp\left(-\sum_{k=1}^{m-1}\tilde{\mathbf{w}}_k^\top\bar{\mathbf{P}}_{k-1}\mathbf{x}_n\right)$, so we can write the above equation as

$$\sum_{n\in\mathcal{S}_1}\mathbf{x}_n\beta_n\exp\left(-\mathbf{x}_n^\top\tilde{\mathbf{w}}_1\right)=\hat{\mathbf{w}}_1$$

In the following Lemma 17 we prove this equation $\forall\boldsymbol{\beta}\in\mathbb{R}_{>0}^{|\mathcal{S}_1|}$.

Lemma 17 *$\forall \boldsymbol{\beta} \in \mathbb{R}_{>0}^{|\mathcal{S}_1|}$ we can find a unique $\tilde{\mathbf{w}}$ such that*

$$\sum_{n \in \mathcal{S}_1} \mathbf{x}_n \beta_n \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}_1\right) = \hat{\mathbf{w}}_1 \tag{102}$$

and for $\forall \mathbf{z} \in \mathbb{R}^d$ such that $\mathbf{z}^\top \mathbf{X}_{\mathcal{S}_1} = 0$ we would have $\tilde{\mathbf{w}}_1^\top \mathbf{z} = 0$.

Proof Let $K = \operatorname{rank}(\mathbf{X}_{\mathcal{S}_1})$. Let and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d] \in \mathbb{R}^{d \times d}$ be a set of orthonormal vectors (*i.e.*, $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top \mathbf{U} = \mathbf{I}$) such that $\mathbf{u}_1 = \hat{\mathbf{w}}_1 / \|\hat{\mathbf{w}}_1\|$, and

$$\forall \mathbf{z} \neq 0, \forall n \in \mathcal{S}_1 : \mathbf{z}^\top [\mathbf{u}_1, \ldots, \mathbf{u}_K]^\top \mathbf{x}_n \neq 0\,, \tag{103}$$

while

$$\forall i > K : \forall n \in \mathcal{S}_1 : \mathbf{u}_i^\top \mathbf{x}_n = 0\,. \tag{104}$$

In other words, $\mathbf{u}_1$ is in the direction of $\hat{\mathbf{w}}_1$, $[\mathbf{u}_1, \ldots, \mathbf{u}_K]$ are in the space spanned by the columns of $\mathbf{X}_{\mathcal{S}_1}$, and $[\mathbf{u}_{K+1}, \ldots, \mathbf{u}_d]$ are orthogonal to the columns of $\mathbf{X}_{\mathcal{S}_1}$.

We define $\mathbf{v}_n = \mathbf{U}^\top \mathbf{x}_n$ and $\mathbf{s} = \mathbf{U}^\top \tilde{\mathbf{w}}_1$. Note that $\forall i > K : v_{i,n} = 0 \, \forall n \in \mathcal{S}_1$ from eq. 104, and $\forall i > K : s_i = 0$, since for $\forall \mathbf{z} \in \mathbb{R}^d$ such that $\mathbf{z}^\top \mathbf{X}_{\mathcal{S}_1} = 0$ we would have $\tilde{\mathbf{w}}_1^\top \mathbf{z} = 0$. Lastly, equation 102 becomes

$$\sum_{n \in \mathcal{S}_1} \mathbf{x}_n \beta_n \exp\left(-\sum_{j=1}^{K} s_j v_{j,n}\right) = \hat{\mathbf{w}}_1\,. \tag{105}$$

Multiplying by $\mathbf{U}^\top$ from the left, we obtain

$$\forall i \leq K : \sum_{n \in \mathcal{S}_1} v_{i,n} \beta_n \exp\left(-\sum_{j=1}^{K} s_j v_{j,n}\right) = \mathbf{u}_i^\top \hat{\mathbf{w}}_1\,.$$

Since $\mathbf{u}_1 = \hat{\mathbf{w}}_1 / \|\hat{\mathbf{w}}_1\|$, we have that

$$\forall i \leq K : \sum_{n \in \mathcal{S}_1} v_{i,n} \beta_n \exp\left(-\sum_{j=1}^{K} s_j v_{j,n}\right) = \|\hat{\mathbf{w}}_1\| \, \delta_{i,1}\,. \tag{106}$$

We recall that $v_{1,n} = \hat{\mathbf{w}}_1^\top \mathbf{x}_n / \|\hat{\mathbf{w}}_1\| = 1 / \|\hat{\mathbf{w}}_1\|\,, \forall n \in \mathcal{S}_1$. Given $\{s_j\}_{j=2}^K$, we examine eq. 106 for $i = 1$,

$$\exp\left(-\frac{s_1}{\|\hat{\mathbf{w}}_1\|}\right) \left[\sum_{n \in \mathcal{S}_1} \beta_n \exp\left(-\sum_{j=2}^{K} s_j v_{j,n}\right)\right] = \|\hat{\mathbf{w}}_1\|^2\,.$$

This equation always has the unique solution

$$s_1 = \|\hat{\mathbf{w}}_1\| \log\left[\|\hat{\mathbf{w}}_1\|^{-2} \sum_{n \in \mathcal{S}_1} \beta_n \exp\left(-\sum_{j=2}^{K} s_j v_{j,n}\right)\right]\,, \tag{107}$$

given $\{s_j\}_{j=2}^K$. Next, we similarly examine eq. 106 for $2 \leq i \leq K$ as a function of s_i

$$\sum_{n \in \mathcal{S}_1} \beta_n v_{i,n} \exp\left(-s_1 / \|\hat{\mathbf{w}}_1\| - \sum_{j=2}^{K} s_j v_{j,n}\right) = 0\,. \tag{108}$$

multiplying by $\exp\left(s_1/\left\|\hat{\mathbf{w}}_1\right\|\right)$ we obtain

$$0=\sum_{n\in\mathcal{S}_1}\beta_n v_{i,n}\exp\left(-\sum_{j=2}^{K}s_j v_{j,n}\right)=-\frac{\partial}{\partial s_i}\left[E\left(s_2,\ldots,s_K\right)\right],$$

where we defined

$$E\left(s_2,\ldots,s_K\right)=\sum_{n\in\mathcal{S}_1}\beta_n\exp\left(-\sum_{j=2}^{K}s_j v_{j,n}\right).$$

Therefore, any critical point of $E\left(s_2,\ldots,s_K\right)$ would be a solution of eq. 108 for $2\leq i\leq K$, and substituting this solution into eq. 107 we obtain s_1. Since $\beta_n>0$, $E\left(s_2,\ldots,s_K\right)$ is a convex function, as positive linear combination of convex function (exponential). Therefore, any finite critical point is a global minimum. All that remains is to show that a finite minimum exists and that it is unique.

From the definition of $\mathcal{S}_1$, $\exists\boldsymbol{\alpha}\in\mathbb{R}_{>0}^{|\mathcal{S}_1|}$ such that $\hat{\mathbf{w}}_1=\sum_{n\in\mathcal{S}_1}\alpha_n\mathbf{x}_n$. Multiplying this equation by $\mathbf{U}^\top$ we obtain that $\exists\boldsymbol{\alpha}\in\mathbb{R}_{>0}^{|\mathcal{S}_1|}$ such that $2\leq i\leq K$

$$\sum_{n\in\mathcal{S}_1}v_{i,n}\alpha_n=0\,. \tag{109}$$

Therefore, $\forall\left(s_{2,}\ldots,s_K\right)\neq\mathbf{0}$ we have that

$$\sum_{n\in\mathcal{S}_1}\left(\sum_{j=2}^{K}s_j v_{j,n}\right)\alpha_n=0\,. \tag{110}$$

Recall, from eq. 103 that $\forall\left(s_{2,}\ldots,s_K\right)\neq\mathbf{0},\exists n\in\mathcal{S}_1:\sum_{j=2}^{K}s_j v_{j,n}\neq 0$, and that $\alpha_n>0$. Therefore, eq. 110 implies that $\exists n\in\mathcal{S}_1$ such that $\sum_{j=2}^{K}s_j v_{j,n}>0$ and also $\exists m\in\mathcal{S}_1$ such that $\sum_{j=2}^{K}s_j v_{j,m}<0$.

