Inductive Bias of Gradient Descent for Weight Normalized Smooth Homogeneous Neural Nets

Depen Morwani DEPENMORWANI@GMAIL.COM
Harish G. Ramaswamy HARIGURU@CSE.IITM.AC.IN
Department of Computer Science and Engineering, and RBCDSAI,
Indian Institute of Technology Madras, India.

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

We analyze the inductive bias of gradient descent for weight normalized smooth homogeneous neural nets, when trained on exponential or cross-entropy loss. We analyse both standard weight normalization (SWN) and exponential weight normalization (EWN), and show that the gradient flow path with EWN is equivalent to gradient flow on standard networks with an adaptive learning rate. We extend these results to gradient descent, and establish asymptotic relations between weights and gradients for both SWN and EWN. We also show that EWN causes weights to be updated in a way that prefers asymptotic relative sparsity. For EWN, we provide a finite-time convergence rate of the loss with gradient flow and a tight asymptotic convergence rate with gradient descent. We demonstrate our results for SWN and EWN on synthetic data sets. Experimental results on simple datasets support our claim on sparse EWN solutions, even with SGD. This demonstrates its potential applications in learning neural networks amenable to pruning.

Keywords: Deep Learning Theory, Inductive Bias, Gradient Descent, Weight Normalization

1. Introduction

The prevailing hypothesis for explaining the generalization ability of deep neural nets, despite their ability to fit even random labels (Zhang et al., 2017), is that the optimisation/training algorithms such as gradient descent have a 'bias' towards 'simple' solutions. This property is often called inductive bias, and has been an active research area over the past few years.

It has been shown that gradient descent does indeed seem to prefer 'simpler' solutions over more 'complex' solutions, where the notion of complexity is often problem/architecture specific. The predominant line of work typically shows that gradient descent prefers a least norm solution in some variant of the L_2-norm. This is satisfying, as gradient descent over the parameters abides by the rules of L_2 geometry, i.e. the weight vector moves along direction of steepest descent, with length measured using the Euclidean norm. However, there is nothing special about the Euclidean norm in the parameter space, and hence several other notions of 'length' and 'steepness' are equally valid. In recent years, several alternative parameterizations of the weight vector, such as Batch normalization and Weight normalization, have seen immense success and these do not seem to respect L_2 geometry in the 'weight space'. We pose the question of inductive bias of gradient descent for some of these parameterizations, and demonstrate interesting inductive biases. In particular, it can still be argued that gradient descent with these reparameterizations prefers simpler solutions, but the notion of complexity is different.

Our Contributions. The main contributions of the paper are as follows.

- We establish that the gradient flow path with exponential weight normalization is equal to the gradient flow path of an unnormalized network using an adaptive neuron dependent learning rate. This provides a crisp description of the difference between exponential weight normalized networks and unnormalized networks.
- While most of the previous works on the inductive bias of non-linear deep learning architectures work under the assumption of directional convergence of weights and gradients, we show that gradient convergence implies weight convergence, even for gradient descent, for both standard and exponentially weight normalized network (which is not homogeneous in its parameters)[1].
- We establish the asymptotic relations between weights and gradients for gradient descent on standard weight normalized and exponentially weight normalized networks and show that exponential weight normalization is likely to lead to asymptotic sparsity in weights. We demonstrate the relative sparsity of exponential weight normalization on MNIST dataset, by showing that it leads to networks with better pruning efficacy.
- We establish finite-time convergence rates for gradient flow and tight asymptotic convergence rates for gradient descent on exponentially weight normalized networks.

2. Related Work

The literature most closely related to this paper can be broadly classified into two categories - the inductive biases established for neural networks, and the theoretical studies of normalization methods in deep learning.

2.1. Inductive Bias

Soudry et al. (2018) showed that gradient descent(GD) on the logistic loss with linearly separable data converges to the L_2 maximum margin solution for almost all datasets. These results were extended to loss functions with super-polynomial tails in Nacson et al. (2019b). Nacson et al. (2019c) extended these results to hold for stochastic gradient descent(SGD) and Gunasekar et al. (2018a) extended the results for other optimization geometries. Ji and Telgarsky (2019b) provided tight convergence bounds in terms of dataset size as well as training time. Ji and Telgarsky (2019a) provide similar results when the data is not linearly separable.

Ji and Telgarsky (2019c) showed that for deep linear nets, under certain conditions on the initialization, for almost all linearly separable datasets, the network, in function space, converges to the maximum margin solution. Gunasekar et al. (2018b) established that for linear convolutional nets, under certain assumptions regarding convergence of gradients etc, the function converges to a KKT point of the maximum margin problem in fourier space. Nacson et al. (2019a) shows that for smooth homogeneous nets, the network converges to a KKT point of the maximum margin problem in parameter space. Lyu and Li (2020) established these results with weaker assumptions and also provide asymptotic convergence rates for the loss. Chizat and Bach (2020) explore the inductive bias for a 2-layer infinitely wide ReLU neural net in function space and show that the function learnt is a

1. The assumptions about weights and gradients converging in direction have been recently shown to hold for gradient flow on homogeneous neural nets without normalization under some regularity conditions related to o-minimality of the architecture(Ji and Telgarsky, 2020)

max-margin classifier for variation norm. Moroshko et al. (2020) established the inductive bias for linear diagonal networks and showed that the network transitions between max-margin and L_1-norm margin, depending on the relation between the initialization scale and training accuracy.

2.2. Normalization

Salimans and Kingma (2016) introduced weight normalization and demonstrated that it replicates the convergence speedup of BatchNorm. Similarly, other normalization techniques have been proposed as well(Ba et al., 2016)(Qiao et al., 2020)(Li et al., 2019), but only a few have been theoretically explored. Santurkar et al. (2018) demonstrated that batch normalization makes the loss surface smoother and L_2 normalization in batchnorm can even be replaced by L_1 and L_∞ normalizations. Kohler et al. (2019) showed that for GD, batchnorm speeds up convergence in the case of GLM by splitting the optimization problem into learning the direction and the norm. Cai et al. (2019) analyzed GD on BN for squared loss and showed that it converges for a wide range of lr. Bjorck et al. (2018) showed that the primary reason BN allows networks to achieve higher accuracy is by enabling higher learning rates. Arora et al. (2019) showed that in case of GD or SGD with batchnorm, lr for scale-invariant parameters does not affect the convergence rate towards stationary points. Du et al. (2018) showed that for GD over one-hidden-layer weight normalized CNN, with a constant probability over initialization, iterates converge to global minima. Qiao et al. (2019) compared different normalization techniques from the perspective of whether they lead to points, where neurons are consistently deactivated. Wu et al. (2020) established the inductive bias of gradient flow with weight normalization for overparameterized least squares and showed that for a wider range of initializations as compared to normal parameterization, it converges to the minimum L_2 norm solution. Dukler et al. (2020) analyzed weight normalization for multilayer ReLU net in the infinite width regime and showed that it may speedup convergence. Some other papers(Luo et al., 2019; Roburin et al., 2020) also provide other perspectives to think about normalization techniques.

