THE IMPLICIT BIAS OF GRADIENT DESCENT ON SEPARABLE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that gradient descent on an unregularized logistic regression problem with separable data converges to the max-margin solution. The result generalizes also to other monotone decreasing loss functions with an infimum at infinity, and we also discuss a multi-class generalizations to the cross entropy loss. 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.

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 any explicit regularization, over models with more parameters and more capacity than the number of training examples, often yields good generalization, despite 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 a decent understanding of the implicit regularization introduced by early stopping of stochastic methods or, at an extreme, of one-pass (no repetition) stochastic optimization. However, as discussed above, in deep learning we often benefit from implicit bias even when optimizing the (unregularized) training error to convergence, 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, we know we will converge to the minimum Euclidean norm solution. But the logistic loss, and its generalization the cross-entropy loss which is often used in deep learning, do not admit a finite minimizer on separable problems. Instead, to drive the loss toward zero and thus minimize it, 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 \( w(t)/\|w(t)\| \), is important. How does \( w(t)/\|w(t)\| \) 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, when minimizing linearly separable logistic regression problems using gradient descent, we have that \( w(t)/\|w(t)\| \) converges to the \( L_2 \) maximum margin separator, i.e. to the solution of the hard margin SVM. This happens even though the norm \( \|w\| \), nor the margin constraint, are in no way part of the objective nor 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
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 \( \{w_t\} \) be the iterates of gradient descent (eq. 2.2) with \( \eta < 2\beta^{-1} \) and any starting point \( w(0) \). Under Assumptions 1 and 2, we have: (1) \( \lim_{t \to \infty} L(w(t)) = 0 \), (2) \( \lim_{t \to \infty} \|w(t)\| = \infty \), (3) \( \forall n : \lim_{t \to \infty} w(t) \top x_n = 0 \), and (4) \( \exists t_0 : \forall t > t_0 : w(t) \top x_n > 0 \), that is, within a finite time we reach a predictor that perfectly separates the data.

**Proof.** Since the data is strictly linearly separable, \( \exists w_* \) which linearly separates the data, and therefore

\[
0 = \sum_{n=1}^{N} \ell'(w_* \top x_n) x_n \top.
\]

For any finite \( w \), this cannot be equal to zero, as a sum of negative terms, since \( \forall n : w_* \top x_n > 0 \) and \( \forall u : \ell'(u) < 0 \). Therefore, there are no finite critical points \( w \) for which \( \nabla L(w) = 0 \). But gradient descent on a smooth loss with an appropriate stepsize is always guaranteed to converge to a critical point: \( \nabla L(w(t)) \to 0 \) (see, e.g. Lemma 5 in Appendix A.4, slightly adapted from Ganti [2015], Theorem 2). This necessarily implies that \( \|w(t)\| \to \infty \) while \( \forall n : w(t) \top x_n > 0 \) — since only then \( \ell'(w(t) \top x_n) \to 0 \). Therefore, \( L(w) \to 0 \), so GD converges to the global minimum. \( \square \)

---

2 MAIN RESULTS

Consider a dataset \( \{x_n, y_n\}_{n=1}^{N} \) with binary labels \( y_n \in \{-1, 1\} \). We analyze learning by minimizing an empirical loss of the form

\[
\ell(y, w) = \sum_{i=1}^{N} \ell(y_i, w),
\]

where \( w \in \mathbb{R}^d \) is the weight vector. A bias term could be added in the usual way, extending \( x_n \) by an additional ‘1’ component. To simplify notation, we assume that \( \forall n : y_n = 1 \) — this is true without loss of generality, since we can always re-define \( y_n x_n \) as \( x_n \).

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

**Assumption 1.** The dataset is strictly linearly separable: \( \exists w_* \) such that \( \forall n : w_* \top x_n > 0 \).

**Assumption 2.** \( \ell(u) \) is a positive, differentiable, monotonically decreasing to zero \( \ell(u) = 0 \), \( \ell'(u) < 0 \) and \( \lim_{u \to -\infty} \ell(u) = \lim_{u \to \infty} \ell'(u) = 0 \) and a \( \beta \)-smooth function, i.e. its derivative is \( \beta \)-Lipschitz.

Many common loss functions, including the logistic, exp-loss, probit and sigmoidal losses, follow Assumption 2. Under these conditions, the infimum of the optimization problem is zero, but it is not attained at any finite \( w \). Furthermore, no finite critical point \( w \) exist. We consider minimizing eq. 2.1 using Gradient Descent (GD) with a fixed learning rate \( \eta \), i.e., with steps of the form:

\[
w(t + 1) = w(t) - \eta \nabla L(w) = w(t) - \eta \sum_{n=1}^{N} \ell'(w(t) \top x_n) x_n \top.
\]

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 \( \{w_t\} \) be the iterates of gradient descent (eq. 2.2) with \( \eta < 2\beta^{-1} \) and any starting point \( w(0) \). Under Assumptions 1 and 2, we have: (1) \( \lim_{t \to \infty} L(w(t)) = 0 \), (2) \( \lim_{t \to \infty} \|w(t)\| = \infty \), (3) \( \forall n : \lim_{t \to \infty} w(t) \top x_n = 0 \), and (4) \( \exists t_0 : \forall t > t_0 : w(t) \top x_n > 0 \), that is, within a finite time we reach a predictor that perfectly separates the data.

**Proof.** Since the data is strictly linearly separable, \( \exists w_* \) which linearly separates the data, and therefore

\[
w_* \top \nabla L(w) = \sum_{n=1}^{N} \ell'(w_* \top x_n) w_* \top x_n.
\]

For any finite \( w \), this sum cannot be equal to zero, as a sum of negative terms, since \( \forall n : w_* \top x_n > 0 \) and \( \forall u : \ell'(u) < 0 \). Therefore, there are no finite critical points \( w \) for which \( \nabla L(w) = 0 \). But gradient descent on a smooth loss with an appropriate stepsize is always guaranteed to converge to a critical point: \( \nabla L(w(t)) \to 0 \) (see, e.g. Lemma 5 in Appendix A.4, slightly adapted from Ganti [2015], Theorem 2). This necessarily implies that \( \|w(t)\| \to \infty \) while \( \forall n : w(t) \top x_n > 0 \) — since only then \( \ell'(w(t) \top x_n) \to 0 \). Therefore, \( L(w) \to 0 \), so GD converges to the global minimum. \( \square \)
We are now ready to state our main result:

\[ \rho \]

**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 \(\ell (u) = \log (1 + e^{-u})\) both follow this assumption with \(a = c = 1\) — without loss of generality, since these constants can be always absorbed by re-scaling \(x_n\) and \(\eta\).

We are now ready to state our main result:

**Theorem 3.** Given strictly linearly separable data (Assumption 1) and a \(\beta\)-smooth decreasing loss function (Assumption 2) with an exponential tail (Assumption 3), then gradient descent (as in eq. 2.2) with stepsize \(\eta < 2\beta^{-1}\) and any starting point \(w(0)\) will behave as:

\[
\begin{align*}
\frac{\partial L(w)}{\partial w} &= \sum_{n=1}^{N} \exp \left(-w^T x_n\right) x_n^T = \sum_{n=1}^{N} \exp \left(-g(t) w^\infty x_n\right) \exp \left(-\rho(t)^T x_n\right) x_n^T.
\end{align*}
\]

As \(g(t) \rightarrow \infty\) 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 smallest margin \(\min_n w_n^\infty x_n\), aka the “support vectors”. The negative gradient (eq. 2.5) would then asymptotically become a non-negative linear combination of support vectors. Since the limit will be dominated by these gradients, \(w^\infty\) will also be non-negative linear combination of support vectors, and so will its scaling \(w = w^\infty / \left(\min_n w_n^\infty x_n\right)\). We therefore have:

\[
\begin{align*}
\bar{w} = \sum_n \alpha_n x_n \quad &\forall n \left(\alpha_n \geq 0 \text{ and } \bar{w}^T x_n = 1\right) \quad \text{OR} \quad \left(\alpha_n = 0 \text{ and } \bar{w}^T x_n > 1\right)
\end{align*}
\]

These are precisely the KKT condition for the SVM problem (eq. 2.4) and we can conclude that \(\bar{w}\) is indeed its solution and \(w^\infty\) is thus proportional to it.