Thus, in any direction we take a limit in which $|s_i|\to\infty\;\forall 2\leq i\leq K$, we obtain that $E\left(s_2,\ldots,s_K\right)\to\infty$, since at least one exponent in the sum diverge. Since $E\left(s_2,\ldots,s_K\right)$, is a continuous function, it implies it has a finite global minimum. This proves the existence of a finite solution. To prove uniqueness we will show the function is strictly convex, since the hessian is (strictly) positive definite, *i.e.*, that the following expression is strictly positive:

$$\begin{aligned}&\sum_{i=2}^{K}\sum_{k=2}^{K}q_i q_k\frac{\partial}{\partial s_i}\frac{\partial}{\partial s_k}E\left(s_2,\ldots,s_K\right).\\&=\sum_{n\in\mathcal{S}_1}\beta_n\left(\sum_{i=2}^{K}q_i v_{i,n}\right)\left(\sum_{k=2}^{K}q_k v_{k,n}\right)\exp\left(-\sum_{j=2}^{K}s_j v_{j,n}\right)\\&=\sum_{n\in\mathcal{S}_1}\beta_n\left(\sum_{i=2}^{K}q_i v_{i,n}\right)^2\exp\left(-\sum_{j=2}^{K}s_j v_{j,n}\right).\end{aligned}$$

the last expression is indeed strictly positive since $\forall\mathbf{q}\neq\mathbf{0},\exists n\in\mathcal{S}_1:\sum_{j=2}^{K}q_j v_{j,n}\neq 0$, from eq. 103. Thus, there exists a unique solution $\tilde{\mathbf{w}}_1$. ■

C.5 Proof of the existence and uniqueness of the solution to eqs. 64-65

Lemma 18 *For $\forall m > k \geq 1$, the equations*

$$\sum_{n \in \mathcal{S}_m} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{P}_{m-1} \mathbf{x}_n = \sum_{k=1}^{m-1} \left[\sum_{n \in \mathcal{S}_k} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{x}_n \mathbf{x}_n^\top\right] \check{\mathbf{w}}_{k,m} \tag{111}$$

under the constraints

$$\mathbf{P}_{k-1} \check{\mathbf{w}}_{k,m} = 0 \,\text{and}\, \bar{\mathbf{P}}_k \check{\mathbf{w}}_{k,m} = 0 \tag{112}$$

have a unique solution $\check{\mathbf{w}}_{k,m}$.

Proof For this proof we denote $\mathbf{X}_{\mathcal{S}_k}$ as the matrix which columns are $\{\mathbf{x}_n | n \in \mathcal{S}_k\}$, the orthogonal projection matrix $\mathbf{Q}_k = \mathbf{P}_k \bar{\mathbf{P}}_{k-1}$where $\mathbf{Q}_k \mathbf{Q}_m = 0 \; \forall k \neq m$, $\mathbf{Q}_k \bar{\mathbf{P}}_m = 0 \; \forall k < m$, and

$$\forall m : \; \mathbf{I} = \mathbf{P}_m + \bar{\mathbf{P}}_m = \sum_{k=1}^{m} \mathbf{Q}_k + \bar{\mathbf{P}}_m \tag{113}$$

We will write $\check{\mathbf{w}}_{k,m} = \mathbf{W}_{k,m} \mathbf{u}_{k,m}$, where $\mathbf{u}_{k,m} \in \mathbb{R}^{d_k}$ and $\mathbf{W}_{k,m} \in \mathbb{R}^{d \times d_k}$ is a full rank matrix such that $\mathbf{Q}_k \mathbf{W}_{k,m} = \mathbf{W}_{k,m}$, so

$$\check{\mathbf{w}}_{k,m} = \mathbf{Q}_k \check{\mathbf{w}}_{k,m} = \mathbf{Q}_k \mathbf{W}_{k,m} \mathbf{u}_{k,m} \,. \tag{114}$$

and, furthermore,

$$\operatorname{rank}\left[\mathbf{X}_{\mathcal{S}_k}^\top \mathbf{Q}_k \mathbf{W}_{k,m}\right] = \operatorname{rank}\left(\mathbf{X}_{\mathcal{S}_k}^\top \mathbf{Q}_k\right) = d_k \,. \tag{115}$$

Recall that $\forall m : \; \bar{\mathbf{P}}_m \mathbf{P}_m = \mathbf{0}$ and $\forall k \geq 1$, $\forall n \in \mathcal{S}_m \; \bar{\mathbf{P}}_{m+k} \mathbf{x}_n = \mathbf{0}$. Therefore, $\forall \mathbf{v} \in \mathbb{R}^d$, $\mathbf{P}_{k-1} \mathbf{Q}_k \mathbf{v} = \mathbf{0}$, $\bar{\mathbf{P}}_k \mathbf{Q}_k \mathbf{v} = \mathbf{0}$. Thus, $\check{\mathbf{w}}_{k,m}$ eq. 114 implies the constraints in eq. 112 hold.

Next, we prove the existence and uniqueness of the solution $\check{\mathbf{w}}_{k,m}$ for each $k = 1, \ldots, m$ separately. We multiply eq. 111 from the left by the identity matrix, decomposed to orthogonal projection matrices as in eq. 113. Since each matrix projects to an orthogonal subspace, we can solve each product separately.

The product with $\bar{\mathbf{P}}_m$ is equal to zero for both sides of the equation. The product with $\mathbf{Q}_k$ is equal to

$$\sum_{n \in \mathcal{S}_m} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{Q}_k \mathbf{P}_{m-1} \mathbf{x}_n = \left[\sum_{n \in \mathcal{S}_k} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{Q}_k \mathbf{x}_n \mathbf{x}_n^\top\right] \check{\mathbf{w}}_{k,m} \,.$$

Substituting eq. 114, and multiplying by $\mathbf{W}_{k,m}^\top$ from the right, we obtain

$$\sum_{n \in \mathcal{S}_m} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{W}_{k,m}^\top \mathbf{Q}_k \mathbf{P}_{m-1} \mathbf{x}_n = \left[\sum_{n \in \mathcal{S}_k} \exp\left(-\tilde{\mathbf{w}}^\top \mathbf{x}_n\right) \mathbf{W}_{k,m}^\top \mathbf{Q}_k \mathbf{x}_n \mathbf{x}_n^\top \mathbf{Q}_k \mathbf{W}_{k,m}\right] \mathbf{u}_{k,m} \,. \tag{116}$$

Denoting $\mathbf{E}_k \in \mathbb{R}^{|\mathcal{S}_k|\times|\mathcal{S}_k|}$ as diagonal matrix for which $E_{nn,k} = \exp\left(-\frac{1}{2}\tilde{\mathbf{w}}^\top \mathbf{x}_n\right)$, the matrix in the square bracket in the left hand side can be written as

$$\mathbf{W}_{k,m}^\top \mathbf{Q}_k \mathbf{X}_{\mathcal{S}_k} \mathbf{E}_k \mathbf{E}_k \mathbf{X}_{\mathcal{S}_k}^\top \mathbf{Q}_k \mathbf{W}_{k,m}\,. \tag{117}$$

Since $\operatorname{rank}\left(\mathbf{A}\mathbf{A}^\top\right) = \operatorname{rank}\left(\mathbf{A}\right)$ for any matrix $\mathbf{A}$, the rank of this matrix is equal to

$$\operatorname{rank}\left[\mathbf{E}\mathbf{X}_{\mathcal{S}_k}\mathbf{Q}_k\mathbf{W}_{k,m}\right] \overset{(1)}{=} \operatorname{rank}\left[\mathbf{X}_{\mathcal{S}_k}\mathbf{Q}_k\mathbf{W}_{k,m}\right] \overset{(2)}{=} d_k$$

where in (1) we used that $\mathbf{E}_k$ is diagonal and non-zero, and in (2) we used eq. 115. This implies that the $d_k \times d_k$ matrix in eq. 117 is full rank, and so eq. 116 has a unique solution $\mathbf{u}_{k,m}$. Therefore, there exists a unique solution $\check{\mathbf{w}}_{k,m}$. ■