3. Problem Setup

We use a standard view of neural networks as a collection of nodes/neurons grouped by layers. Each node u is associated with a weight vector w_u, that represents the incoming weight vector for that node. In case of CNNs, weights can be shared across different nodes. w represents all the parameters of the network arranged in form of a vector (In general, for any vector v associated with the entire network, v_u represent its components corresponding to the node u). The training dataset consists of (x_i, y_i) pairs with a total of m points in the dataset. The function represented by the neural network is denoted by $\phi(w, \cdot)$. The loss for a single data point x_i is given by $\ell(y_i, \phi(w, x_i))$ and the loss vector is represented by ℓ. The overall loss is represented by $L(w)$ and is given by $L(w) = \sum_{i=1}^{m} \ell(y_i, \phi(w, x_i))$. We sometimes abbreviate $L(w(t))$ as L when the context is clear.

In standard weight normalisation (SWN), each weight vector w_u is reparameterized as $g_u \frac{v_u}{\|v_u\|}$. This was proposed by Salimans and Kingma (2016), as a substitute for Batch Normalization and has been practically used in multiple papers such as Sokolic et al. (2017), Dauphin et al. (2017), Kim et al. (2018) and Hieber et al. (2018). The corresponding update equations for gradient descent

are given by

$$g_u(t+1) = g_u(t) - \eta(t)\frac{v_u(t)^\top \nabla_{w_u} L}{\|v_u(t)\|} \tag{1}$$

$$v_u(t+1) = v_u(t) - \eta(t)\frac{g_u(t)}{\|v_u(t)\|}\left(I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\right)\nabla_{w_u} L \tag{2}$$

In exponential weight normalisation (EWN), each weight vector is reparameterized as $w_u = \frac{v_u}{\|v_u\|}$. This was mentioned in Salimans and Kingma (2016), but to the best of our knowledge, has not been widely used. The corresponding update equations for gradient descent with learning rate $\eta(t)$ are given by

$$g_u(t+1) = g_u(t) - \eta(t)e^{g_u(t)}\frac{v_u(t)^\top \nabla_{w_u} L}{\|v_u(t)\|} \tag{3}$$

$$v_u(t+1) = v_u(t) - \eta(t)\frac{e^{g_u(t)}}{\|v_u(t)\|}\left(I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\right)\nabla_{w_u} L \tag{4}$$

The update equations for gradient flow are the continuous counterparts for the same. In gradient flow, for both SWN and EWN, we set $\|v_u(0)\| = 1$, to simplify the update equations.

4. Inductive Bias of Weight Normalization

In this section, we state our main results for weight normalized smooth homogeneous models on exponential loss $\ell(y_i, f(w, x_i)) = e^{-y_i f(w,x_i)}$. The results for cross-entropy loss and proofs have been deferred to the appendix due to space constraints. First, we state the main proposition that helps in establishing these results for EWN.

Theorem 1 The gradient flow path with learning rate $\eta(t)$ for EWN and SWN are given as follows:

$$\text{EWN: } \frac{dw_u(t)}{dt} = -\eta(t)\|w_u(t)\|^2 \nabla_{w_u} L \tag{5}$$

$$\text{SWN: } \frac{dw_u(t)}{dt} = -\eta(t)\left[\|w_u(t)\|^2 \nabla_{w_u} L + \frac{1 - \|w_u(t)\|^2}{\|w_u(t)\|^2}\left(w_u(t)^\top \nabla_{w_u} L\right) w_u(t)\right] \tag{6}$$

Thus, the gradient flow path of EWN can be replicated by an adaptive (neuron dependent) learning rate given by $\eta(t)\|w_u(t)\|^2$ on the unnormalized network (Unnorm).

4.1. Assumptions

The assumptions in the paper can be broadly divided into loss function/architecture based assumptions and trajectory based assumptions. The loss functions/architecture based assumptions are shared across both gradient flow and gradient descent.

Loss function/Architecture based assumptions

1 $\ell(y_i, f(w, x_i)) = e^{-y_i f(w,x_i)}$

2 $f(\cdot, x)$ is a C^1 function (i.e. continuously differentiable), for any fixed x

$3\ (\ \ w;x) = \ ^{L}(\ w;x)$, for some > 0 and $L > 0$

Gradient ow . For gradient ow, we make the following trajectory based assumptions

(A1) There exists a time t_0 such that $L(w(t_0)) < 1$.

(A2) $\lim_{t!1} \frac{r\ _w L(w(t))}{kr\ _w L(w(t))k} := g$.

The rst trajectory assumption is simply a separability assumption and means that the network is able to correctly classify the dataset at some point during the training process. This is not a completely unreasonable assumption, given recent papers demonstrating neural networks with sufcient overparameterization can t even random labels (Zhang et al., 2017; Jacot et al., 2018). The second assumption has been used in multiple previous works (Gunasekar et al., 2018b; Chizat and Bach, 2020; Nacson et al., 2019a), and is standard in the literature related to the inductive bias of non-linear deep learning architectures. Moreover, we remove one of the assumptions related to directional convergence of weights used in these works, and instead show that it is implied by the directional convergence of gradients.

Gradient Descent For gradient descent, we require the learning rate (t) to not grow too fast, and a slightly stronger assumption on loss.

(B1) $\lim_{t!1} L(w(t)) = 0$ (B2) $\lim_{t!1} \frac{r\ _w L(w(t))}{kr\ _w L(w(t))k} := g$

(B3) $\lim_{t!1} (t)kw_u(t)kr\ _{w_u} L(w(t))k = 0$ for all u in the network.

The assumption (B3) is mild, as the norm of the gradient of the exponential loss goes down exponentially fast as compared to norm of the weights. We demonstrate that these assumptions hold for multiple datasets including MNIST in Appendix P.

4.2. Asymptotic relations between weights and gradients

This section contains the main theorems that establish asymptotic relations between weights and gradients for SWN and EWN. First, we will state a common proposition for both SWN and EWN.

Proposition 2 Under assumption (A1) for gradient ow, for both SWN and EWN, $\lim_{t!1} L(w(t)) = 0$.

Although the above proposition was established for homogeneous nets by Lyu and Li[2] (2020) extend it for the non-homogeneous parameterization of EWN. Now, we provide one of our main theorem that establishes gradient convergence implies weight convergence.

Theorem 3 Consider a node u in the network with $kg_u k > 0$ and $\lim_{t!1} kw_u(t)k = 1$. Under assumptions (A1), (A2) for gradient ow and (B1)-(B3) for gradient descent, for both SWN and EWN

(i) $\lim_{t!1} \frac{w_u(t)}{kw_u(t)k} := w_u$ exists. (ii) $w_u = g_u$ for some > 0.

The above theorem relaxes one of the assumptions regarding weight convergence used in many of the previous works, by showing that even for non-homogeneous parameterization under gradient descent, gradient convergence implies weight convergence. Moreover, it also shows, that weights and gradients eventually get aligned opposite to each other.