To prove Theorem rigorously, we need to show that \(w(t) / \|w(t)\|\) 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. 2.3 into the gradient descent dynamics (eq. 2.2), with \(w^\infty = \bar{w}\) being the max margin vector and \(g(t) = \log (t)\).

We then show that the increment in the norm of \(\rho(t)\) is bounded by \(C_1 t^{-\beta}\) for some \(C_1 > 0\) and \(\beta > 1\), which is a converging series. This happens because the increment in the max margin term, \(\bar{w} \log (t+1) - \log (t) \approx \bar{w} t^{-1}\), cancels out the dominant \(t^{-1}\) term in the gradient \(-\nabla L(w(t))\) (eq. 2.5) with \(g(t) = \log (t)\) and \(w^\infty x_n = 1\). A complete proof can be found in Appendix A.
More refined analysis: characterizing the residual  We can furthermore characterize the asymptotic behavior of \( \rho (t) \). To do so, we need to refer to the KKT conditions (eq. 2.6) of the SVM problem (eq. 2.4) and the associated support vectors \( S = \arg\min_{\eta} \mathbf{w}^\top \mathbf{x}_n \). The following refinement of Theorem 3 is also proved in Appendix A.

**Theorem 4.** Under the conditions of Theorem 3 if, in addition the support vectors span the data \( \text{i.e.} \text{rank}(X_S) = \text{rank}(X) \) where the columns of \( X \) are all samples and of \( X_S \) are the support vectors, then \( \lim_{t \to \infty} \rho (t) = \hat{w} \), where \( \hat{w} \) is the unique solution to

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

3 Implications: Rates of Convergence

The solution in eq. 2.3 implies that \( \mathbf{w} (t) / \| \mathbf{w} (t) \| \) converges to the normalized max margin vector \( \hat{w} / \| \hat{w} \| \). Moreover, this convergence is very slow—logarithmic in the number of iterations. Specifically, in Appendix B we show that Theorem 3 implies the following tight rates of convergence:

The normalized weight vector converges to normalized max margin vector in \( L_2 \) norm

\[
\left\| \frac{\mathbf{w} (t)}{\| \mathbf{w} (t) \|} - \frac{\hat{w}}{\| \hat{w} \|} \right\| = O \left( \frac{1}{\log t} \right),
\]

and in angle

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

and the margin converges as

\[
1 - \min_{\mathbf{w}} \frac{\mathbf{x}_n^\top \mathbf{w} (t)}{\| \mathbf{w} (t) \|} = O \left( \frac{1}{\log t} \right).
\]

This slow convergence is in sharp contrast to the convergence of the (training) loss:

\[
\mathcal{L} (\mathbf{w} (t)) = O \left( \frac{1}{t} \right).
\]

A simple construction (also in Appendix B) shows that the rates in the above equations are tight. Thus, the convergence of \( \mathbf{w} (t) \) to the max-margin \( \hat{w} \) 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 3.1. 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, 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 \( \hat{w} \). Although \( \hat{w} \) 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 \( \hat{w} \) classifies all training points correctly, we have that on the training set \( \frac{1}{n} \sum_{n} h(\hat{w}^\top \mathbf{x}_n) = 0 \). However, on the population we would expect some errors and so \( \mathbb{E} [h(\hat{w}^\top \mathbf{x})] > 0 \). Since \( \mathbf{w}(t) \approx \hat{w} \log t + \ell (\alpha u) \to \alpha h(u) \) as \( \alpha \to \infty \), we have:

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

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.
We discuss several possible extensions of our results. 

\[ \text{Implementation details:} \]

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 \( L_{\text{val}}(w(t)) = \sum_{x \in \mathcal{V}} \ell(w(t)^T x) \) calculated on an independent validation set \( \mathcal{V} \), will increase logarithmically with \( t \) (since we would not expect zero validation error), which might cause us to think we are over-fitting or otherwise encourage us to stop the optimization. But this increase does not actually represent the model getting worse, merely \( \|w(t)\| \) getting larger, and in fact the model might be getting better (with larger margin and possibly smaller error rate).

### 4 Extensions

We discuss several possible extensions of our results.

#### 4.1 Multi-Class Classification with Cross-Entropy Loss

So far, we have discussed the problem of binary classification. For multi-class problems commonly encountered, we frequently learn a predictor \( w_k \) for each class, and use the cross-entropy loss with a softmax output, which is a generalization of the logistic loss. What do the linear predictors \( w_k(t) \) converge to if we minimize the cross-entropy loss by gradient descent on the predictors? In Appendix \( C \) we analyze this problem for separable data, and show that again, the predictors diverge to infinity and the loss converges to zero. Furthermore, we show that, generically, the loss converges to a logistic loss for transformed data, for which our theorems hold. This strongly suggests that gradient descent converges to a scaling of the \( K \)-class SVM solution:

\[
\arg \min_{w_1, \ldots, w_K} \sum_{k=1}^{K} \|w_k\|^2 \text{ s.t. } \forall n, y_n \neq y_n : w_{y_n}^T x_n \geq w_k^T x_n + 1 \quad (4.1)
\]

We believe this can also be established rigorously and for generic exponential tailed multi-class loss.
ADAM does not. The implicit bias of adaptive methods has been a recent topic of interest, with the red dashed line does not converge to $L_2$ max margin solution (black line), in contrast to GD (blue dashed line), or GDMO.

4.2 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 3.1, we see that adding momentum does not qualitative effect the bias induced by gradient descent. In Figure E.1 in Appendix E we also repeat the experiment using stochastic gradient descent, and observe a similar bias. This is consistent with the fact that momentum, acceleration and stochasticity do not change the bias when using gradient descent to optimize an underdetermined least squares problems. 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.

Employing adaptive methods, such as AdaGrad [Duchi et al., 2011] and ADAM [Kingma & Ba, 2015], does significantly affect the bias. In Figure 4.1 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 method has been a recent topic of interest, with Hoffer et al. [2017] and Wilson et al. [2017] suggesting they lead to worse generalization. Wilson et al. discuss the limit of AdaGrad on least square problems, but fall short of providing an actual characterization of the limit. This is not surprising, as the limit of AdaGrad on least square problems is fragile and depends on the choice of stepsize and other parameters, and thus complicated to characterize. We expect our methodology could be used to precisely characterize the implicit bias of such methods on logistic regression problems. The asymptotic nature of the analysis is appealing here, as it is insensitive to the initial point, initial conditioning matrix, and large initial steps.

More broadly, it would be interesting to study the behavior of mirror descent and natural gradient descent, and relate the bias they induce to the potential function or divergence underlying them. A reasonable conjecture, which we have not yet investigated, is that for any potential function $\Psi(w)$, these methods converge to the maximum $\Psi$-margin solution $\arg\min_w \Psi(w) \text{s.t.} \forall n : w^\top x_n \geq 1$. Since mirror descent can be viewed as regularizing progress using $\Psi(w)$, it is worth noting the results of Rosset et al. [2004b]: they considered the regularization path $w_\lambda = \arg\min L(w) + \lambda \|w\|_p^p$ for similar loss function as we do and showed that $\lim_{\lambda \to 0} w_\lambda / \|w_\lambda\|_p$ is proportional to the maximum $L_p$ margin solution. 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.

Our analysis also covers the exp-loss used in boosting, as its tail is similar to that of the logistic loss. However, boosting is a coordinate descent procedure, and not a gradient descent procedure. Indeed, the coordinate descent interpretation of AdaBoost shows that coordinate descent on the exp-loss for a
Figure 4.2: 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.