C.6 Proof of Lemma 15

Lemma 15 *Let $\phi(t), h(t), z(t)$ be three functions from $\mathbb{N}$ to $\mathbb{R}_{\geq 0}$, and C_1, C_2, C_3 be three positive constants. Then, if $\sum_{t=1}^{\infty} h(t) \leq C_1 < \infty$, and*

$$\phi^2(t+1) \leq z(t) + h(t)\,\phi(t) + \phi^2(t) \tag{74}$$

we have

$$\phi^2(t+1) \leq C_2 + C_3 \sum_{u=1}^{t} z(u) \tag{75}$$

Proof We define $\psi(t) = z(t) + h(t)$, and start from eq. 74

$$\begin{aligned}
&\phi^2(t+1)\\
&\leq z(t) + h(t)\,\phi(t) + \phi^2(t)\\
&\leq z(t) + h(t)\max\left[1, \phi^2(t)\right] + \phi^2(t)\\
&\leq z(t) + h(t) + h(t)\,\phi^2(t) + \phi^2(t)\\
&\leq \psi(t) + (1+h(t))\,\phi^2(t)\\
&\leq \psi(t) + (1+h(t))\,\psi(t-1) + (1+h(t))\,(1+h(t-1))\,\phi^2(t-1)\\
&\leq \psi(t) + (1+h(t))\,\psi(t-1) + (1+h(t))\,(1+h(t-1))\,\psi(t-2)\\
&+ (1+h(t))\,(1+h(t-1))\,(1+h(t-2))\,\phi^2(t-2)
\end{aligned}$$

we keep iterating eq. 74, until we obtain

$$
\begin{aligned}
&\leq \left[\prod_{m=1}^{t-1}\left(1+h\left(t-m\right)\right)\right]\phi\left(t_1\right)+\sum_{k=0}^{t-t_1}\left[\prod_{m=0}^{k-1}\left(1+h\left(t-m\right)\right)\right]\psi\left(t-k\right)\\
&\leq \left[\exp\left(\sum_{m=1}^{t-1}h\left(t-m\right)\right)\right]\phi\left(t_1\right)+\sum_{k=0}^{t-1}\left[\exp\left(\sum_{m=1}^{k-1}h\left(t-m\right)\right)\right]\psi\left(t-k\right)\\
&\leq \exp\left(C\right)\left[\phi\left(1\right)+\sum_{k=0}^{t-1}\psi\left(t-k\right)\right]\\
&\leq \exp\left(C\right)\left[\phi\left(1\right)+\sum_{u=1}^{t}\psi\left(u\right)\right]\\
&\leq \exp\left(C\right)\left[\phi\left(1\right)+\sum_{u=1}^{t}\left(z\left(u\right)+h\left(u\right)\right)\right]\\
&\leq \exp\left(C\right)\left[\phi\left(1\right)+C+\sum_{u=1}^{t}z\left(u\right)\right]
\end{aligned}
$$

Therefore, the Lemma holds with $C_2=\left(\phi\left(1\right)+C\right)\exp\left(C\right)$ and $C_3=\exp\left(C\right)$. ■

Appendix D. Calculation of convergence rates

In this section we calculate the various rates mentioned in section 3.

D.1 Proof of Theorem 5

From Theorems 4 and 13, we can write $\mathbf{w}\left(t\right)=\hat{\mathbf{w}}\log t+\boldsymbol{\rho}\left(t\right)$, where $\boldsymbol{\rho}\left(t\right)$ has a bounded norm for almost all datasets, while in zero measure case $\boldsymbol{\rho}\left(t\right)$ contains additional $O(\log\log(t))$ components which are orthogonal to the support vectors in $\mathcal{S}_1$, and, asymptotically, have a positive angle with the other support vectors. In this section we first calculate the various convergence rates for the non-degenerate case of Theorem 4, and then write the correction in the zero measure cases, if there is such a correction.

First, we calculated of the normalized weight vector (eq. 8), for almost every dataset:

$$\begin{aligned}
&\frac{\mathbf{w}\left(t\right)}{\left\|\mathbf{w}\left(t\right)\right\|}\\
&=\frac{\boldsymbol{\rho}\left(t\right)+\hat{\mathbf{w}}\log t}{\sqrt{\boldsymbol{\rho}\left(t\right)^{\top}\boldsymbol{\rho}\left(t\right)+\hat{\mathbf{w}}^{\top}\hat{\mathbf{w}}\log^{2}t+2\boldsymbol{\rho}\left(t\right)^{\top}\hat{\mathbf{w}}\log t}}\\
&=\frac{\boldsymbol{\rho}\left(t\right)/\log t+\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|\sqrt{1+2\boldsymbol{\rho}\left(t\right)^{\top}\hat{\mathbf{w}}/\left(\left\|\hat{\mathbf{w}}\right\|^{2}\log t\right)+\left\|\boldsymbol{\rho}\left(t\right)\right\|^{2}/\left(\left\|\hat{\mathbf{w}}\right\|^{2}\log^{2}t\right)}}\\
&=\frac{1}{\left\|\hat{\mathbf{w}}\right\|}\left(\boldsymbol{\rho}\left(t\right)\frac{1}{\log t}+\hat{\mathbf{w}}\right)\left[1-\frac{\boldsymbol{\rho}\left(t\right)^{\top}\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|^{2}\log t}+\left[\frac{3}{2}\left(\frac{\boldsymbol{\rho}\left(t\right)^{\top}\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|^{2}}\right)^{2}-\frac{\left\|\boldsymbol{\rho}\left(t\right)\right\|^{2}}{2\left\|\hat{\mathbf{w}}\right\|^{2}}\right]\frac{1}{\log^{2}t}+O\left(\frac{1}{\log^{3}t}\right)\right] \qquad (118)\\
&=\frac{\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|}+\left(\frac{\boldsymbol{\rho}\left(t\right)}{\left\|\hat{\mathbf{w}}\right\|}-\frac{\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|}\frac{\boldsymbol{\rho}\left(t\right)^{\top}\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|^{2}}\right)\frac{1}{\log t}+O\left(\frac{1}{\log^{2}t}\right)\\
&=\frac{\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|}+\left(\mathbf{I}-\frac{\hat{\mathbf{w}}\hat{\mathbf{w}}^{\top}}{\left\|\hat{\mathbf{w}}\right\|^{2}}\right)\frac{\boldsymbol{\rho}\left(t\right)}{\left\|\hat{\mathbf{w}}\right\|}\frac{1}{\log t}+O\left(\frac{1}{\log^{2}t}\right),
\end{aligned}$$

where to obtain eq. 118 we used $\frac{1}{\sqrt{1+x}}=1-\frac{1}{2}x+\frac{3}{4}x^{2}+O\left(x^{3}\right)$, and in the last line we used the fact that $\boldsymbol{\rho}\left(t\right)$ has a bounded norm for almost every dataset. Thus, in this case

$$\left\|\frac{\mathbf{w}\left(t\right)}{\left\|\mathbf{w}\left(t\right)\right\|}-\frac{\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|}\right\|=O\left(\frac{1}{\log t}\right).$$

For the measure zero cases, we instead have from eq. 61, $\mathbf{w}(t)=\sum_{m=1}^{M}\hat{\mathbf{w}}\log^{\circ m}(t)+\rho(t)$, where $\|\rho(t)\|$ is bounded (Theorem 3). Let $\tilde{\rho}(t)=\sum_{m=2}^{M}\hat{\mathbf{w}}\log^{\circ m}(t)+\rho(t)$, such that $\mathbf{w}(t)=\hat{\mathbf{w}}\log(t)+\tilde{\rho}(t)$ with $\tilde{\rho}(t)=O(\log\log(t))$. Repeating the same calculations as above, we have for the degenerate cases,

$$\left\|\frac{\mathbf{w}\left(t\right)}{\left\|\mathbf{w}\left(t\right)\right\|}-\frac{\hat{\mathbf{w}}}{\left\|\hat{\mathbf{w}}\right\|}\right\|=O\left(\frac{\log\log t}{\log t}\right)$$

Next, we use eq. 118 to calculate the angle (eq. 9)

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

for almost every dataset. Thus, in this case

$$\frac{\mathbf{w}(t)^\top \hat{\mathbf{w}}}{\|\mathbf{w}(t)\| \|\hat{\mathbf{w}}\|}=O\left(\frac{1}{\log^2 t}\right)$$