2. Homogeneous networks in the w space are also homogeneous in the v space. Therefore results regarding convergence rates and monotonic margin hold from Lyu and Li (2020). However, the results for convergence to a KKT point of the max margin problem do not hold. For details, refer Appendix L

(a) (b) (c)

Figure 1: Demonstration of Results for EWN in Lin-Sep experiment: (a) Evolution of $\|w_u\|$ - norm of the incoming weights for neuron u (b) Cosine between weights and gradients for neurons 5, 7 and 8. (c) Weight and gradient norms for weights 5, 7 and 8.

Now, we provide the main theorem that distinguishes SWN and EWN.

Theorem 4 Consider two nodes u and v in the network with $\|g_u\| \neq \|g_v\| > 0$; $\lim_{t\to\infty} \|w_u(t)\| = \infty$ and $\lim_{t\to\infty} \|w_v(t)\| = \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c. Under assumptions (A1), (A2) for gradient flow and (B1)-(B3) for gradient descent,

(i) for SWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

(ii) for EWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|}$ is either $0, \infty$ or $\frac{1}{c}$

Thus, if $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ converges to a finite non-zero value for EWN, then $\|w_u(t)\|\|\nabla_{w_u} L(w(t))\| = \|w_v(t)\|\|\nabla_{w_v} L(w(t))\| = k_1(t)$ asymptotically. While, for SWN, $\frac{\|w_u(t)\|}{\|\nabla_{w_u} L(w(t))\|} = \frac{\|w_v(t)\|}{\|\nabla_{w_v} L(w(t))\|} = k_2(t)$ asymptotically, where $k_1(t)$ and $k_2(t)$ are independent of u and v. The exact conditions under which $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ tends to 0 or ∞ for EWN are provided in Proposition 5.

We demonstrate Theorem 3 and Theorem 4 for EWN on a linearly separable dataset (Lin-Sep) in Figure 1. In this experiment, a 2-layered neural network, with 8 neurons in the hidden layer and a ReLU-squared activation function, is trained on a linearly separable dataset. The learning rate schedule used was $\frac{1}{L^{0.97}}$ and the network was trained till a loss of 10^{-300}. As can be seen in Figure 1, for weights 5, 7 and 8, whose norms keep on growing, weights and gradients eventually become oppositely aligned, and their norms are inversely proportional to each other. The results for SWN, along with results on other datasets including MNIST have been deferred to Appendix N.

4.3. Sparsity Inducing Bias for Exponential Weight Normalisation

The inverse relation between $\|w_u(t)\|$ and $\|\nabla_{w_u} L(w(t))\|$ in the EWN trajectory results in an interesting inductive bias that favours movement along sparse directions.

Proposition 5 Consider two nodes u and v in the network such that $\|g_v\| \neq \|g_u\| > 0$ and $\|w_u(t)\|, \|w_v(t)\| \to \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c (note that g and c will be different for SWN and EWN trajectory). Consider any ϵ such that $0 < \epsilon < c$ and $0 < \epsilon < \frac{c}{2}$. Then, the following holds:

Figure 2: (a) Network architecture for the Simple-Traj experiment . (b) Trajectories of the two weights for EWN and Unnorm, starting from 5 different initialization points.

(i) There exists a time t_1, such that for all $t > t_1$ both SWN and EWN trajectories have the following properties:

(a) $\frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_v} L(w(t))\|} \in [c - \epsilon, c + \epsilon]$ (b) $\frac{w_u(t)}{\|w_u(t)\|} \cdot \frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|} \geq \cos(\theta)$

(c) $\frac{w_v(t)}{\|w_v(t)\|} \cdot \frac{\nabla_{w_v} L(w(t))}{\|\nabla_{w_v} L(w(t))\|} \geq \cos(\theta)$.

(ii) for SWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

(iii) for EWN, if at some time $t_2 > t_1$,

(a) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} > \frac{1}{(c-\epsilon)\cos(\theta)} \Rightarrow \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 1$

(b) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} < \frac{\cos(\theta)}{c+\epsilon} \Rightarrow \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 0$

The above proposition shows that the limit property of the weights in Theorem 4, makes non-sparse w an unstable convergent direction for EWN. But that is not the case for SWN. We demonstrate the relative sparsity between EWN, SWN and Unnorm through two toy experiments Simple-Traj and XOR.

In the Simple-Traj experiment, we illustrate the notion of asymptotic relative sparsity occuring in the EWN parameterization. In this dataset, we have a single data point $(2, 1)$, that is labelled positive and train a network with linear activations. The architecture is shown in Figure 2, where weights in blue and red are frozen to values 1 and 0 respectively. Thus, there are effectively only two scalar parameters w_1 and w_2. The network is trained till a loss value of e^{-50} starting from 5 different initialization points. The weight trajectories in Figure 2 shows that EWN prefers to converge either along the x or y axis, and hence has an asymptotic relative sparsity property. We provide a theoretical proof for the same in the general d-dimensional case.

Proposition 6 Consider a linear model over $\mathbb{R}^d$ given by $f(x) = w^\top x$, where each w_i is further reparameterized as v_i. Consider a dataset consisting of a single data point $(0, 1)$, that is labelled

(a) EWN (b) SWN (c) Unnorm

Figure 3: (a), (b) and (c) demonstrate the evolution of weight norms for each neuron in the XOR experiment. EWN weights grow sparsely when compared to Unnorm and SWN

(a) EWN (b) SWN (c) Unnorm

Figure 4: Variation of convergence rate of train loss with number of layers for multilayer linear nets on a linearly separable dataset

as +1. According to the initialization of , de ne a relation R on $\{1, \ldots, d\}$, given by $i \; j$ if $w_i(0) z_i = w_j(0) z_j$. Then, R is an equivalence relation on $\{1, \ldots, d\}$. Let these equivalent sets be denoted by $I_1, I_2, \ldots, I_k$. De ne a total order on these sets given by $I_a > I_b$ if $\exists i \in I_a, j \in I_b$ such that $w_i(0) z_i > w_j(0) z_j$. Let the maximum set according to this order be denoted by . Then, for gradient ow on exponential loss, the following holds

(i) For any $i \in I$, $\lim_{t \to \infty} w_i(t) = 1$

(ii) For $i, j \in I$, $\frac{w_i(t)}{w_j(t)} = \frac{x_j}{x_i}$.

(iii) For any $i \notin I$, $\lim_{t \to \infty} w_i(t) = \left(\frac{1}{w_i(0)} \; \frac{x_i}{w_j(0) x_j} \right)^{\frac{1}{2}}$, where j is any element in .

Thus, if w is initialized from a continuous distribution, then with probability 1, the cardinality of is 1 and hence $\hat{w}(t) = w(t) / \|w(t)\|$ will approach a sparse vector.

In the XOR experiment, we train a 2-layer ReLU network with 20 hidden neurons on XOR dataset, till a loss value of 10^{-50}. The second layer is xed to the values 1 or -1 randomly. For attaining 100% accuracy on this dataset with this architecture, at least 4 hidden units are needed. As

can be seen in Figure 3, EWN asymptotically uses exactly 4 neurons out of 20, while Unnorm and SWN use almost all the 20 neurons.