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

Table 1: Sample values from various epochs in the experiment depicted in Fig. 4.2.

linearly separable problem is related to finding the maximum $L_1$ margin solution (Schapire et al., 1998; Rosset et al., 2004a; Shalev-Shwartz & Singer, 2010).

4.3 Deep networks

In this paper, we only consider 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 improving even as the validation loss increases. These effects are demonstrated in Figure 4.2 and Table 1 which portray typical training of a convolutional neural network using unregularized gradient descent.

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 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 two special cases, our results may be directly applied to multi-layered networks. First, our results may be applied exactly, as we show in Appendix D, 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. Second, our results may also be applied directly to the last weight layer 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., 2016).

4.4 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 we minimize the

\[ \text{Code available here: } \text{https://github.com/paper-submissions/MaxMargin} \]
**5 SUMMARY**

We characterized the implicit bias induced by gradient descent 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 the logistic loss, or exp-loss, or any other monotone decreasing loss with appropriate tail. In particular, the non-tail part does not affect the bias and so the logistic loss and the exp-loss, although very different on non-separable problems, behave the same for separable problems. The bias is also independent of the step-size used (as long as it is small enough to ensure convergence) and (unlike for least square problem) is also independent on the initialization.

2. This convergence is very slow. This 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 slow decrease of the training loss, or on no decrease of the validation loss, to decide when to stop. We might improve the validation, and test, errors even when the validation loss increases and even when the decrease in the training loss is tiny.

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. Perhaps 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.

**REFERENCES**


I Hubara, M Courbariaux, D. Soudry, R El-yaniv, and Y Bengio. Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. *Accepted to JMLR*, 2016.


Appendix

A Proof of Main Results

In the following proofs, for any solution \( w(t) \), we define \( r(t) = w(t) - \hat{w} \log t - \bar{w} \), where \( \bar{w} \) and \( \hat{w} \) follow the conditions of the Theorems[2] and [3], i.e. \( \hat{w} \) is the \( L_2 \) max margin vector and (eq. 2.4), so
\[
\forall n \in S : \ x_n^T \bar{w} = 1 ; \ \theta = \max_{n \not\in S} x_n^T \bar{w} > 1 . \quad (A.1)
\]
and eq. 2.7 holds
\[
\forall n \in S : \ \eta \exp (-x_n^T \bar{w}) = \alpha_n , \quad (A.2)
\]
where we recall that we denoted \( X_S \in \mathbb{R}^{S \times d} \) as the matrix of support vectors, which columns are the support vectors, and a subset \( S \subset \{1, \ldots, N\} \) of the columns of \( X = [x_1, \ldots, x_N] \in \mathbb{R}^{d \times N} \). Note that, in general, if the support vectors do not span the data, the solution \( \bar{w} \) to eq. A.2 might not be unique, and we can use any solution in the proof.

We furthermore denote by \( C_{i, \epsilon_1, t} (i \in \mathbb{N}) \) various positive constants which are independent of \( t \). Lastly, we define \( P_1 \in \mathbb{R}^{d \times d} \) as the orthogonal projection matrix[4] to the subspace spanned by the support vectors (the columns of \( X_S \)), and \( P_2 = I - P_1 \) as the complementary projection (to the left nullspace of \( X_S \)).

A.1 Auxiliary Lemmata

In our proofs, we use the following Lemmata.

**Lemma 5.** Let \( L(w) \) be a general \( \beta \)-smooth non-negative objective. If \( \eta < 2\beta^{-1} \), then for any solution of GD
\[
w(t + 1) = w(t) - \eta \nabla L(w) \quad (A.3)
\]
we have that \( \sum_{t=0}^{T} \| \nabla L(w(t)) \|^2 < \infty \) and \( \lim_{t \to \infty} t \| \nabla L(w(t)) \|^2 = 0 \).

**Lemma 6.** We have
\[
\exists C_1, t_1 : \ \forall t > t_1 : \ (r(t) + 1 - r(t)) \leq C_1 t^{-\min(\theta, -1 - 1.5 \epsilon_1, -1 - 0.5 \epsilon_2)} . \quad (A.4)
\]
Additionally, \( \forall \epsilon_1 > 0 \), \( \exists C_2, t_2 \), such that \( \forall t > t_2 \), such that if
\[
\| P_1 r(t) \| \geq \epsilon_1 , \quad (A.5)
\]
then we can improve this bound to
\[
(r(t + 1) - r(t)) \leq C_2 t^{-1} < 0 . \quad (A.6)
\]

We prove both Lemmata below, in appendix sections A.4 and A.5 after we prove of both theorems.

A.2 Proof of Theorem[3]

Our goal is to show that \( \| r(t) \| \) is bounded, and therefore \( \rho(t) = r(t) + \hat{w} \) is bounded. To show this, we will upper bound the following equation
\[
\| r(t + 1) \|^2 = \| r(t + 1) - r(t) \|^2 + 2 \langle r(t + 1) - r(t), r(t) \rangle + \| r(t) \|^2 \quad (A.7)
\]
First, we note that first term in this equation can be upper-bounded by
\[
| r(t + 1) - r(t) |^2 = \|
\]
\[
\| w(t + 1) - \hat{w} \log (t + 1) - w(t) + \hat{w} \log (t) + \bar{w} \|^2
\]
\[
= \| -\eta \nabla L(w(t)) - w \log (t + 1) - \log (t) \|^2
\]
\[
= \eta^2 \| \nabla L(w(t)) \|^2 + \| \hat{w} \|^2 \log^2 (1 + t^{-1}) + 2 \eta \hat{w}^T \nabla L(w(t)) \log (1 + t^{-1})
\]
\[
\leq \eta^2 \| \nabla L(w(t)) \|^2 + \| \hat{w} \|^2 t^{-2} \quad (A.8)
\]

This matrix can be written as \( P_1 = X_S X_S^+ \), where \( M^+ \) is penrose-moore pseudo inverse of \( M \).
where in (1) we used eq. (2.3) in (2) we used eq. (2.2) and in (3) we used \( \forall x > 0 : x \geq \log (1 + x) > 0 \), and also that
\[
\hat{w}^T \nabla L (w (t)) = \sum_{n=1}^{N} \ell \left( w (t)^T x_n \right) \hat{w}^T x_n \leq 0 \tag{A.9}
\]
since \( \hat{w}^T x_n \geq 1 \) (eq. A.1).

Also, from Lemma 5 we know that
\[
\| \nabla L (w (t)) \|^2 = o \left( t^{-1} \right) \text{ and } \sum_{t=0}^{\infty} \| \nabla L (w (t)) \|^2 < \infty . \tag{A.10}
\]

Substituting eq. (A.10) into eq. (A.8) and recalling that a \( t^{-\beta} \) power series converges for some \( \beta > 1 \), we can find \( C_0 \) such that
\[
\| r (t + 1) - r (t) \|^2 = o \left( t^{-1} \right) \text{ and } \sum_{t=0}^{\infty} \| r (t + 1) - r (t) \|^2 = C_0 < \infty . \tag{A.11}
\]

Note that this equation also implies that
\[
\exists \epsilon_0 : \forall t > t_0 : \| r (t + 1) \| - \| r (t) \| < \epsilon_0 . \tag{A.12}
\]

Next, we would like to bound the second term in eq. (A.7). From eq. (A.4) in Lemma 5, we can find \( t_1, C_1 \) such that \( \forall t > t_1 \): \[
(r (t + 1) - r (t))^\top r (t) \leq C_1 t^{- \min(\theta, -1 - 1.5 \epsilon_+, -1 - 0.5 \epsilon_-)} . \tag{A.13}
\]

Thus, by combining eqs. (A.13) and (A.11) into eq. (A.7) we find
\[
\| r (t) \|^2 - \| r (t_1) \|^2 \\
= \sum_{u=t_1}^{t-1} \| r (u + 1) \|^2 - \| r (u) \|^2 \\
\leq C_0 + 2 \sum_{u=t_1}^{t-1} C_1 u^{- \min(\theta, -1 - 1.5 \epsilon_+, -1 - 0.5 \epsilon_-)}
\]
which is a bounded, since \( \theta > 1 \) (eq. A.1). Therefore, \( \| r (t) \| \) is bounded. \[ \blacksquare \]

### A.3 Proof of Theorem 4

All that remains now is to show that \( \| r (t) \| \to 0 \) if \( \text{rank} (X_S) = \text{rank} (X) \). To do so, this proof will continue where the proof of Theorem 4 stopped, using notations and equations from that proof.