Repeating the same calculation for the measure zero case, we have instead

$$\frac{\mathbf{w}(t)^\top \hat{\mathbf{w}}}{\|\mathbf{w}(t)\| \|\hat{\mathbf{w}}\|}=O\left(\left(\frac{\log\log t}{\log t}\right)^2\right)$$

Next, we calculate the margin (eq. 10)

$$\begin{aligned}
&\min_n \frac{\mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|}-\frac{1}{\|\hat{\mathbf{w}}\|} \\
&=\min_n \mathbf{x}_n^\top\left[\left(\frac{\boldsymbol{\rho}(t)}{\|\hat{\mathbf{w}}\|}-\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}\frac{\boldsymbol{\rho}(t)^\top \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|^2}\right)\frac{1}{\log t}+O\left(\frac{1}{\log^2 t}\right)\right] \\
&=\frac{1}{\|\hat{\mathbf{w}}\|}\left(\min_n \mathbf{x}_n^\top \boldsymbol{\rho}(t)-\frac{\boldsymbol{\rho}(t)^\top \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|^2}\right)\frac{1}{\log t}+O\left(\frac{1}{\log^2 t}\right) \qquad (119)
\end{aligned}$$

for almost every dataset, where in eq. 119 we used eq. 19. Interestingly the measure zero case has a similar convergence rate, since after a sufficient number of iterations, the $O(\log\log(t))$ correction is orthogonal to $\mathbf{x}_k$, where $k=\mathrm{argmin}_n \mathbf{x}_n^\top \mathbf{w}(t)$. Thus, for all datasets,

$$\min_n \mathbf{x}_n^\top \mathbf{w}(t)-\frac{1}{\|\hat{\mathbf{w}}\|}=O\left(\frac{1}{\log t}\right) \qquad (120)$$

Calculation of the training loss (eq. 11):

$$\begin{aligned}
\mathcal{L}(\mathbf{w}(t)) &\leq \sum_{n=1}^N \left(1+\exp\left(-\mu_+ \mathbf{w}(t)^\top \mathbf{x}_n\right)\right)\exp\left(-\mathbf{w}(t)^\top \mathbf{x}_n\right) \\
&=\sum_{n=1}^N \left(1+\exp\left(-\mu_+ (\boldsymbol{\rho}(t)+\hat{\mathbf{w}}\log t)^\top \mathbf{x}_n\right)\right)\exp\left(-(\boldsymbol{\rho}(t)+\hat{\mathbf{w}}\log t)^\top \mathbf{x}_n\right) \\
&=\sum_{n=1}^N \left(1+t^{-\mu_+ \hat{\mathbf{w}}^\top \mathbf{x}_n}\exp\left(-\mu_+ \boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)\right)\exp\left(-\boldsymbol{\rho}(t)^\top \mathbf{x}_n\right)t^{-\hat{\mathbf{w}}^\top \mathbf{x}_n} \\
&=\frac{1}{t}\sum_{n\in\mathcal{S}} e^{-\boldsymbol{\rho}(t)^\top \mathbf{x}_n}+O\left(t^{-\max(\theta,1+\mu_+)}\right).
\end{aligned}$$

Thus, for all datasets $\mathcal{L}\left(\mathbf{w}\left(t\right)\right)=O(t^{-1})$. Note that the zero measure case has the same behavior, since after a sufficient number of iterations, the $O(\log\log(t))$ correction has a non-negative angle with all the support vectors.

Next, we give an example demonstrating the bounds above, for the non-degenerate case, are strict. Consider optimization with and exponential loss $\ell\left(u\right)=e^{-u}$, and a single data point $\mathbf{x}=(1,0)$. In this case $\hat{\mathbf{w}}=(1,0)$ and $\|\hat{\mathbf{w}}\|=1$. We take the limit $\eta\to 0$, and obtain the continuous time version of GD:

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

We can analytically integrate these equations to obtain

$$w_1\left(t\right)=\log\left(t+\exp\left(w_1\left(0\right)\right)\right)\ ;\ w_2\left(t\right)=w_2\left(0\right).$$

Using this example with $w_2\left(0\right)>0$, it is easy to see that the above upper bounds are strict in the non-degenerate case. ■

D.2 Validation error lower bound

Lastly, recall that $\mathcal{V}$ is a set of indices for validation set samples. We calculate of the validation loss for logistic loss, if the error of the L_2 max margin vector has some classification errors on the validation, *i.e.,* $\exists k\in\mathcal{V}:\ \hat{\mathbf{w}}^\top\mathbf{x}_k<0$*:*

$$\begin{aligned}
\mathcal{L}_{\text{val}}\left(\mathbf{w}\left(t\right)\right)&=\sum_{n\in\mathcal{V}}\log\left(1+\exp\left(-\mathbf{w}\left(t\right)^\top\mathbf{x}_n\right)\right)\\
&\geq\log\left(1+\exp\left(-\mathbf{w}\left(t\right)^\top\mathbf{x}_k\right)\right)\\
&=\log\left(1+\exp\left(-\left(\boldsymbol{\rho}\left(t\right)+\hat{\mathbf{w}}\log t\right)^\top\mathbf{x}_k\right)\right)\\
&=\log\left(\exp\left(-\left(\boldsymbol{\rho}\left(t\right)+\hat{\mathbf{w}}\log t\right)^\top\mathbf{x}_k\right)\left(1+\exp\left(\left(\boldsymbol{\rho}\left(t\right)+\hat{\mathbf{w}}\log t\right)^\top\mathbf{x}_k\right)\right)\right)\\
&\geq-\left(\boldsymbol{\rho}\left(t\right)+\hat{\mathbf{w}}\log t\right)^\top\mathbf{x}_k+\log\left(1+\exp\left(\left(\boldsymbol{\rho}\left(t\right)+\hat{\mathbf{w}}\log t\right)^\top\mathbf{x}_k\right)\right)\\
&\geq-\log t\hat{\mathbf{w}}^\top\mathbf{x}_k+\boldsymbol{\rho}\left(t\right)^\top\mathbf{x}_k
\end{aligned}$$

Thus, for all datasets $\mathcal{L}_{\text{val}}\left(\mathbf{w}\left(t\right)\right)=\Omega(\log(t))$.

Appendix E. Softmax output with cross-entropy loss

We examine multiclass classification. In the case the labels are the class index $y_n\in\{1,\ldots,K\}$ and we have a weight matrix $\mathbf{W}\in\mathbb{R}^{K\times d}$ with $\mathbf{w}_k$ being the k-th row of $\mathbf{W}$.

Furthermore, we define $\mathbf{w}=\text{vec}\left(\mathbf{W}^\top\right)$, a basis vector $\mathbf{e}_k\in\mathbb{R}^K$ so that$\left(\mathbf{e}_k\right)_i=\delta_{ki}$, and the matrix $\mathbf{A}_k\in\mathbb{R}^{dK\times d}$ so that $\mathbf{A}_k=\mathbf{e}_k\otimes\mathbf{I}_d$, where $\otimes$ is the Kronecker product and $\mathbf{I}_d$ is the d-dimension identity matrix. Note that $\mathbf{A}_k^\top\mathbf{w}=\mathbf{w}_k$.