5. Convergence Rates

In this section, we provide convergence rate of loss for EWN.

Gradient Flow: We provide a nite-time convergence rate of loss for gradient ow in case of EWN.

Theorem 7 For Exponential Weight Normalization, under assumption (A1), the following hold for $t > t_0$ in case of gradient ow

(i) $\|w(t)\|$ grows witht as$o((\log t)^{\frac{1}{L}})$ (ii) $L(t)$ goes down with as $O\left(\frac{1}{t}\right)$

Gradient Descent: For establishing convergence rates for gradient descent, we are going to make an additional assumption that the overall weight vector converges in direction, i.e, $\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|}$ exists (B4). Although we have already shown this is indeed true for nodes with $g(t) > 0$, we need this assumption to take into account the nodes with $g(t) = 0$. Under this assumption, w can be represented as $w = g(t)\hat{w} + r(t)$, where$\lim_{t\to\infty} \frac{\|r(t)\|}{g(t)} = 0$. Let $d : \mathbb{N} \to \mathbb{R}$, given by $d(t) = \sum_{\tau=0}^{t-1} \eta(\tau)$ denote total step size. Let $\gamma = \min_i y_i (\hat{w}, x_i)$ be the normalized margin at convergence.

The asymptotic convergence rate of loss for SWN and Unnorm have already been established in Lyu and Li (2020) as $\frac{1}{d(t)(\log d(t))^{\frac{2}{L}}}$. For EWN, the corresponding theorem is provided below

Theorem 8 For Exponential Weight Normalization, under Assumptions (B1)-(B4), $\eta(t) = O\left(\log\frac{1}{L}^{c}\right)$ for $c < 1$ and$\lim_{t\to\infty} \frac{\|r(t+1)\|}{g(t+1)} - \frac{\|r(t)\|}{g(t)} = 0$, the following hold

(i) $\|w(t)\|$ asymptotically grows with as $(\log d(t))^{\frac{1}{L}}$

(ii) $L(w(t))$ asymptotically goes down with as $\frac{1}{d(t)(\log d(t))^2}$.

Although the additional assumption$\lim_{t\to\infty} \frac{\|r(t+1)\|}{g(t+1)} - \frac{\|r(t)\|}{g(t)} = 0$ is not standard, we empirically demonstrate that, for EWN, the convergence rate is almost independent of the number of layers. Moreover, the learning rate assumption used still covers the constant$\eta(t)$ case, that is generally used in practice.

For multilayer linear nets, the variation of convergence rate with number of layers for a linearly separable dataset is illustrated in Figure 4. All of these networks were explicitly initialized to represent the same point in function space. It can be seen that EWN, SWN and unnormalized networks all converge faster with more layers, but the effect is much less pronounced for EWN.

6. Pruning Experiments

As EWN leads to asymptotically sparse solutions, it is likely that a suf ciently trained EWN network would be comparatively robust to pruning. In this section, we compare the pruning ef cacy of

(a) $L = e^{-10}$ (b) $L = e^{-100}$ (c) $L = e^{-300}$

Figure 5: Variation of test accuracy vs percentage of neurons pruned in first layer at different loss values for MNIST experiment

EWN, SWN and Unnorm on MNIST (LeCun et al., 2010) dataset. We use a 2-layer ReLU network with 1024 neurons in the hidden layer. In case of EWN and SWN, only the first layer is weight normalized as only this layer needs to be pruned.

Pruning Strategy: The natural pruning strategy of removing neurons with least $\|w_u\|$ gives inordinate importance to the initialisation and the initial optimization epochs. In order to minimize the effect of initialization and initial movement of weights, we prune according to the weight norm increase from a reference point. For example, when pruning at a loss value of e^{-300}, we consider 4 reference points, 0, weight at initialization, weight when $L = e^{-10}$ and $L = e^{-100}$. We then choose the pruning strategy that gives maximum testing accuracy for a given level of pruning. Similarly, for pruning at a loss value of e^{-100}, we consider three reference points, 0, weight at initialization and weight when $L = e^{-10}$, and for pruning at a loss value of e^{-10}, we consider 0, weight at initialization and weight when $L = e^{-5}$. More detailed description of the pruning strategy is provided in Appendix Q.

The pruning graphs for MNIST at different loss values averaged across multiple seeds are shown in Figure 5. It can be seen that when the loss levels are sufficiently low, the EWN network becomes better adapted for pruning, significantly outperforming SWN and the unnormalized network in terms of test accuracy for a given level of pruning. Further details along with convergence rate plots are provided in Appendix O.

6.1. Combining EWN with other sparsity regularizers

EWN by itself has a sparsifying effect, and in addition it can be combined with other sparsity regularizers. e.g. the neural net can be trained using any existing sparsity regularizer with standard parameterization for the initial few epochs, and use EWN (without the regularizer) for the later phase.

We conducted experiment on MNIST dataset, by initially training the network till convergence with $\ell_{2,1}$ group sparsity regularizer ($\sum_{u \in \text{nodes of network}} \|w_u\|$), and later on using EWN till a loss value of e^{-30}. In this case, the pruning strategy is based on the final norms of the weights as the sparsity regularizer already induces an asymmetry among different neurons. The pruning results are

(a) (b)

Figure 6: (a) Variation of test accuracy vs percentage of neurons pruned on MNIST dataset after training the network initially with $\ell_{2,1}$ regularizer and later on continuing training with either EWN, SWN or unnormalized net with $\lambda = e^{-30}$ (b) Zoomed in view on EWN

shown in Figure 6. As can be seen, training further with EWN improves the pruning efficacy of the network.

7. Proof Sketch

In this section, we provide the proof for part (iii)a of Proposition 5 for gradient flow, i.e, for EWN, $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} > \frac{1}{(c-\epsilon)\cos(\theta)} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = \infty$. Remaining proofs are given in appendix.

The update equations for w_u and v_u in case of gradient flow for EWN are given by

$$\frac{d g_u(t)}{dt} = \eta(t) e^{-g_u(t)} \frac{v_u(t)^\top \nabla_{w_u} L}{\|v_u(t)\|} \tag{7}$$

$$\frac{dv_u(t)}{dt} = \eta(t) \frac{e^{-g_u(t)}}{\|v_u(t)\|} \left(I - \frac{v_u(t) v_u(t)^\top}{\|v_u(t)\|^2} \right) \nabla_{w_u} L \tag{8}$$

Using Equation (7) (along with the fact that $\|v_u(t)\|$ does not change with time and $\|v_u(0)\| = 1$),

$$\frac{d\|w_u(t)\|}{dt} = \frac{de^{g_u(t)}}{dt} = \eta(t) \|w_u(t)\|^2 (v_u(t)^\top \nabla_{w_u} L(w(t))) \tag{9}$$