Since \( r (t) \) has bounded norm, its two orthogonal components \( r (t) = P_1 r (t) + P_2 r (t) \) also have bounded norms (recall where \( P_1, P_2 \) defined in the beginning of appendix section A). From eq. (2.2), \( \nabla L (w) \) is spanned by the columns of \( X \). If \( \text{rank} (X_S) = \text{rank} (X) \), then it is also spanned by the columns of \( X_S \), and so \( P_2 \nabla L (w) = 0 \). Therefore, \( P_2 r (t) \) is not updated during GD, and remains constant. Since \( \hat{w} \) in eq. (2.3) is also bounded, we can absorb this constant \( P_2 r (t) \) into \( \hat{w} \) without affecting eq. (2.7) (since \( \forall n \in S : x_n P_2 r (t) = 0 \)). Thus, without loss of generality, we can assume that \( r (t) = P_1 r (t) \).

Now, recall eq. (A.6) in Lemma 6
\[
\exists c_2, t_2 : \forall t > t_2 : (r (t + 1) - r (t))^\top r (t) \leq -c_2 t^{-1} < 0 .
\]

Combining this with eqs. (A.7) and (A.11) implies that \( \exists t_3 > \max \{ t_2, t_0 \} \) such that \( \forall t > t_3 \) such that \( \| r (t) \| > \epsilon_1 \), we have that \( \| r (t + 1) \|^2 - \| r (t) \|^2 \) is a decreasing function since then
\[
\| r (t + 1) \|^2 - \| r (t) \|^2 \leq -c_3 t^{-1} < 0 . \tag{A.14}
\]
Additionally, this result also implies that we cannot have \( \| \mathbf{r}(t) \| > \varepsilon_1 \forall t > t_3 \), since then we arrive to the contradiction.

\[
\| \mathbf{r}(t) \|^2 - \| \mathbf{r}(t_3) \|^2 = \sum_{u=t_3}^{t-1} (\| \mathbf{r}(u+1) \|^2 - \| \mathbf{r}(u) \|^2) \leq -\sum_{u=t_3}^{t-1} C_3 u^{-1} \to -\infty,
\]

Therefore, \( \exists t_4 > t_3 \) such that \( \| \mathbf{r}(t_4) \| \leq \varepsilon_1 \). Recall also that \( \| \mathbf{r}(t) \| \) is a decreasing function whenever \( \| \mathbf{r}(t) \| \geq \varepsilon_1 \) (eq. [A.14]). Also, recall that \( t_4 > t_0 \), so from eq. [A.12] we have that \( \forall t > t_4, \| \mathbf{r}(t+1) \| - \| \mathbf{r}(t) \| < \varepsilon_0 \). Combining these three facts we conclude that \( \forall t > t_4 : \| \mathbf{r}(t) \| \leq \varepsilon_1 + \varepsilon_0 \). Since this reasoning holds \( \forall \varepsilon_1, \varepsilon_0 \), this implies that \( \| \mathbf{r}(t) \| \to 0. \)

**A.4 Proof of Lemma 5**

**Lemma 5.** Let \( \mathcal{L}(\mathbf{w}) \) be a general \( \beta \)-smooth non-negative objective. If \( \eta < 2\beta^{-1} \), then for any solution of GD

\[
\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla \mathcal{L}(\mathbf{w}) \tag{A.3}
\]

we have that \( \sum_{t=0}^{T} \| \nabla \mathcal{L}(\mathbf{w}(u)) \|_2^2 < \infty \) and \( \lim_{t \to \infty} \| \nabla \mathcal{L}(\mathbf{w}(t)) \|_2^2 = 0. \)

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

\[
\left| f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \right| \leq \frac{\beta}{2} \| \mathbf{x} - \mathbf{y} \|_2^2. \tag{A.15}
\]

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

\[
\mathcal{L}(\mathbf{w}(t+1)) \leq \mathcal{L}(\mathbf{w}(t)) + \nabla \mathcal{L}(\mathbf{w}(t)) (\mathbf{w}(t+1) - \mathbf{w}(t)) + \frac{\beta}{2} \| \mathbf{w}(t+1) - \mathbf{w}(t) \|_2^2
\]

\[
= \mathcal{L}(\mathbf{w}(t)) - \eta \| \nabla \mathcal{L}(\mathbf{w}(t)) \|_2^2 + \frac{\beta \eta^2}{2} \| \nabla \mathcal{L}(\mathbf{w}(t)) \|_2^2
\]

\[
= \mathcal{L}(\mathbf{w}(t)) - \eta \left( 1 - \frac{\beta \eta}{2} \right) \| \nabla \mathcal{L}(\mathbf{w}(t)) \|_2^2
\]

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

\[
\sum_{u=0}^{t} \| \nabla \mathcal{L}(\mathbf{w}(u)) \|^2 \leq \sum_{u=0}^{t} \frac{\mathcal{L}(\mathbf{w}(u)) - \mathcal{L}(\mathbf{w}(u+1))}{\eta \left( 1 - \frac{\beta \eta}{2} \right)} = \frac{\mathcal{L}(\mathbf{w}(0)) - \mathcal{L}(\mathbf{w}(t+1))}{\eta \left( 1 - \frac{\beta \eta}{2} \right)} < \infty,
\]

since \( \mathcal{L}(\mathbf{w}(0)) \to \infty \) and \( 0 \leq \mathcal{L}(\mathbf{w}(t+1)) \). Furthermore, since

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

we also have \( t \| \nabla \mathcal{L}(\mathbf{w}(t)) \|^2 \to 0 \), from Cauchy Condensation test, so \( \| \nabla \mathcal{L}(\mathbf{w}(t)) \| \to o(t^{-1/2}). \)

**A.5 Proof of Lemma 6**

Recall that we defined \( \mathbf{r}(t) = \mathbf{w}(t) - \hat{\mathbf{w}} \log (t) - \hat{\mathbf{w}} \), with \( \hat{\mathbf{w}} \) and \( \hat{\mathbf{w}} \) follow the conditions of the Theorems 3 and 4 i.e. \( \hat{\mathbf{w}} \) is the \( L_2 \) max margin vector and (eq. [2.4], and eq. [2.7]) holds

\[
\forall n \in \mathcal{S} : \eta \exp(-\mathbf{x}_n^\top \hat{\mathbf{w}}) = \alpha_n.
\]

**Lemma 6.** We have

\[
\exists C_1, t_1 : \forall t > t_1 : (\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq C_1 t^{-\min(\theta, -1-1.5\varepsilon_+,-1-0.5\varepsilon_-)} \tag{A.4}
\]

Additionally, \( \forall \varepsilon_1 > 0, \exists C_2, t_2, such that \forall t > t_2, such that if

\[
\| \mathbf{P}_1 \mathbf{r}(t) \| \geq \varepsilon_1 \tag{A.5}
\]

then we can improve this bound to

\[
(\mathbf{r}(t+1) - \mathbf{r}(t))^\top \mathbf{r}(t) \leq -C_2 t^{-1} < 0 \tag{A.6}
\]
We examine the three terms in eq. A.18. The first term can be upper bounded by

\[
\forall t > t_+: -\ell' (w^\top (t) x_n) \leq (1 + \exp (-\epsilon_+ w^\top (t) x_n)) \exp (-w^\top (t) x_n) \tag{A.16}
\]

\[
\forall t > t_-: -\ell' (w^\top (t) x_n) \geq (1 - \exp (-\epsilon_- w^\top (t) x_n)) \exp (-w^\top (t) x_n) \tag{A.17}
\]