Consider the cross entropy loss with softmax output

$$\mathcal{L}\left(\mathbf{W}\right)=-\sum_{n=1}^{N}\log\left(\frac{\exp\left(\mathbf{w}_{y_n}^\top\mathbf{x}_n\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^\top\mathbf{x}_n\right)}\right)$$

Using our notation, this loss can be re-written as

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= -\sum_{n=1}^{N}\log\left(\frac{\exp\left(\mathbf{w}^\top\mathbf{A}_{y_n}\mathbf{x}_n\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{w}^\top\mathbf{A}_k\mathbf{x}_n\right)}\right)\\ &= \sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}\exp\left(\mathbf{w}^\top\left(\mathbf{A}_k-\mathbf{A}_{y_n}\right)\mathbf{x}_n\right)\right)\end{aligned} \tag{121}$$

Therefore

$$\begin{aligned}\nabla\mathcal{L}(\mathbf{w}) &= \sum_{n=1}^{N}\frac{\sum_{k=1}^{K}\exp\left(\mathbf{w}^\top\left(\mathbf{A}_k-\mathbf{A}_{y_n}\right)\mathbf{x}_n\right)\left(\mathbf{A}_k-\mathbf{A}_{y_n}\right)\mathbf{x}_n}{\sum_{r=1}^{K}\exp\left(\mathbf{w}^\top\left(\mathbf{A}_r-\mathbf{A}_{y_n}\right)\mathbf{x}_n\right)}\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{1}{\sum_{r=1}^{K}\exp\left(\mathbf{w}^\top\left(\mathbf{A}_r-\mathbf{A}_k\right)\mathbf{x}_n\right)}\left(\mathbf{A}_k-\mathbf{A}_{y_n}\right)\mathbf{x}_n\,.\end{aligned}$$

If, again, we make the assumption that the data is linearly separable, *i.e.*, in our notation

Assumption 4 $\exists\mathbf{w}_*$ *such that* $\mathbf{w}_*^\top\left(\mathbf{A}_k-\mathbf{A}_{y_n}\right)\mathbf{x}_n<0\;\forall k\neq y_n$.

then the expression

$$\mathbf{w}_*^\top\nabla\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\mathbf{w}_*^\top\left(\mathbf{A}_k-\mathbf{A}_{y_n}\right)\mathbf{x}_n}{\sum_{r=1}^{K}\exp\left(\mathbf{w}^\top\left(\mathbf{A}_r-\mathbf{A}_k\right)\mathbf{x}_n\right)}\,.$$

is strictly negative for any finite $\mathbf{w}$. However, from Lemma 10, in gradient descent with an appropriately small learning rate, we have that $\nabla L(\mathbf{w}(t))\to\mathbf{0}$. This implies that: $\|\mathbf{w}(t)\|\to\infty$, and $\forall k\neq y_n,\exists r:\;\mathbf{w}(t)^\top\left(\mathbf{A}_r-\mathbf{A}_k\right)\mathbf{x}_n\to\infty$, which implies $\forall k\neq y_n,\max_k\mathbf{w}(t)^\top\left(\mathbf{A}_k-\mathbf{A}_{y_n}\right)\mathbf{x}_n\to-\infty$. Examining the loss (eq. 121) we find that $\mathcal{L}(\mathbf{w}(t))\to\mathbf{0}$ in this case. Thus, we arrive to an equivalent Lemma to Lemma 1, for this case:

Lemma 19 *Let* $\mathbf{w}(t)$ *be the iterates of gradient descent (eq. 2) with an appropriately small learning rate, for cross-entropy loss operating on a softmax output, under the assumption of strict linear separability (Assumption 4), then: (1)* $\lim_{t\to\infty}\mathcal{L}(\mathbf{w}(t))=0$*, (2)* $\lim_{t\to\infty}\|\mathbf{w}(t)\|=\infty$*, and (3)* $\forall n,k\neq y_n:\;\lim_{t\to\infty}\mathbf{w}(t)^\top\left(\mathbf{A}_{y_n}-\mathbf{A}_k\right)\mathbf{x}_n=\infty$.

Using Lemma 10 and Lemma 19, we prove the following Theorem (equivalent to Theorem 3) in the next section:

Theorem 7 *For almost all multiclass datasets (*i.e.*, except for a measure zero) which are linearly separable (i.e. the constraints in eq. 15 below are feasible), any starting point* $\mathbf{w}(0)$ *and any small enough stepsize, the iterates of gradient descent on 13 will behave as:*

$$\mathbf{w}_k(t)=\hat{\mathbf{w}}_k\log(t)+\boldsymbol{\rho}_k(t), \tag{14}$$

where the residual $\boldsymbol{\rho}_k(t)$ *is bounded and* $\hat{\mathbf{w}}_k$ *is the solution of the K-class SVM:*

$$\text{argmin}_{\mathbf{w}_1,\ldots,\mathbf{w}_k}\sum_{k=1}^{K}||\mathbf{w}_k||^2\ \textit{s.t.}\ \forall n,\forall k\neq y_n:\mathbf{w}_{y_n}^\top\mathbf{x}_n\geq\mathbf{w}_k^\top\mathbf{x}_n+1. \tag{15}$$

E.1 Notations and Definitions

To prove Theorem 7 we require additional notation. we define $\tilde{\mathbf{x}}_{n,k} \triangleq (\mathbf{A}_{y_n} - \mathbf{A}_k)\mathbf{x}_n$. Using this notation, we can re-write eq. 15 (K-class SVM) as

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 \text{ s.t. } \forall n, \forall k \neq y_n : \mathbf{w}^\top \tilde{\mathbf{x}}_{n,k} \geq 1 \tag{122}$$

From the KKT optimality conditions, we have for some $\alpha_{n,k} \geq 0$,

$$\hat{\mathbf{w}} = \sum_{n=1}^{N}\sum_{k=1}^{K} \alpha_{n,k}\tilde{\mathbf{x}}_{n,k}\mathbf{1}_{\{n\in\mathcal{S}_k\}} \tag{123}$$

In addition, for each of the K classes, we define $\mathcal{S}_k = \arg\min_n(\hat{\mathbf{w}}_{y_n} - \hat{\mathbf{w}}_k)^\top \mathbf{x}_n$ (the k'th class support vectors).
Using this definition, we define $\mathbf{X}_{\mathcal{S}_k} \in \mathcal{R}^{dK\times|S_k|}$ as the matrix which columns are $\tilde{\mathbf{x}}_{n,k},\ \forall n \in \mathcal{S}_k$.
We also define $\mathcal{S} \triangleq \bigcup_{k=1}^{K} \mathcal{S}_k$ and $\tilde{\mathbf{X}}_{\mathcal{S}} \triangleq \bigcup_{k=1}^{K} \mathbf{X}_{\mathcal{S}_k}$.
We recall that we defined $\mathbf{W} \in \mathbb{R}^{K\times d}$ with $\mathbf{w}_k$ being the k-th row of $\mathbf{W}$ and $\mathbf{w} = \mathrm{vec}(\mathbf{W}^\top)$. Similarly, we define:
1. $\hat{\mathbf{W}} \in \mathbb{R}^{K\times d}$ with $\hat{\mathbf{w}}_k$ being the k-th row of $\hat{\mathbf{W}}$
2. $\mathbf{P} \in \mathbb{R}^{K\times d}$ with $\boldsymbol{\rho}_k$ being the k-th row of $\mathbf{P}$
3. $\tilde{\mathbf{W}} \in \mathbb{R}^{K\times d}$ with $\tilde{\mathbf{w}}_k$ being the k-th row of $\tilde{\mathbf{W}}$
and $\hat{\mathbf{w}} = \mathrm{vec}(\hat{\mathbf{W}}^\top), \boldsymbol{\rho} = \mathrm{vec}(\mathbf{P}^\top), \tilde{\mathbf{w}} = \mathrm{vec}(\tilde{\mathbf{W}}^\top)$.
Using our notations, eq. 14 can be re-written as $\mathbf{w} = \hat{\mathbf{w}}\log(t) + \boldsymbol{\rho}(t)$ when $\boldsymbol{\rho}(t)$ is bounded.
For any solution $\mathbf{w}(t)$, we define

$$\mathbf{r}(t) = \mathbf{w}(t) - \hat{\mathbf{w}}\log t - \tilde{\mathbf{w}}, \tag{124}$$

where $\hat{\mathbf{w}}$ is the concatenation of $\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_k$ which are the K-class SVM solution, so

$$\forall k,\ \forall n \in \mathcal{S}_k : \tilde{\mathbf{x}}_{n,k}^\top \hat{\mathbf{w}} = 1\ ;\ \theta = \min_k \left[\min_{n\notin\mathcal{S}_k} \tilde{\mathbf{x}}_{n,k}^\top \hat{\mathbf{w}}\right] > 1 \tag{125}$$

and $\tilde{\mathbf{w}}$ satisfies the equation:

$$\forall k,\ \forall n \in \mathcal{S}_k\ :\ \eta\exp((\tilde{\mathbf{w}}_k - \tilde{\mathbf{w}}_{y_n})^\top \mathbf{x}_n) = \alpha_{n,k} \tag{126}$$