Using Equation (9) and part 1 of Proposition 5, we can say for $t \geq t_1$,

$$\begin{aligned} \frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} = & \frac{\|w_v(t)\| \frac{d\|w_u(t)\|}{dt} - \|w_u(t)\| \frac{d\|w_v(t)\|}{dt}}{\|w_v(t)\|^2} \\ & \eta(t) \frac{\|w_u(t)\|}{\|w_v(t)\|} (\|w_u(t)\| \|\nabla_{w_u} L(w(t))\| \cos(\theta) - \|w_v(t)\| \|\nabla_{w_v} L(w(t))\|) \\ & \eta(t) \|w_u(t)\| \|\nabla_{w_u} L(w(t))\| \left(\frac{\|w_u(t)\|}{\|w_v(t)\|} \cos(\theta) - \frac{1}{c} \right) \end{aligned} \tag{10}$$

In this case, using Equation (10), we can see $\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} > 0$ at t_2. Thus, $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ always remains greater than $\frac{1}{(c\ )\cos(\ )}$ and keeps on increasing. Let's denote $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|}$ by . Then, for $t > t_2$,

$$\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} \quad \cos(\ ) \quad \frac{1}{c} \quad (t)\|w_u(t)\|\|r_{w_u}L(w(t))\|$$

As $_u \to 1$, $\int_{t_2}^{R_1} (t)\|w_u(t)\|\|r_{w_u}L(w(t))\|dt = 1$ using Equation (7). Thus, integrating both the sides of the equation above from t_2 to 1 , we get

$$\int_{t_2}^{1} \frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt}dt \quad 1$$

Thus $\lim_{t\to 1} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 1$. A similar proof works for the part (iii)b of Proposition 5 as well.

8. Conclusion

In this paper, we analyze the inductive bias of weight normalization for smooth homogeneous neural nets and show that exponential weight normalization is likely to lead to asymptotically sparse solutions and has a faster convergence rate than unnormalized or standard weight normalized networks.

The smooth homogeneity assumptions made in the paper are satisfied by any positive power of ReLU greater than 1. The primary issue with ReLU is that the assumption of gradients converging in direction does not make sense for the non-smooth case. However, as our experiments demonstrate, the implication of relative sparsity holds for ReLU activation as well. Therefore, extending the results in the paper for the non-smooth case is a promising research direction.

Although the trajectory based assumptions have been shown to hold for gradient flow on unnormalized nets under certain regularity conditions, establishing similar conditions for weight normalized networks remains an open question. Moreover, extending the directional convergence results from gradient flow to gradient descent is also an interesting research direction.

References

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkxQ-nA9FX .

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7694–7705. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7996-understanding-batch-normalization.pdf .

Yongqiang Cai, Qianxiao Li, and Zuowei Shen. A quantitative analysis of the effect of batch normalization on gradient descent. volume 97 of Proceedings of Machine Learning Research, pages 882–890, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/cai19a.html.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. volume 125 of Proceedings of Machine Learning Research, pages 1305–1338. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/chizat20a.html.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. volume 70 of Proceedings of Machine Learning Research, pages 933–941, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/dauphin17a.html.

Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. volume 80 of Proceedings of Machine Learning Research, pages 1339–1348, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/du18b.html.

Yonatan Dukler, Quanquan Gu, and Guido Montúfar. Optimization theory for relu neural networks trained with normalization layers. In International Conference on Machine Learning, 2020.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. volume 80 of Proceedings of Machine Learning Research, pages 1832–1841, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018a. PMLR. URL http://proceedings.mlr.press/v80/gunasekar18a.html.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 9461–9471. Curran Associates, Inc., 2018b. URL http://papers.nips.cc/paper/8156-implicit-bias-of-gradient-descent-on-linear-convolutional-networks.pdf.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. Sockeye: A toolkit for neural machine translation, 2018.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 8571–8580. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8076-neural-tangent-kernel-convergence-and-generalization-in-neural-networks.pdf.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. volume 99 of Proceedings of Machine Learning Research, pages 1772–1798, Phoenix, USA, 25–28 Jun 2019a. PMLR. URL http://proceedings.mlr.press/v99/ji19a.html.

Ziwei Ji and Matus Telgarsky. A re ned primal-dual analysis of the implicit bias, 2019b.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In International Conference on Learning Representations, 2019c. URL https://openreview.net/forum?id=HJflg30qKX .

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. arXiv preprint arXiv:2006.06657, 2020.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, volume 31, pages 1564–1574. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/96ea64f3a1aa2fd00c72faacf0cb8ac9-Paper.pdf .

Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Ming Zhou, and Klaus Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. volume 89 of Proceedings of Machine Learning Research, pages 806–815. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/kohler19a.html .

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d`Alch-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 1622–1634. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8440-positional-normalization.pdf .

Ping Luo, Xinjiang Wang, Wenqi Shao, and Zhanglin Peng. Towards understanding regularization in batch normalization. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=HJlLKjR9FQ .

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS .

Edward Moroshko, Suriya Gunasekar, Blake Woodworth, Jason D. Lee, Nathan Srebro, and Daniel Soudry. Implicit bias in deep linear classi cation: Initialization scale vs training accuracy, 2020.

M. Muresan. A Concrete Approach to Classical Analysis. CMS Books in Mathematics. Springer New York, 2015. ISBN 9780387789330. URL https://books.google.co.in/books?id=N8rBgtIu_qgC .

Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. volume 97 of Proceedings of Machine Learning Research, pages 4683–4692, Long Beach, California, USA, 09–15 Jun 2019a. PMLR. URL http://proceedings.mlr.press/v97/nacson19a.html .

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. volume 89 of Proceedings of Machine Learning Research, pages 3420–3428. PMLR, 16–18 Apr 2019b. URL http://proceedings.mlr.press/v89/nacson19b.html .

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. volume 89 of Proceedings of Machine Learning Research, pages 3051–3059. PMLR, 16–18 Apr 2019c. URL http://proceedings.mlr.press/v89/nacson19a.html .

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Rethinking normalization and elimination singularity in neural networks. arXiv preprint arXiv:1911.09738, 2019.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization, 2020.

Simon Roburin, Yann de Mont-Marin, Andrei Bursuc, Renaud Marlet, Patrick Pérez, and Mathieu Aubry. Spherical perspective on learning with batch norm. arXiv preprint arXiv:2006.13382, 2020.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 901–909. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6114-weight-normalization-a-simple-reparameterization-to-accelerate-training-of-deep-neural-networks.pdf .

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 2483–2493. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7515-how-does-batch-normalization-help-optimization.pdf .

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. IEEE Transactions on Signal Processing, 65(16):4265–4280, Aug 2017. ISSN 1941-0476. doi: 10.1109/tsp.2017.2708039. URL http://dx.doi.org/10.1109/TSP.2017.2708039 .

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=r1q7n9gAb .

Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, and Qiang Liu. Implicit regularization and convergence for weight normalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 2835–2847. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1de7d2b90d554be9f0db1c338e80197d-Paper.pdf .

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In International Conference on Learning Representations, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx .