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

\[
(r(t + 1) - r(t))^\top r(t)
\]

\[
= (-\eta \nabla \mathcal{L} (w(t)) - \bar{w} [\log (t + 1) - \log (t)])^\top r(t)
\]

\[
= -\eta \sum_{n=1}^{N} \ell' (w(t)^\top x_n) x_n^\top r(t) - \bar{w}^\top r(t) \log (1 + t^{-1})
\]

\[
= \hat{w}^\top r(t) [t^{-1} - \log (1 + t^{-1})] - \eta \sum_{n \notin S} \ell' (w(t)^\top x_n) x_n^\top r(t)
\]

\[
- \eta \sum_{n \in S} [t^{-1} \exp (-\bar{w}^\top x_n) + \ell' (w(t)^\top x_n)] x_n^\top r(t) \tag{A.18}
\]

where in last line we used eqs. 2.6 and 2.7 to obtain

\[
\hat{w} = \sum_{n \in S} \alpha_n x_n = \eta \sum_{n \in S} \exp (-\bar{w}^\top x_n) x_n .
\]

We examine the three terms in eq. A.18. The first term can be upper bounded by

\[
\hat{w}^\top r(t) [t^{-1} - \log (1 + t^{-1})]
\]

\[
\leq \max \left[ \hat{w}^\top r(t), 0 \right] [t^{-1} - \log (1 + t^{-1})] \tag{1}
\]

\[
\leq \max \left[ \hat{w}^\top P_1 r(t), 0 \right] t^{-2}
\]

\[
\leq \begin{cases} 
\|\hat{w}\|_1 t^{-2}, & \text{if } \|P_1 r(t)\| \leq \epsilon_1 \\
\Theta (t^{-1}), & \text{if } \|P_1 r(t)\| > \epsilon_1
\end{cases} \tag{2}
\]

where in (1) we used that \( P_2 \hat{w} = P_2 X_S \alpha = 0 \) from eq. 2.6 and in (2) we used that \( \hat{w}^\top r(t) = o(t) \), since

\[
t^{-1} \hat{w}^\top r(t) = t^{-1} \hat{w}^\top \left( w(0) - \eta \sum_{u=0}^{t} \nabla \mathcal{L} (w(u)) - \bar{w} \log (t) - \hat{w} \right)
\]

\[
\leq o(1) - \eta t^{-1} \hat{w}^\top \left( \sum_{u=0}^{[\sqrt{t}]} \nabla \mathcal{L} (w(u)) + \sum_{u=\lceil \sqrt{t} \rceil + 1}^{t} \nabla \mathcal{L} (w(u)) \right)
\]

\[
\leq o(1) - \eta t^{-1} \left( [\sqrt{t}] \max_{u \leq \sqrt{t}} \hat{w}^\top \nabla \mathcal{L} (w(u)) + t \max_{[\sqrt{t}] \leq u \leq t} \hat{w}^\top \nabla \mathcal{L} (w(u)) \right)
\]

\[
\leq o(1)
\]

where in the last line we used that \( \nabla \mathcal{L} (w(t)) = o(1) \), from Lemma 5.
Next, we upper bound the second term in eq. \[ A.18 \] \( \forall t > t_+ \):

\[
- \eta \sum_{n \in S} \ell' \left( \mathbf{w} (t) ^\top x_n \right) x_n ^\top r (t) \\
\leq - \eta \sum_{n \in S: x_n ^\top r (t) \geq 0} \ell' \left( \mathbf{w} (t) ^\top x_n \right) x_n ^\top r (t) \\
\overset{(1)}{\leq} \eta \sum_{n \in S: x_n ^\top r (t) \geq 0} \left( 1 + \exp \left( - \epsilon_+ \mathbf{w} ^\top (t) x_n \right) \right) \exp \left( - \mathbf{w} (t) ^\top x_n \right) x_n ^\top r (t) \\
\overset{(2)}{\leq} \eta \sum_{n \in S: x_n ^\top r (t) \geq 0} \left( 1 + t^{-\epsilon_+ x_n ^\top \tilde{w}} \exp \left( - \epsilon_+ \tilde{w} ^\top x_n - x_n ^\top r (t) \right) \right) t^{-x_n ^\top \tilde{w}} \exp \left( - \tilde{w} ^\top x_n \right) x_n ^\top r (t) \\
\overset{(3)}{\leq} \eta \sum_{n \in S: x_n ^\top r (t) \geq 0} \left( 1 + t^{-\epsilon_+ x_n ^\top \tilde{w}} \exp \left( - \epsilon_+ \tilde{w} ^\top x_n \right) \right) t^{-x_n ^\top \tilde{w}} \exp \left( - \tilde{w} ^\top x_n \right) \\
\overset{(4)}{\leq} \eta N \left( 1 + \left[ t^{-\theta} \exp \left( - \frac{n}{\min \tilde{w} ^\top x_n} \right) \right] ^{\epsilon_1} \right) \exp \left( - \frac{n}{\min \tilde{w} ^\top x_n} \right) t^{-\theta} \\
\leq 2\eta N \exp \left( - \frac{n}{\min \tilde{w} ^\top x_n} \right) t^{-\theta}, \forall t > t_+ \quad (A.20)
\]

where in (1) we used eq. \[ A.16 \] in (2) we used \( \mathbf{w} (t) = \tilde{w} \log t + \tilde{w} + r (t) \), in (3) we used \( \tilde{x} e^{-\mathbf{w} ^\top x_n} \leq 1 \) and \( x_n ^\top r (t) \geq 0 \), in (4) we used \( \theta > 1 \), from eq. \[ A.1 \] and in (5) we defined \( t'_+ = \max \left[ t_+, \exp \left( \min \tilde{w} ^\top x_n \right) \right] \).

Lastly, the third term in eq. \[ A.18 \] can be upper bounded by

\[
\eta |S| \max_{n \in \tilde{S}} \left[ -\ell' \left( \mathbf{w} (t) ^\top x_n \right) t^{-1} \exp \left( - \tilde{w} ^\top x_n \right) \right] x_n ^\top r (t) \\
\overset{(5)}{=} \eta |S| \max_{n \in \tilde{S}} \left[ -\ell' \left( \mathbf{w} (t) ^\top x_n \right) t^{-1} \exp \left( - \tilde{w} ^\top x_n \right) \right] x_n ^\top r (t) \quad (A.21)
\]

where

\[
\tilde{S} \triangleq \left\{ n \in S : |x_n ^\top r (t)| > 0 \right\}
\]

and \( k \) is the index that maximizes the expression in eq. \[ A.21 \] Additionally, if \( \| \mathbf{P}_1 r (t) \| \geq \epsilon_1 \) (as in Eq. \[ A.5 \]), we have that

\[
\min_{n \in \tilde{S}} \left| x_n ^\top r (t) \right|^2 \overset{(1)}{=} \min_{n \in S} \left| x_n ^\top \mathbf{P}_1 r (t) \right|^2 \overset{(2)}{\geq} \sigma_{\min}^2 (X_S) \epsilon_1^2 \overset{(3)}{=} \epsilon_2 > 0 \quad (A.22)
\]

where in (1) we used \( \mathbf{P}_1 x_n = x_n \forall n \in \tilde{S} \), in (2) we denoted by \( \sigma_{\min} (X_S) \), the minimal non-zero singular value of \( X_S \) and used eq. \[ A.5 \] Thus, the condition \( \| \mathbf{P}_1 r (t) \| \geq \epsilon_1 \) implies that the set \( \tilde{S} \) is not empty. Using these notations, we upper bound eq. \[ A.21 \] in various cases:

First, if \( x_n ^\top r (t) \geq 0 \), then eq. \[ A.21 \] can be upper bounded \( \forall t > t_+ \), using eq. \[ A.16 \] by

\[
\eta t^{-1} |S| \exp \left( - \tilde{w} ^\top x_k \right) \left[ \left( 1 + t^{-\epsilon_+} \exp \left( - \epsilon_+ \tilde{w} ^\top x_n \right) \right) \exp \left( - x_k ^\top r (t) \right) - 1 \right] x_k ^\top r (t) \quad (A.23)
\]

We divide into cases:

1. If \( \| x_k ^\top r \| \leq C_0 t^{-0.5 \epsilon_+} \), then we can upper bound eq. \[ A.23 \] with

\[
\eta |S| \exp \left( - \tilde{w} ^\top x_k \right) \exp \left( - \epsilon_+ \max_n \tilde{w} ^\top x_n \right) C_0 t^{-1-1.5 \epsilon_+} \quad (A.24)
\]

2. If \( \| x_k ^\top r \| > C_0 t^{-0.5 \epsilon_+} \), then we can find \( t''_+ > t'_+ \) to upper bound eq. \[ A.23 \] \( \forall t > t''_+ \):

\[
\eta t^{-1} |S| e^{-\tilde{w} ^\top x_k \left[ \left( 1 + t^{-\epsilon_+} \exp \left( - \epsilon_+ \tilde{w} ^\top x_n \right) \right) \exp \left( - C_0 t^{-0.5 \epsilon_+} \right) - 1 \right] x_k ^\top r (t) \\
\leq \eta t^{-1} |S| e^{-\tilde{w} ^\top x_k \left[ \left( 1 + t^{-\epsilon_+} \exp \left( - \epsilon_+ \tilde{w} ^\top x_n \right) \right) \left( 1 - C_0 t^{-0.5 \epsilon_+} + C_0^2 t^{-1.5 \epsilon_+} \right) - 1 \right] x_k ^\top r (t) \\
\overset{(1)}{\leq} \eta t^{-1} |S| e^{-\tilde{w} ^\top x_k \left[ 1 - C_0 t^{-0.5 \epsilon_+} + C_0^2 t^{-1.5 \epsilon_+} \right] \exp \left( - \epsilon_+ \min \tilde{w} ^\top x_n t^{-\epsilon_+} - C_0 t^{-0.5 \epsilon_+} \right) x_k ^\top r (t) \\
\leq \eta t^{-1} |S| e^{-\tilde{w} ^\top x_k \left[ 1 - C_0 t^{-0.5 \epsilon_+} + C_0^2 t^{-1.5 \epsilon_+} \right] \exp \left( - \epsilon_+ \min \tilde{w} ^\top x_n t^{-\epsilon_+} - C_0 t^{-0.5 \epsilon_+} \right) x_k ^\top r (t) \\
\overset{(2)}{\leq} 0 \quad \forall t > t''_+ \quad (A.25)
\]

where in (1) we use the fact that \( \epsilon^{-x} \leq 1 - x + x^2 \) and in (2) we define \( t''_+ \) so that the previous expression is negative — this is possible since \( t^{-0.5 \epsilon_+} \) decreases slower then \( t^{-\epsilon_+} \).
3. If \( \left\| \mathbf{P}_t \mathbf{r}(t) \right\| \leq \epsilon_1 \) (as in Eq. A.5), then eq. A.22 can be used to upper bound eq. A.23 by 
\[
- \eta \mid \mathcal{S} \mid \exp \left( - \max_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) \left( 1 - \left( 1 + \epsilon t^{-\epsilon} \exp \left( - \epsilon \left( \min_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) \right) \right) \epsilon_2 t^{-1} \\
\leq - \eta \mid \mathcal{S} \mid \exp \left( - \max_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) \left( 1 - e^{-0.5 \epsilon_2} \right) \epsilon_2 t^{-1}, \forall t > t'' \\
\]
(A.26)

Where we defined \( t'' > t' \) such that \( t'' > \exp \left( \min_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) \left[ e^{0.5 \epsilon_2} - 1 \right]^{-1/\epsilon} \), and therefore \( \forall t > t'' \), we have \( \left( 1 + \epsilon t \exp \left( - \epsilon \left( \min_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) \right) \right) e^{-\epsilon_2 t} < e^{-0.5 \epsilon_2} \).

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

1. If \( \mathbf{x}_k^\top \mathbf{r} \leq C_0 t^0 - 0.5 \epsilon_2 \), then, since \( - \ell' \left( \mathbf{w} (t) \mathbf{x}_n \right) > 0 \), we can upper bound eq. A.21 with 
\[
\eta t^{-1} \mid \mathcal{S} \mid \exp \left( - \bar{\mathbf{w}}^\top \mathbf{x}_k \right) \mathbf{x}_k^\top \mathbf{r}(t) \leq \eta \mid \mathcal{S} \mid \exp \left( - \min_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) C_0 t^{-1 - 0.5 \epsilon_2} \\
\]
(A.27)

2. If \( \mathbf{x}_k^\top \mathbf{r} > C_0 t^0 - 0.5 \epsilon_2 \), then, using eq. A.17 \( \forall t > t_- \), we can find \( t'' > t_- \) such that upper bound eq. A.21 with zero \( \forall t > t_- \), since 
\[
\eta \mid \mathcal{S} \mid \left[ - t^{-1} e^{- \bar{\mathbf{w}}^\top \mathbf{x}_k} x_k^\top \mathbf{r}(t) - \ell' \left( \mathbf{w} (t) \mathbf{x}_n \right) \right] \mathbf{x}_k^\top \mathbf{r}(t) \\
\leq \eta \mid \mathcal{S} \mid \left[ - t^{-1} e^{- \bar{\mathbf{w}}^\top \mathbf{x}_k} + \left( 1 - \exp \left( - \epsilon \mathbf{w}^\top (t) \mathbf{x}_n \right) \right) \exp \left( - \mathbf{w}^\top (t) \mathbf{x}_n \right) \right] \mathbf{x}_k^\top \mathbf{r}(t) \\
= \eta \mid \mathcal{S} \mid e^{- \bar{\mathbf{w}}^\top \mathbf{x}_k} t^{-1} \left[ 1 - \left( 1 - e^{- \epsilon \bar{\mathbf{w}}^\top \mathbf{x}_k} t^{-\epsilon} \right) \exp \left( C_0 t^0 - 0.5 \epsilon_2 \right) \right] C_0 t^{-0.5 \epsilon_2} \\
\leq \eta C_0 \mid \mathcal{S} \mid e^{- \bar{\mathbf{w}}^\top \mathbf{x}_k} t^{-1-0.5 \epsilon_2} \left( 1 - \left( 1 - e^{- \epsilon \bar{\mathbf{w}}^\top \mathbf{x}_k} \right) \right) \left( 1 + C_0 t^0 - 0.5 \epsilon_2 \right) \\
= \eta C_0 \mid \mathcal{S} \mid e^{- \bar{\mathbf{w}}^\top \mathbf{x}_k} t^{-1-0.5 \epsilon_2} \left( t^{-\epsilon} e^{- \epsilon \bar{\mathbf{w}}^\top \mathbf{x}_k} t^{-\epsilon} - C_0 t^0 - 0.5 \epsilon_2 + C_0 e^{- \epsilon \bar{\mathbf{w}}^\top \mathbf{x}_k} t^{-1-0.5 \epsilon_2} \right) \\
< 0, \forall t > t_- \\
\]
(A.28)

where in (1) we used \( \forall x > 0 : e^x \geq 1 + x \), and in the last line we were able to find \( t_- \) such that this expression is negative since \( t^0 \) decreases slower then \( t^- \), such that this expression is negative given sufficiently large \( t \).