This equation has a unique solution for almost every data set according to Lemma 12.
For each of the K classes, we define $\mathbf{P}_1^k \in \mathcal{R}^{d\times d}$ as the orthogonal projection matrix to the subspace spanned by the support vector of the k'th class, and $\bar{\mathbf{P}}_1^k = \mathbf{I} - \mathbf{P}_1^k$ as the complementary projection.
Finally, we define $\mathbf{P}_1 \in \mathcal{R}^{Kd\times Kd}$ and $\bar{\mathbf{P}}_1 \in \mathcal{R}^{Kd\times Kd}$ as follows:

$$\mathbf{P}_1 = \mathrm{diag}(\mathbf{P}_1^1, \mathbf{P}_1^2, ..., \mathbf{P}_1^K)\ ,\ \bar{\mathbf{P}}_1 = \mathrm{diag}(\bar{\mathbf{P}}_1^1, \bar{\mathbf{P}}_1^2, ..., \bar{\mathbf{P}}_1^K)$$

$$(\mathbf{P}_1 + \bar{\mathbf{P}}_1 = \mathbf{I} \in \mathcal{R}^{Kd\times Kd})$$

In the following section we will also use $\mathbf{1}_{\{A\}}$, the indicator function, which is 1 if A is satisfied and 0 otherwise.

E.2 Auxiliary Lemma

Lemma 20 *We have*

$$\exists C_1, t_1 \ : \ \forall t > t_1 : (\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-\theta} + C_2 t^{-2} \tag{127}$$

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

$$||\mathbf{P}_1 \mathbf{r}(t)|| > \epsilon_1 \tag{128}$$

then we can improve this bound to

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq -C_3 t^{-1} < 0 \tag{129}$$

We prove the Lemma below, in appendix section E.4

E.3 Proof of Theorem 7

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

$$\|\mathbf{r}(t+1)\|^2 = \|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 + 2(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) + \|\mathbf{r}(t)\|^2 \tag{130}$$

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

$$\begin{aligned}
&\|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 \\
&\overset{(1)}{=} \|\mathbf{w}(t+1) - \hat{\mathbf{w}}\log(t+1) - \tilde{\mathbf{w}} - \mathbf{w}(t) + \hat{\mathbf{w}}\log(t) + \tilde{\mathbf{w}}\|^2 \\
&\overset{(2)}{=} \|-\eta\nabla\mathcal{L}(\mathbf{w}(t)) - \hat{\mathbf{w}}[\log(t+1) - \log(t)]\|^2 \\
&= \eta^2\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \|\hat{\mathbf{w}}\|^2\log^2\left(1+t^{-1}\right) + 2\eta\hat{\mathbf{w}}^\top\nabla\mathcal{L}(\mathbf{w}(t))\log\left(1+t^{-1}\right) \\
&\overset{(3)}{\leq} \eta^2\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \|\hat{\mathbf{w}}\|^2 t^{-2},
\end{aligned} \tag{131}$$

where in (1) we used eq. 124, in (2) we used eq 2.2, and in (3) we used $\forall x > 0 : x \geq \log(1+x) > 0$, and also that

$$\hat{\mathbf{w}}^\top\nabla\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\hat{\mathbf{w}}^\top(\mathbf{A}_{y_n} - \mathbf{A}_k)\mathbf{x}_n}{\sum_{r=1}^{K}\exp(\mathbf{w}^\top(\mathbf{A}_r - \mathbf{A}_k)\mathbf{x}_n)} < 0 \tag{132}$$

since $\hat{\mathbf{w}}^\top(\mathbf{A}_r - \mathbf{A}_k)\mathbf{x}_n = (\hat{\mathbf{w}}_r - \hat{\mathbf{w}}_{y_n})\mathbf{x}_n < 0, \forall k \neq y_n$ (we recall that $\hat{\mathbf{w}}_k$ is the K-class SVM solution).
Also, from Lemma 10 we know that

$$\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 = o(1) \text{ and } \sum_{t=0}^{\infty}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 < \infty. \tag{133}$$

Substituting eq. 133 into eq. 131, and recalling that a $t^{-\nu}$ power series converges for any $\nu > 1$, we can find C_0 such that

$$\|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 = o(1) \text{ and } \sum_{t=0}^{\infty}\|\mathbf{r}(t+1) - \mathbf{r}(t)\|^2 = C_0 < \infty. \tag{134}$$

Note that this equation also implies that $\forall \epsilon_0$

$$\exists t_0 : \forall t > t_0 : |\|\mathbf{r}(t+1)\| - \|\mathbf{r}(t)\|| < \epsilon_0 \,. \tag{135}$$

Next, we would like to bound the second term in eq. 130. From eq. 127 in Lemma 20, we can find t_1, C_1 such that $\forall t > t_1$:

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-\theta} + C_2 t^{-2} \tag{136}$$

Thus, by combining eqs. 136 and 134 into eq. 130, we find:

$$\begin{aligned}
&||\mathbf{r}(t)||^2 - ||\mathbf{r}(t_1)||^2 \\
&= \sum_{u=t_1}^{t-1} \left[||\mathbf{r}(u+1)||^2 - ||\mathbf{r}(u)||^2 \right] \\
&\leq C_0 + 2 \sum_{u=t_1}^{t-1} \left[C_1 u^{-\theta} + C_2 u^{-2} \right]
\end{aligned}$$

which is bounded, since $\theta > 1$ (eq. 125). Therefore, $||\mathbf{r}(t)||$ is bounded.

E.4 Proof of Lemma 20

Lemma 20 *We have*

$$\exists C_1, t_1 \; : \; \forall t > t_1 : (\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-\theta} + C_2 t^{-2} \tag{127}$$

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

$$||\mathbf{P}_1 \mathbf{r}(t)|| > \epsilon_1 \tag{128}$$

then we can improve this bound to

$$(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq -C_3 t^{-1} < 0 \tag{129}$$

We wish to bound $(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t)$. First, we recall we defined $\tilde{\mathbf{x}}_{n,k} \triangleq (\mathbf{A}_{y_n} - \mathbf{A}_k)\mathbf{x}_n$.

$$\begin{aligned}
&(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) = (-\eta \nabla \mathcal{L}(\mathbf{w}(t)) - \hat{\mathbf{w}}[\log(t+1) - \log(t)])^\top \mathbf{r}(t) \\
&= \left(\eta \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,k}) \tilde{\mathbf{x}}_{n,k}}{\sum_{r=1}^{K} \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r})} - \hat{\mathbf{w}} \log(1 + t^{-1}) \right)^\top \mathbf{r}(t) \\
&= \hat{\mathbf{w}}^\top \mathbf{r}(t)[t^{-1} - \log(1 + t^{-1})] \qquad (137) \\
&+ \eta \sum_{n=1}^{N} \sum_{k=1}^{K} \left[\frac{\exp\left(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t)}{\sum_{r=1}^{K} \exp\left(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r}\right)} - t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{n \in \mathcal{S}_k\}} \right], \qquad (138)
\end{aligned}$$

where in the last line we used eqs. 123 and 126 to obtain

$$\hat{\mathbf{w}} = \eta \sum_{n=1}^{N} \sum_{k=1}^{K} \alpha_{n,k} \tilde{\mathbf{x}}_{n,k} \mathbf{1}_{\{n \in \mathcal{S}_k\}} = \eta \sum_{n=1}^{N} \sum_{k=1}^{K} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k})\right) \tilde{\mathbf{x}}_{n,k} \mathbf{1}_{\{n \in \mathcal{S}_k\}},$$

where $\mathbf{1}_{\{A\}}$ is the indicator function which is 1 if A is satisfied and 0 otherwise.