Appendix A. Proof of Theorem 1

Theorem The gradient flow path with learning rate $\eta(t)$ for EWN and SWN are given as follows:

$$\text{EWN: } \frac{dw_u(t)}{dt} = -\eta(t)\|w_u(t)\|^2 \nabla_{w_u} L$$

$$\text{SWN: } \frac{dw_u(t)}{dt} = -\eta(t)\left[\|w_u(t)\|^2 \nabla_{w_u} L + \frac{1 - \|w_u(t)\|^2}{\|w_u(t)\|^2}(w_u(t)^\top \nabla_{w_u} L) w_u(t)\right]$$

The proof for the two parts will be provided in different subsections, where the corresponding part will be restated for ease of the reader.

A.1. Exponential Weight Normalization

Theorem The gradient flow path with learning rate $\eta(t)$ for EWN is given by:

$$\frac{dw_u(t)}{dt} = -\eta(t)\|w_u(t)\|^2 \nabla_{w_u} L$$

Proof In case of EWN, weights are reparameterized as $w_u = e^{g_u}\frac{v_u}{\|v_u\|}$. Then

$$\nabla_{g_u} L = e^{g_u}\frac{v_u^\top \nabla_{w_u} L}{\|v_u\|}$$

$$\nabla_{v_u} L = \frac{e^{g_u}}{\|v_u\|}\left[I - \frac{v_u v_u^\top}{\|v_u\|^2}\right]\nabla_{w_u} L$$

Now, in case of gradient flow with learning rate $\eta(t)$, we can say

$$\frac{dg_u(t)}{dt} = -\eta(t)\nabla_{g_u} L = -\eta(t) e^{g_u(t)}\frac{v_u(t)^\top \nabla_{w_u} L}{\|v_u(t)\|}$$

$$\frac{dv_u(t)}{dt} = -\eta(t)\nabla_{v_u} L = -\eta(t)\frac{e^{g_u(t)}}{\|v_u(t)\|}\left[I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\right]\nabla_{w_u} L$$

Now, using these equations, we can say

$$\frac{d\|v_u(t)\|^2}{dt} = 2v_u(t)^\top \frac{dv_u(t)}{dt} = 0$$

Thus, $\|v_u(t)\|$ does not change with time. As we assumed $\|v_u(0)\|$ to be 1, therefore for any t, $\|v_u(t)\| = 1$. Using this simplification, we can write

$$\begin{aligned}\frac{dw_u(t)}{dt} &= \frac{d(e^{g_u(t)}v_u(t))}{dt} \\ &= e^{g_u(t)}\left(-\eta(t)e^{g_u(t)}(I - v_u(t)v_u(t)^\top)\nabla_{w_u} L\right) - \eta(t)e^{2g_u(t)}(v_u(t)^\top \nabla_{w_u} L)v_u(t) \\ &= -\eta(t)e^{2g_u(t)}\nabla_{w_u} L\end{aligned}$$

Thus, the gradient flow path with exponential weight normalization can be replicated by an adaptive learning rate given by $(t)\|w_u(t)\|^2$. ■

A.2. Standard Weight Normalization

Theorem The gradient flow path with learning rate (t) for SWN is given by:

$$\frac{dw_u(t)}{dt} = \quad (t) \quad \|w_u(t)\|^2 r_{w_u} L + \frac{1 \quad \|w_u(t)\|^2}{\|w_u(t)\|^2} \quad (w_u(t)^> r_{w_u} L) w_u(t)$$

Proof In case of SWN, weights are reparameterized as $w_u = \quad_u \frac{v_u}{\|v_u\|}$. Then

$$r \quad_u L = \frac{v_u^> r_{w_u} L}{\|v_u\|}$$

$$r_{v_u} L = \frac{\quad_u}{\|v_u\|} \quad I \quad \frac{v_u v_u^>}{\|v_u\|^2} \quad r_{w_u} L$$

Now, in case of gradient flow with learning rate (t), we can say

$$\frac{d \quad_u(t)}{dt} = \quad (t) r \quad_u L = \quad (t) \frac{v_u(t)^> r_{w_u} L}{\|v_u(t)\|}$$

$$\frac{dv_u(t)}{dt} = \quad (t) r_{v_u} L = \quad (t) \frac{\quad_u(t)}{\|v_u(t)\|} \quad I \quad \frac{v_u(t) v_u(t)^>}{\|v_u(t)\|^2} \quad r_{w_u} L$$

Now, similar to EWN, $\|v_u(t)\|$ does not change with time. Using the fact that $\|w_u(t)\| = 1$ for all t, we can say

$$\begin{aligned}
\frac{dw_u(t)}{dt} &= \frac{d(\quad_u(t) v_u(t))}{dt} \\
&= \quad_u(t)(\quad (t) \quad_u(t)(I \quad v_u(t) v_u(t)^>) r_{w_u} L) \quad (t)(v_u(t)^> r_{w_u} L) v_u(t) \\
&= \quad (t)(\quad_u(t)^2 r_{w_u} L + (1 \quad \quad_u(t)^2)(v_u(t)^> r_{w_u} L) v_u(t)) \\
&= \quad (t) \quad \quad_u(t)^2 r_{w_u} L + \frac{1 \quad \quad_u(t)^2}{\quad_u(t)^2} \quad (w_u(t)^> r_{w_u} L) w_u(t)
\end{aligned}$$

Replacing $\quad_u(t)$ by $\|w_u(t)\|$ gives the required expression. ■

Appendix B. Proof of Proposition 2

Proposition Under assumption (A1) for gradient flow, for both SWN and EWN, $\lim_{t!1} L(w(t)) = 0$.

The proof for SWN, as it is homogeneous in its parameters, was provided by Lyu and Li (2020). We provide the proof for EWN.

First of all, for exponential loss

$$\frac{dL(t)}{dw} = \sum_i e^{-y_i(w(t);x_i)} y_i r_w(w(t);x_i) \tag{11}$$

Now, using Theorem 1,

$$\frac{dL(t)}{dt} = \frac{dL(t)}{dw}^\top \frac{dw(t)}{dt} = \sum_u \|w_u(t)\|^2 \frac{dL(t)}{dw_u}^2$$

Let k be the total number of neurons in the network. Then using the elementary inequality $(\sum_{i=1}^n a_i)^2 \le n \sum_{i=1}^n a_i^2$, we get

$$\frac{dL(t)}{dt} \ge \frac{1}{k} \left(\sum_u \|w_u(t)\| \frac{dL(t)}{dw_u} \right)^2$$

Again using the fact that $w(t)^\top \frac{dL(t)}{dw} = \sum_u \|w_u(t)\| \frac{dL(t)}{dw_u}$, we get

$$\frac{dL(t)}{dt} \ge \frac{1}{k} \left(w(t)^\top \frac{dL(t)}{dw} \right)^2 \tag{12}$$

Taking the dot product with w on both sides of Equation (11) and using $w^\top r_w(w;x_i) = L(w;x_i)$ (Euler's homogeneity theorem), we get

$$w(t)^\top \frac{dL(t)}{dw} = -\sum_i e^{-y_i(w(t);x_i)} y_i(w(t);x_i)$$

Now, using the fact, that at time t_0, $L(t_0) < 1$, which means $\min_i y_i(w(t_0);x_i) = \gamma > 0$. Also, as we know, for gradient flow, the loss cannot go up, therefore, for any time $t > t_0$, $\min_i y_i(w(t);x_i) > \gamma > 0$. Using this, we can say, for any $t > t_0$,

$$w(t)^\top \frac{dL(t)}{dw} \le -\gamma L \quad L(t))$$

Substituting this in Equation (12), we get

$$\frac{dL(t)}{dt} \ge \frac{\gamma^2}{k} L(t)^2$$

Integrating this equation from t_0 to t, we get

$$\frac{1}{L(t)} \ge \frac{1}{L(t_0)} + \frac{\gamma^2}{k}(t - t_0) \tag{13}$$

Clearly as t tends to ∞, RHS tends to ∞ and thus L tends to 0.