3. If \( \left\| \mathbf{P}_t \mathbf{r}(t) \right\| \geq \epsilon_1 \) then, using eqs. A.17 and A.22 \( \forall t > t_- \), we can find \( t'' > t_- \) and \( C_0' > C_0 \) such that we can upper bound eq. A.21 with \( \forall t > t_- \), with 
\[
\eta \mid \mathcal{S} \mid \left[ - t^{-1} \exp \left( - \bar{\mathbf{w}}^\top \mathbf{x}_k \right) + \left( 1 - \exp \left( - \epsilon \mathbf{w}^\top (t) \mathbf{x}_n \right) \right) \exp \left( - \mathbf{w}^\top (t) \mathbf{x}_n \right) \right] \mathbf{x}_k^\top \mathbf{r}(t) \\
\leq - t^{-1} \exp \left( - \bar{\mathbf{w}}^\top \mathbf{x}_k \right) \eta \mid \mathcal{S} \mid \left[ \left( 1 - t^{-\epsilon} \exp \left( - \epsilon \bar{\mathbf{w}}^\top \mathbf{x}_k \right) \right) \exp \left( \epsilon_2 \right) - 1 \right] \epsilon_2 \\
\leq - t^{-1} \exp \left( - \max_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) \eta \mid \mathcal{S} \mid \left[ \exp \left( \epsilon_2 \right) - 1 - t^{-\epsilon} \exp \left( \epsilon_2 - \epsilon \min_n \bar{\mathbf{w}}^\top \mathbf{x} \right) \right] \epsilon_2 . \\
\leq - C_0' t^{-1}, \forall t > t'' \\
\]
(A.29)

To conclude, we choose \( t_0 = \max \left[ t''_+, t''_x \right] \):

1. If \( \left\| \mathbf{P}_t \mathbf{r}(t) \right\| \geq \epsilon_1 \), we denote \( C_0'' \) as the minimum between \( C_0' \) (eq. A.29) and

\[
\eta \mid \mathcal{S} \mid \exp \left( - \max_n \bar{\mathbf{w}}^\top \mathbf{x}_n \right) (1 - e^{-0.5 \epsilon_2}) \epsilon_2 (eq. A.26). 
\]

Then we find that eq. A.21 can be upper bounded by \( -C_0'' t^{-1} \), \( \forall t > t_0 \), given eq. A.5. Substituting this result, together with eqs. A.19 and A.20 into eq. A.18, we obtain \( \forall t > t_0 \)
\[
\left( t (t + 1) - \mathbf{r}(t) \right)^\top \mathbf{r}(t) \leq -C_0'' t^{-1} + o \left( t^{-1} \right). 
\]

This implies that \( \exists t_2 < C_0'' \) and \( t_2 > t_0 \) such that eq. A.5 holds. This implies also that eq. A.4 holds for \( \left\| \mathbf{P}_t \mathbf{r}(t) \right\| \geq \epsilon_1 \).

2. If \( \left\| \mathbf{P}_t \mathbf{r}(t) \right\| < \epsilon_1 \), we find that \( \forall t > t_0 \), eq. A.21 can be upper bounded by either zero (eqs. A.25 and A.28), or terms proportional to \( t^{-\epsilon+1.5 \epsilon+} \) (eq. A.24) or \( t^{-1-0.5 \epsilon_2} \) (eq. A.27). Combining this together with eqs. A.19 and A.20 into eq. A.18, we obtain (for some positive constants \( C_5, C_4, C_3, C_6 \))
\[
\left( t (t + 1) - \mathbf{r}(t) \right)^\top \mathbf{r}(t) \leq C_5 t^{-1-1.5 \epsilon+} + C_4 t^{-1-0.5 \epsilon_2} + C_5 t^{-2} + C_0 t^{-\theta}. 
\]

Therefore, \( \exists t_1 > t_0 \) and \( C_1 \) such that eq. A.4 holds. \( \square \)
B Calculation of convergence rates

From Theorem 3, we can write $w(t) = \hat{w} \log t + \rho(t)$, where $\rho(t)$ has bounded norm.

Calculation of normalized weight vector (eq. 3.1):

\[
\frac{w(t)}{\|w(t)\|} = \frac{\rho(t) + \hat{w} \log t}{\sqrt{\rho(t) \mathbf{1} + \hat{w}^\top \hat{w} \log^2 t + 2 \rho(t) \hat{w} \log t}}
\]

\[
= \frac{\rho(t) / \log t + \hat{w}}{\|\hat{w}\| \sqrt{1 + 2 \rho(t) \hat{w} / (\|\hat{w}\|^2 \log t) + \|\rho(t)\|^2 / (\|\hat{w}\|^2 \log^2 t)}}
\]

\[
= \frac{1}{\|\hat{w}\|} \left( \frac{\rho(t)}{\log t} + \hat{w} \right) \left[ 1 - \frac{\rho(t) \hat{w}}{\|\hat{w}\|^2 \log t} + \left[ \frac{3}{4} \left( \frac{\rho(t) \hat{w}}{\|\hat{w}\|^2} \right)^2 - \frac{\|\rho(t)\|^2}{2 \|\hat{w}\|^2} \right] \frac{1}{\log^2 t} + O \left( \frac{1}{\log^3 t} \right) \right]
\]

(B.1)

where to obtain eq. B.1 we used $\frac{1}{\sqrt{1 + x}} = 1 - \frac{1}{2} x + \frac{3}{4} x^2 + O(x^3)$, and in the last line we used the fact that $\rho(t)$ has a bounded norm.

We use eq. 3.1 to calculate of angle (eq. 3.2):

\[
\frac{w(t)^\top \hat{w}}{\|w(t)\| \|\hat{w}\|} = \hat{w}^\top \left( \frac{\rho(t)}{\log t} + \hat{w} \right) \left[ 1 - \frac{\rho(t) \hat{w}}{\|\hat{w}\|^2 \log t} + \left[ \frac{3}{4} \left( \frac{\rho(t) \hat{w}}{\|\hat{w}\|^2} \right)^2 - \frac{\|\rho(t)\|^2}{2 \|\hat{w}\|^2} \right] \frac{1}{\log^2 t} + O \left( \frac{1}{\log^3 t} \right) \right]
\]

= \frac{1}{\|\hat{w}\|^2 \log t} \left[ 1 - \frac{3}{2} \left( \frac{\rho(t) \hat{w}}{\|\hat{w}\| \|\rho(t)\|} \right)^2 \right] \frac{1}{\log^2 t} + O \left( \frac{1}{\log^3 t} \right)

Calculation of margin (eq. 3.3):

\[
\min_n x_n^\top \hat{w} (t)
\]

= \min_n x_n^\top \left[ \frac{\hat{w}}{\|\hat{w}\|} + \left( \frac{\rho(t)}{\|\hat{w}\|} - \frac{\hat{w}}{\|\hat{w}\|} \rho(t) \hat{w} \right) \frac{1}{\log t} + O \left( \frac{1}{\log^2 t} \right) \right]

= \frac{1}{\|\hat{w}\|} + \frac{1}{\|\hat{w}\|} \left( \min_n x_n^\top \rho(t) \hat{w} \right) \frac{1}{\log t} + O \left( \frac{1}{\log^2 t} \right)

(B.2)

where in eq. B.2 we used eq. A.1
We examine multiclass classification. In the case the labels are the class index $y$, we can analytically integrate these equations to obtain an identity matrix. Note that $w = vec(W)$.

Furthermore, we define $w = vec(W^\top)$, a basis vector $e_k \in \mathbb{R}^K$ so that $(e_k)_i = \delta_{ki}$, and the matrix $A_k \in \mathbb{R}^{dK \times d}$ so that $A_k = e_k \otimes I_d$, where $\otimes$ is the Kronecker product and $I_d$ is the $d$-dimensional identity matrix. Note that $A_k^\top w = w_k$.