The first term can be upper bounded by

$$\begin{aligned}
&\hat{\mathbf{w}}^\top \mathbf{r}\left(t\right)\left[t^{-1}-\log\left(1+t^{-1}\right)\right] \\
&\leq \max\left[\hat{\mathbf{w}}^\top \mathbf{r}\left(t\right),0\right]\left[t^{-1}-\log\left(1+t^{-1}\right)\right] \\
&\overset{(1)}{\leq} \max\left[\hat{\mathbf{w}}^\top \mathbf{P}_1 \mathbf{r}\left(t\right),0\right] t^{-2} \\
&\overset{(2)}{\leq} \begin{cases} \|\hat{\mathbf{w}}\|\epsilon_1 t^{-2} & \text{, if } \|\mathbf{P}_1\mathbf{r}\left(t\right)\| \leq \epsilon_1 \\ o\left(t^{-1}\right) & \text{, if } \|\mathbf{P}_1\mathbf{r}\left(t\right)\| > \epsilon_1 \end{cases}
\end{aligned} \tag{139}$$

where in (1) we used that $\mathbf{P}_2\hat{\mathbf{w}}=0$, and in (2) we used that $\hat{\mathbf{w}}^\top \mathbf{r}\left(t\right)=o\left(t\right)$, since

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

where in the last line we used that $\nabla\mathcal{L}\left(\mathbf{w}\left(t\right)\right)=o\left(1\right)$, from Lemma 10.
Next, we wish to upper bound the second term in eq. 137:

$$\eta\sum_{n=1}^{N}\sum_{k=1}^{K}\left[\frac{\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,k}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)}{\sum_{r=1}^{K}\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,r}\right)}-t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)\mathbf{1}_{\{n\in\mathcal{S}_k\}}\right] \tag{140}$$

We examine each term n in the sum:

$$\sum_{k=1}^{K}\left[\frac{\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,k}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)}{\sum_{r=1}^{K}\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,r}\right)}-t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)\mathbf{1}_{\{n\in\mathcal{S}_k\}}\right]$$

$$=\sum_{k=1}^{K}\left[\frac{\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,k}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)}{1+\sum\limits_{\substack{r=1\\r\neq y_n}}^{K}\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,r}\right)}-t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)\mathbf{1}_{\{n\in\mathcal{S}_k\}}\right]$$

$$\overset{(1)}{\leq}\sum_{k=1}^{K}\left(\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,k}\right)-t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\mathbf{1}_{\{n\in\mathcal{S}_k\}}\right)\mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)\geq 0\}}\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)$$

$$+\sum_{k=1}^{K}\left(\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,k}\right)\left(1-\sum_{\substack{r=1\\r\neq y_n}}^{K}\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,r}\right)\right)\right.$$

$$\left.-t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\mathbf{1}_{\{n\in\mathcal{S}_k\}}\right)\mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)<0\}}\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)$$

$$=\sum_{k=1}^{K}\left(\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,k}\right)-t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\mathbf{1}_{\{n\in\mathcal{S}_k\}}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)$$

$$-\sum_{k=1}^{K}\sum_{\substack{r=1\\r\neq y_n}}^{K}\exp\left(-\mathbf{w}(t)^\top(\tilde{\mathbf{x}}_{n,k}+\tilde{\mathbf{x}}_{n,r})\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)\mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)<0\}}$$

$$\overset{(2)}{\leq}\sum_{k=1}^{K}\left(\exp\left(-\mathbf{w}(t)^\top\tilde{\mathbf{x}}_{n,k}\right)-t^{-1}\exp\left(-\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\mathbf{1}_{\{n\in\mathcal{S}_k\}}\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)$$

$$-K^2\exp\left(-\mathbf{w}(t)^\top(\tilde{\mathbf{x}}_{n,k_1}+\tilde{\mathbf{x}}_{n,r_1})\right)\tilde{\mathbf{x}}_{n,k_1}^\top\mathbf{r}(t)\mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top\mathbf{r}(t)<0\}}, \tag{141}$$

where in (1) we used $\forall x\geq 0:\ 1-x\leq\frac{1}{1+x}\leq 1$ and in (2) we defined:

$$(k_1,r_1)=\underset{k,r}{\operatorname{argmax}}\left|\exp\left(-\mathbf{w}(t)^\top(\tilde{\mathbf{x}}_{n,k}+\tilde{\mathbf{x}}_{n,r})\right)\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)\mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}(t)<0\}}\right|$$

Recalling that $\mathbf{w}(t) = \hat{\mathbf{w}}\log(t) + \tilde{\mathbf{w}} + \mathbf{r}(t)$, eq. 141 can be upper bounded by

$$\sum_{k=1}^{K} t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \geq 0 \ , \ n \notin \mathcal{S}_k\}}$$
$$+ \sum_{k=1}^{K} t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) - 1\right] \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \geq 0 \ , \ n \in \mathcal{S}_k\}}$$
$$+ \sum_{k=1}^{K} t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0 \ , \ n \notin \mathcal{S}_k\}}$$
$$+ \sum_{k=1}^{K} t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) - 1\right] \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0 \ , \ n \in \mathcal{S}_k\}}$$
$$- K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) < 0\}}$$
$$\overset{(1)}{\leq} K t^{-\theta} \exp\left(-\min_{n,k} \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) + \phi(t), \tag{142}$$

where in (1) we used $xe^{-x} < 1, \ \forall x: \ (e^{-x} - 1)x < 0$, $\theta = \min_k \left[\min_{n \notin \mathcal{S}_k} \tilde{\mathbf{x}}_{n,k}^\top \hat{\mathbf{w}}\right] > 1$ (eq. 125) and denoted:

$$\phi(t) = \sum_{k=1}^{K} t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0, \ n \notin \mathcal{S}_k\}}$$
$$+ \sum_{k=1}^{K} t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) - 1\right] \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0, n \in \mathcal{S}_k\}}$$
$$- K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) < 0\}}.$$

We use the fact that $\forall x: \ (e^{-x} - 1)x < 0$ and therefore $\forall (n,k)$:

$$t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0\}} < 0$$
$$t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) - 1\right] \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0\}} < 0, \tag{143}$$

to show that $\phi(t)$ is strictly negative. If $\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r} \geq 0$ then from the last two equations:

$$\phi(t) = \sum_{k=1}^{K} t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0, \ n \notin \mathcal{S}_k\}}$$
$$+ \sum_{k=1}^{K} t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) - 1\right] \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) < 0, \ n \in \mathcal{S}_k\}} < 0 \tag{144}$$

If $\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r} < 0$ then we note that $-\tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t) \leq -\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)$ since:
1. If $\tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t) \geq 0$ then this is immediate since $-\tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t) \leq 0 \leq -\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)$.
2. If $\tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t) < 0$ then from (k_1, r_1) definition:

$$\left|\exp\left(-\mathbf{w}(t)^\top (\tilde{\mathbf{x}}_{n,k_1} + \tilde{\mathbf{x}}_{n,r_1})\right) \tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t)\right| \leq \left|\exp\left(-\mathbf{w}(t)^\top (\tilde{\mathbf{x}}_{n,k_1} + \tilde{\mathbf{x}}_{n,r_1})\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)\right|,$$

and therefore

$$-\tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t) = \left|\tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t)\right| \leq \left|\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)\right| = -\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t).$$

We divide into cases:

1. If $n \notin \mathcal{S}_{k_1}$ then we examine the sum

$$t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}$$
$$- K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}$$

The first term is negative and the second is positive. From Lemma 19 $\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1} \to \infty$. Therefore $\exists t_3$ so that $\forall t > t_3 :\ \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) < K^2$ and therefore this sum is strictly negative since

$$\left| \frac{K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right)\right) \exp\left(-\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}}{t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}} \right|$$
$$= \left| K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) \right| < 1,\ \forall t > t_3$$

2. If $n \in \mathcal{S}_{k_1}$ then we examine the sum

$$t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) - 1\right] \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}$$
$$- K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}$$

a. If $|\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)| > C_0$ then $\exists t_4$ such that $\forall t > t_4$ this sum can be upper bounded by zero since

$$\left| \frac{K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1}) t^{-\hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right)\right) \exp\left(-\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)\right) \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}}{t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k_1}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right) - 1\right] \tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \mathbf{1}_{\{\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)<0\}}} \right|$$
$$= \frac{K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1})}{1 - \exp\left(\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k_1}\right)} \leq \frac{K^2 \exp(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_{n,r_1})}{1 - \exp\left(-C_0\right)} < 1,\ \forall t > t_4 \tag{145}$$

where in the last transition we used Lemma 19.

b. If $|\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t)| \leq C_0$ then we can find constant C_5 so that eq. 145 can be upper bounded by

$$K^2 t^{-\hat{\mathbf{w}}^\top (\tilde{\mathbf{x}}_{n,k_1} + \tilde{\mathbf{x}}_{n,r_1})} \exp\left(-\tilde{\mathbf{w}}^\top (\tilde{\mathbf{x}}_{n,k_1} + \tilde{\mathbf{x}}_{n,r_1})\right) \exp\left(2C_0\right) C_0 \leq C_5 t^{-2}, \tag{146}$$

since $-\tilde{\mathbf{x}}_{n,r_1}^\top \mathbf{r}(t) \leq -\tilde{\mathbf{x}}_{n,k_1}^\top \mathbf{r}(t) \leq C_0$ and by definition, $\forall (n,k) : \hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k} \geq 1$.