Appendix C. Proof of Theorem 3

Theorem Consider a node u in the network with $\|g_u\| > 0$ and $\lim_{t\to\infty} \|w_u(t)\| = 1$. Under assumptions (A1), (A2) for gradient flow and (B1)-(B3) for gradient descent, for both SWN and EWN

(i) $\lim_{t\to\infty} \frac{w_u(t)}{\|w_u(t)\|} := w_{\infty,u}$ exists. (ii) $w_{\infty,u} = g_u$ for some > 0.

The proof for different cases will be split into different subsections and corresponding theorem will be stated there for ease of the reader. The proof will depend on the Stolz Cesaro theorem(stated in Appendix K),Integral Form of Stolz-Cesaro Theorem(stated and proved in Appendix K) and following lemmas that have been proved in Appendix J.

Lemma 9 Consider sequence a satisfying the following properties

1. $a_k > 0$
2. $\sum_{k=0}^{\infty} a_k = 1$
3. $\lim_{k\to\infty} a_k = 0$

Then $\sum_{k=0}^{\infty} \frac{a_k}{\sqrt{\sum_{j=0}^{k} a_j^2}} = 1$

Lemma 10 Consider two sequences a and b satisfying the following properties

1. $a_k > 0$; $\sum_{k=0}^{\infty} a_k = 1$ and $\lim_{k\to\infty} a_k = 0$
2. $b_0 > 0$, b is increasing and $b_{k+1}^2 \leq b_k^2 + \left(\frac{a_k}{b_k}\right)^2$

Then $\sum_{k=0}^{\infty} \frac{a_k}{b_k} = \infty$.

C.1. Exponential Weight Normalization

In this section, we will use $e^{\alpha(t)}$ and $\|w_u(t)\|$ interchangeably.

C.1.1. GRADIENT FLOW

Theorem Consider a node u in the network with $\|g_u\| > 0$ and $\lim_{t\to\infty} \|w_u(t)\| = 1$. Under assumptions (A1), (A2) for gradient flow, for EWN

(i) $\lim_{t\to\infty} \frac{w_u(t)}{\|w_u(t)\|} := w_{\infty,u}$ exists. (ii) $w_{\infty,u} = g_u$ for some > 0.

Update Equations:

$$\frac{d\alpha_u(t)}{dt} = \eta(t) e^{\alpha_u(t)} (v_u(t)^\top \nabla_{w_u} L(w(t))) \tag{14}$$

$$\frac{dv_u(t)}{dt} = \eta(t) e^{\alpha_u(t)} (I - v_u(t)v_u(t)^\top) \nabla_{w_u} L(w(t)) \tag{15}$$

Proof As $\|g_u\| > 0$, therefore $\nabla_{w_u} L(w(t))$ converges in direction. Therefore, for every satisfying $0 < < 2$, there exists a time $t_1(\ )$, such that for $t > t_1$, $\left\langle \frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|}, \frac{g_u}{\|g_u\|} \right\rangle > \cos(\ )$.

Now, Let's assume that $v_u(t)$ does not converge in the direction of g_u. Then, there must exist a θ satisfying $0 < \theta < \frac{\pi}{2}$, such that for this θ, there exists a time $t_2 > t_1(\theta)$ satisfying $v_u(t_2)^\top \frac{g_u}{\|g_u\|} = \cos(\theta)$, where $\theta > \theta$.

Now, we are going to show that for any θ satisfying $\theta < \theta < \frac{\pi}{2}$, there exists a time $t_3 > t_2$ such that $v_u(t_3)^\top \frac{g_u}{\|g_u\|} > \cos(\theta)$. Let's say for a given θ, no such t_3 exists. Then, taking dot product with $\frac{g_u}{\|g_u\|}$ on both sides of Equation (15), we can say

$$\frac{g_u}{\|g_u\|}^\top \frac{dv_u(t)}{dt} = \eta(t) e^{-\gamma_u(t)} \|\nabla_{w_u} L(w(t))\| \frac{g_u}{\|g_u\|}^\top (I - v_u(t) v_u(t)^\top) \frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|}$$

Now, as $\frac{g_u}{\|g_u\|}^\top \frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|} \geq \cos(\theta)$ and $\frac{g_u}{\|g_u\|}^\top v_u \leq \cos(\theta)$, we can say

$$\frac{g_u}{\|g_u\|}^\top \frac{dv_u(t)}{dt} \geq \eta(t) e^{-\gamma_u(t)} \|\nabla_{w_u} L(w(t))\| (\cos(\theta) - \cos(\theta)) \tag{16}$$

Now, using the fact that $\gamma_u \to 1$ and using Equation (14), we can say

$$\int_{t=t_2}^{1} \eta(t) e^{-\gamma_u(t)} \|\nabla_{w_u} L(w(t))\| dt = 1$$

Using this fact and integrating the Equation (16) on both the sides from t_2 to 1, we get a contradiction as vectors on LHS have a finite norm while RHS tends to 1. Thus, for every θ between θ and $\frac{\pi}{2}$, there must exist t_3, such that $v_u(t_3)^\top \frac{g_u}{\|g_u\|} > \cos(\theta)$.

Now, we are going to show for all $t > t_3$, $v_u(t)^\top \frac{g_u}{\|g_u\|} > \cos(\theta)$. Now, consider any θ such that $\theta < \theta < \frac{\pi}{2}$. Using similar argument as in Equation (16), we can say, if for any $t_4 > t_3$, $v_u(t_4)^\top \frac{g_u}{\|g_u\|} < \cos(\theta)$, then

$$\frac{g_u}{\|g_u\|}^\top \frac{dv_u(t_4)}{dt} \geq \eta(t_4) e^{-\gamma_u(t_4)} \|\nabla_{w_u} L(w(t_4))\| (\cos(\theta) - \cos(\theta)) \tag{17}$$

This means that the dot product between $\frac{g_u}{\|g_u\|}$ and $v_u(t)$ goes up, whenever $\frac{g_u}{\|g_u\|}^\top v_u(t) < \cos(\theta)$. Therefore, its not possible that $v_u(t)^\top \frac{g_u}{\|g_u\|} \leq \cos(\theta)$ for any $t > t_3$. As θ can be arbitrarily chosen between θ and $\frac{\pi}{2}$, $w_u(t)$ converges in the direction of g_u. ■

C.1.2. GRADIENT DESCENT

Theorem Consider a node u in the network with $\|g_u\| > 0$ and $\lim_{t\to 1} \|w_u(t)\| = 1$. Under assumptions (B1)-(B3) for gradient descent, for EWN

(i) $\lim_{t\to 1} \frac{w_u(t)}{\|w_u(t)\|} := w_u$ exists. (ii) $w_u = g_u$ for some > 0.