Consider the cross entropy loss with softmax output

$$L(w) = -\sum_{n=1}^{N} \log \left( \frac{\exp(w_{yn}^\top x_n)}{\sum_{k=1}^{K} \exp(w_k^\top x_n)} \right)$$

### C Softmax output with crossentropy loss

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

Furthermore, we define $w = vec(W^\top)$, a basis vector $e_k \in \mathbb{R}^K$ so that $(e_k)_i = \delta_{ki}$, and the matrix $A_k \in \mathbb{R}^{dK \times d}$ so that $A_k = e_k \otimes I_d$, where $\otimes$ is the Kronecker product and $I_d$ is the $d$-dimensional identity matrix. Note that $A_k^\top w = w_k$.
Using our notation, this loss can be re-written as
\[
\mathcal{L}(w) = - \sum_{n=1}^{N} \log \left( \frac{\exp (w^\top A_{y_n} x_n)}{\sum_{k=1}^{K} \exp (w^\top A_k x_n)} \right)
\]
\[
= \sum_{n=1}^{N} \log \left( \sum_{k=1}^{K} \exp (w^\top (A_k - A_{y_n}) x_n) \right)
\]  
(C.1)

Therefore
\[
\nabla \mathcal{L}(w) = \sum_{n=1}^{N} \sum_{k=1}^{K} \exp (w^\top (A_k - A_{y_n}) x_n) (A_k - A_{y_n}) x_n
\]
\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{\sum_{r=1}^{K} \exp (w^\top (A_r - A_k) x_n)} (A_k - A_{y_n}) x_n.
\]

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

**Assumption 4.** \( \exists w_s \) such that \( w_s^\top (A_k - A_{y_n}) x_n < 0 \forall k \neq y_n \) then the expression
\[
w_s^\top \nabla \mathcal{L}(w) = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{w_s^\top (A_k - A_{y_n}) x_n}{\sum_{r=1}^{K} \exp (w^\top (A_r - A_k) x_n)}.
\]

is strictly negative for any finite \( w \). However, from Lemma 5 in gradient descent with learning rate \( \eta > 2\beta^{-1} \), we have that \( \nabla \mathcal{L}(w(t)) \rightarrow 0 \). This implies that: \( \|w(t)\| \rightarrow \infty \), and \( \forall k \neq y_n, \forall r : w^\top (A_r - A_k) x_n \rightarrow \infty \), which implies \( \forall k \neq y_n, \max_k w^\top (A_k - A_{y_n}) x_n \rightarrow -\infty \). Examining the loss (eq. C.1) we find that \( \mathcal{L}(w(t)) \rightarrow 0 \) in this case. Thus, we arrive to an equivalent Lemma to Lemma 1 for this case:

**Lemma 7.** Let \( w(t) \) be the iterates of gradient descent (eq. 2.2) with \( \eta < 2\beta^{-1} \), for cross-entropy loss operating on a softmax output, under the assumption of strict linear separability (Assumption 4), then:

1. \( \lim_{t \rightarrow \infty} \mathcal{L}(w(t)) = 0 \)
2. \( \lim_{t \rightarrow \infty} \|w(t)\| = \infty \)
3. \( \forall n \exists \theta_n : w^\top (A_{y_n} - A_k) x_n > 0 \), and
4. \( \exists t_0, \forall t > t_0 : w^\top (A_{y_n} - A_k) x_n > 0 \), that is, within a finite time we reach a predictor that perfectly separates the data.

Therefore, since
\[
\mathcal{L}(w(t)) = \sum_{n=1}^{N} \log \left( \sum_{k=1}^{K} \exp \left( w^\top (A_k - A_{y_n}) x_n \right) \right)
\]
\[
\simeq \sum_{n=1}^{N} \log \left( 1 + \max_{k \neq y_n} \exp \left( w^\top (A_k - A_{y_n}) x_n \right) \right).
\]
(C.2)

where in the last line we assumed that \( (A_k - A_{y_n}) x_n \) has a unique minimum in \( k \) (which is true if \( x_n \) are in general position), since then the other exponential terms inside the log become negligible. If
\[
\arg \max_{k \neq y_n} \exp \left( w^\top (A_k - A_{y_n}) x_n \right)
\]
has a limit \( k_n \), then we define \( \hat{x}_n = (A_{y_n} - A_{k_n}) x_n \), so eq. C.2 is transformed to the standard logistic regression loss
\[
\sum_{n=1}^{N} \log \left( 1 + \exp \left( -w^\top \hat{x}_n \right) \right),
\]
to which our Theorems directly apply.

Therefore, \( w(t) / \|w(t)\| \rightarrow \hat{w} \) where
\[
\hat{w} = \arg \min_{w} \|w\|^2 \text{ s.t. } w^\top \hat{x}_n \geq 1
\]

Recalling that \( A^\top w = w_k \), we can re-write this as
\[
\arg \min_{w_1, \ldots, w_K} \sum_{k=1}^{K} \|w_k\|^2 \text{ s.t. } \forall n, \forall k \neq y_n : w_{y_n}^\top x_n \geq w_k^\top x_n + 1
\]
D Deep Networks, If Only a Single Layer Is Optimized

We examine a deep neural network (DNN) with \( m = 1, \ldots, L \) layers, piecewise linear activation functions \( f \) and loss function \( \ell \) following assumption [3] parameterized by weights matrices \( W_l \). Since \( f \) are piecewise linear, we can write for almost every \( u \): \( f_l(u) = \nabla f_l(u) \odot u \) (an element-wise product). Given an input sample \( x_n \), for each layer \( l \) the input \( u_{n,l} \) and output \( v_{n,l} \) are calculated sequentially in a “forward propagation”

\[
\begin{align*}
u_{n,0} &= x_n. \quad \text{Then, given the DNN output } u_{n,L} \text{ and target } y_n \in \{-1, 1\} \text{ the loss } \\
\ell(y_n, u_{n,L}) \text{ can be calculated. During training, the gradients of the loss are calculated using the chain rule in a “back-propagation”}
\end{align*}
\]

\[
\begin{align*}
\delta_{n,l-1} &= \left[ \nabla f_l(u_{n,l}) \right] W_l^\top \delta_{n,l} \\
\end{align*}
\]

initialized by \( \delta_{n,L} = 1 \). Finally, the weights are updated with GD. The basic update (without weight sharing) is

\[
\begin{align*}
W_l(t+1) - W_l(t) &= -\eta \sum_{n=1}^N \frac{\partial}{\partial W_l} \ell(y_n, u_{n,L}) \\
&= -\eta \sum_{n=1}^N y_n \ell' \left( y_n, u_{n,L} \right) \delta_{n,l} v_{n,l-1}^\top \\
&= -\eta \sum_{n=1}^N y_n \ell' \left( y_n, \delta_{n,l}^\top W_l v_{n,l-1} \right) \delta_{n,l} v_{n,l-1}^\top ,
\end{align*}
\]

where in the last line we used

\[
\forall l : u_{n,L} = W_L \left( \prod_{m=1}^{L-1} \nabla f_l(u_{n,m}) W_m \right) x_n = \delta_{n,l}^\top W_l v_{n,l-1} .
\]

Denoting \( \bar{x}_{n,l} = y_n \delta_{n,l} \odot v_{n,l-1} \) and \( w_l = \text{vec} \left( W_l^\top \right) \) we obtain

\[
\begin{align*}
w_l(t+1) - w_l(t) &= -\eta \sum_{n=1}^N \ell' \left( w_l^\top \bar{x}_{n,l} \right) \bar{x}_{n,l} .
\end{align*}
\]

We got the same update as in eq. 2.2 Thus, if \( \bar{x}_{n,l} \) does not change between iterations and becomes linearly separable so the training error can go to zero, we can apply Theorem 3. This can happen if optimize only \( W_l \), and the activation units stop crossing their thresholds, after a sufficient number of iterations.

E An Experiment with Stochastic Gradient Descent

Figure E.1: Same as Fig. 3.1 except stochastic gradient decent is used (with mini-batch of size 4), instead of GD.