Therefore, eq. 141 can be upper bounded by

$$K t^{-\theta} \exp\left(-\min_{n,k} \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) + C_5 t^{-2} \tag{147}$$

If, in addition, $\exists k, n \in \mathcal{S}_k :\ |\tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t)| > \epsilon_2$ then

$$t^{-1} \exp\left(-\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[\exp\left(-\mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k}\right) - 1\right] \tilde{\mathbf{x}}_{n,k}^\top \mathbf{r}(t) \tag{148}$$
$$\leq \begin{cases} -t^{-1} \exp\left(-\max_{n,k} \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[1 - \exp\left(-\epsilon_2\right)\right] \epsilon_2 & ,\ \text{if } \mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k} \geq 0 \\ -t^{-1} \exp\left(-\max_{n,k} \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_{n,k}\right) \left[\exp\left(\epsilon_2\right) - 1\right] \epsilon_2 & ,\ \text{if } \mathbf{r}(t)^\top \tilde{\mathbf{x}}_{n,k} < 0 \end{cases} \tag{149}$$

and we can improve this bound to

$$-C''t^{-1} < 0, \tag{150}$$

where C'' is the minimum between $\exp\left(-\max_{n,k}\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\left[1-\exp\left(-\epsilon_2\right)\right]\epsilon_2$ and $\exp\left(-\max_{n,k}\tilde{\mathbf{w}}^\top\tilde{\mathbf{x}}_{n,k}\right)\left[\exp\left(\epsilon_2\right)-1\right]\epsilon_2$. To conclude:
1. If $\|\mathbf{P}_1\mathbf{r}\left(t\right)\| \geq \epsilon_1$ (as in Eq. 139), we have that

$$\max_{k,n\in\mathcal{S}_k}\left|\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}\left(t\right)\right|^2 \overset{(1)}{\geq} \frac{1}{|\mathcal{S}|}\sum_{k,n\in\mathcal{S}_k}\left|\tilde{\mathbf{x}}_{n,k}^\top\mathbf{P}_1\mathbf{r}\left(t\right)\right|^2 = \frac{1}{|\mathcal{S}|}\left\|\mathbf{X}_\mathcal{S}^\top\mathbf{P}_1\mathbf{r}\left(t\right)\right\|^2 \overset{(2)}{\geq} \frac{1}{|\mathcal{S}|}\sigma_{\min}^2\left(\mathbf{X}_\mathcal{S}\right)\epsilon_1^2 \tag{151}$$

where in (1) we used $\mathbf{P}_1^\top\tilde{\mathbf{x}}_{n,k} = \tilde{\mathbf{x}}_{n,k}\ \forall k,\ n\in\mathcal{S}_k$, in (2) we denoted by $\sigma_{\min}\left(\mathbf{X}_\mathcal{S}\right)$, the minimal non-zero singular value of $\mathbf{X}_\mathcal{S}$ and used eq. 128. Therefore, for some (n,k), $\left|\tilde{\mathbf{x}}_{n,k}^\top\mathbf{r}\right| \geq \epsilon_2 \triangleq |\mathcal{S}|^{-1}\sigma_{\min}^2\left(\mathbf{X}_\mathcal{S}\right)\epsilon_1^2$. If $||\mathbf{P}_1\mathbf{r}(t)|| \geq \epsilon_1$, then combining eq. 139 with eq. 150 we find that eq. 137 can be upper bounded by:

$$\left(\mathbf{r}(t+1)-\mathbf{r}(t)\right)^\top\mathbf{r}(t) \leq -C''t^{-1}+o(t^{-1})$$

This implies that $\exists C_2 < C''$ and $\exists t_2 > 0$ such that eq. 129 holds. This implies also that eq. 127 holds for $||\mathbf{P}_1\mathbf{r}(t)|| \geq \epsilon_1$.
2. If $||\mathbf{P}_1 r(t)|| < \epsilon_1$, we obtain (for some positive constants C_3, C_4):

$$\left(\mathbf{r}(t+1)-\mathbf{r}(t)\right)^\top\mathbf{r}(t) \leq C_3t^{-\theta}+C_4t^{-2}$$

Therefore, $\exists t_1 > 0$ and C_1 such that eq. 127 holds.

Appendix F. An experiment with stochastic gradient descent

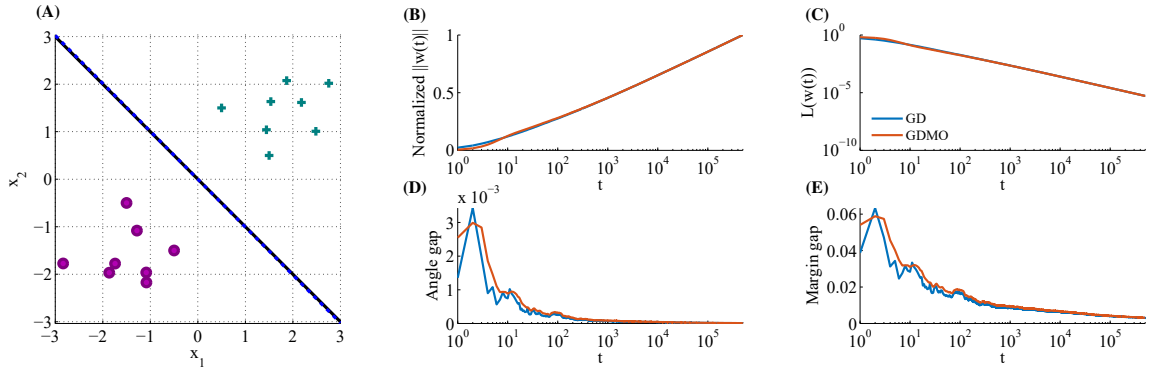

Figure 4: Same as Fig. 1, except stochastic gradient decent is used (with mini-batch of size 4), instead of GD.

References

Mor Shpigel Nacson, Nati Srebro, and Daniel Soudry. Stochastic Gradient Descent on Separable Data Exact Convergence with a Fixed Learning Rate. *AISTATS*, 2018.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit Bias of Gradient Descent on Linear Convolutional Networks. *NIPS*, 2018.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Radha Krishna Ganti. EE6151, Convex optimization algorithms. Unconstrained minimization: Gradient descent algorithm, 2015. URL

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit Regularization in Matrix Factorization. *NIPS*, 2017.

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

Moritz Hardt, Benjamin Recht, and Y Singer. Train faster, generalize better: Stability of stochastic gradient descent. *ICML*, pages 1–24, 2016.

Elad Hoffer, Itay Hubara, and D. Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *NIPS*, pages 1–13, may 2017.

I Hubara, M Courbariaux, D. Soudry, R El-yaniv, and Y Bengio. Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. *JMLR*, 2018.

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

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ICLR*, pages 1–16, 2017.

Diederik P Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization. In *ICLR*, pages 1–13, 2015.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Nathan Srebro, and Daniel Soudry. Convergence of Gradient Descent on Separable Data. *AISTATS*, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv:1412.6614*, 2014.

Behnam Neyshabur, Ruslan R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *NIPS*, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring Generalization in Deep Learning. *arXiv*, jun 2017.

Saharon Rosset, Ji Zhu, and Trevor J Hastie. Margin Maximizing Loss Functions. In *NIPS*, pages 1237–1244, 2004.

Robert E Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and N Srebro. The Implicit Bias of Gradient Descent on Separable Data. In *ICLR*, 2018.

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.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. *arXiv*, pages 1–14, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Tong Zhang, Bin Yu, et al. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.