Update Equations:

$$\gamma_u(t+1) = \gamma_u(t) - \eta(t) e^{-\gamma_u(t)} \frac{v_u(t)^\top \nabla_{w_u} L(w(t))}{\|v_u(t)\|} \tag{18}$$

$$v_u(t+1) = v_u(t) - \eta(t)\frac{e^{\alpha_u(t)}}{\|v_u(t)\|}\left(I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\right)\nabla_{w_u} L(w(t)) \tag{19}$$

Proof

As $\|\tilde{g}_u\| > 0$, therefore $-\nabla_{w_u} L(w(t))$ converges in direction. Therefore, for every ϵ satisfying $0 < \epsilon < \frac{\pi}{2}$, there exists a time $t_1(\epsilon)$, such that for $t > t_1(\epsilon)$, $\left(\frac{-\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|}\right)^\top \frac{\tilde{g}_u}{\|\tilde{g}_u\|} \geq \cos(\epsilon)$. Now, Let's assume that $v_u(t)$ does not converge in the direction of $\tilde{g}_u$. Then, there must exist a δ satisfying $0 < \delta < \frac{\pi}{2}$, such that for this δ, there exists a time $t_2 > t_1(\epsilon)$ satisfying $v_u(t_2)^\top \frac{\tilde{g}_u}{\|\tilde{g}_u\|} = \cos(\delta)$, where $\delta > \epsilon$.

Now, we are going to show that for any β satisfying $\epsilon < \beta < \delta$, there exists a time $t_3 > t_2$ such that $\left(\frac{v_u(t_3)}{\|v_u(t_3)\|}\right)^\top \frac{\tilde{g}_u}{\|\tilde{g}_u\|} > \cos(\beta)$. Let's say for a given β, no such t_3 exists. Then, taking dot product with $\frac{\tilde{g}_u}{\|\tilde{g}_u\|}$ on both sides of Equation (19), we can say

$$\frac{v_u(t+1)^\top \tilde{g}_u}{\|\tilde{g}_u\|} = \frac{v_u(t)^\top \tilde{g}_u}{\|\tilde{g}_u\|} + \eta(t)\frac{e^{\alpha_u(t)}}{\|v_u(t)\|}\|\nabla_{w_u} L(w(t))\|\left(\frac{\tilde{g}_u}{\|\tilde{g}_u\|}\right)^\top\left(I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\right)\left(\frac{-\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|}\right)$$

Now, as $\left(\frac{\tilde{g}_u}{\|\tilde{g}_u\|}\right)^\top\left(\frac{-\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|}\right) \geq \cos(\epsilon)$ and $\left(\frac{\tilde{g}_u}{\|\tilde{g}_u\|}\right)^\top\left(\frac{v_u(t)}{\|v_u(t)\|}\right) \leq \cos(\beta)$, we can say

$$\frac{v_u(t+1)^\top \tilde{g}_u}{\|\tilde{g}_u\|} \geq \frac{v_u(t)^\top \tilde{g}_u}{\|\tilde{g}_u\|} + (\cos(\epsilon) - \cos(\beta))\left(\eta(t)\frac{e^{\alpha_u(t)}}{\|v_u(t)\|}\|\nabla_{w_u} L(w(t))\|\right) \tag{20}$$

However, in this case, $\|v_u(t)\|$ doesn't stay constant and thus increase in dot product doesn't directly correspond to an increase in angle. Now, using Equation (19), we can say

$$\|v_u(t+1)\|^2 \leq \|v_u(t)\|^2 + \left(\eta(t)\frac{e^{\alpha_u(t)}}{\|v_u(t)\|}\|\nabla_{w_u} L(w(t))\|\right)^2 \tag{21}$$

Using the above two equations, we can say, for time $t \geq t_2$,

$$\frac{v_u(t+1)^\top \tilde{g}_u}{\|v_u(t+1)\|\|\tilde{g}_u\|} \geq \frac{\frac{v_u(t)^\top \tilde{g}_u}{\|\tilde{g}_u\|} + (\cos(\epsilon) - \cos(\beta))\,\eta(t)\frac{e^{\alpha_u(t)}}{\|v_u(t)\|}\|\nabla_{w_u} L(w(t))\|}{\sqrt{\|v_u(t)\|^2 + \left(\eta(t)\frac{e^{\alpha_u(t)}}{\|v_u(t)\|}\|\nabla_{w_u} L(w(t))\|\right)^2}}$$

Unrolling the equation above, we get

$$\frac{v_u(t+1)^\top \tilde{g}_u}{\|v_u(t+1)\|\|\tilde{g}_u\|} \geq \frac{\frac{v_u(t_2)^\top \tilde{g}_u}{\|\tilde{g}_u\|} + \sum_{k=t_2}^{k=t}(\cos(\epsilon) - \cos(\beta))\,\eta(k)\frac{e^{\alpha_u(k)}}{\|v_u(k)\|}\|\nabla_{w_u} L(w(k))\|}{\sqrt{\|v_u(t_2)\|^2 + \sum_{k=t_2}^{k=t}\left(\eta(k)\frac{e^{\alpha_u(k)}}{\|v_u(k)\|}\|\nabla_{w_u} L(w(k))\|\right)^2}} \tag{22}$$

Now, as $\alpha_u(t) \to \infty$, therefore, using Equation (18), we can say

$$\sum_{k=t_2}^{k=\infty} \eta(k)e^{\alpha_u(k)}\|\nabla_{w_u} L(w(k))\| = \infty$$

Now, using this identity, along with the Assumption (A5), Equation (21) and Lemma 10, we can say

$$\sum_{k=t_2}^{1} (k)\frac{e^{\ u(k)}}{kv_u(k)k} kr_{\ w_u} L(w(k))k = 1$$

Using this along with Equation (22) and Lemma 9, we can say

$$\lim_{t!1} \frac{v_u(t+1)^{>} \hat{g}_u}{kv_u(t+1)kk\hat{g}_u k} \quad 1$$

However, this is not possible as the vectors on LHS have bounded norm. This contradicts. Thus there must exist a t_3 such that $\frac{v_u(t_3)}{kv_u(t_3)k}^{>} \frac{\hat{g}_u}{k\hat{g}_u k} > \cos(\ )$.

Now, we are going to show that there exists a $t_4 > t_3$, such that for all $t > t_4$, $\frac{v_u(t)}{kv_u(t)k}^{>} \frac{\hat{g}_u}{k\hat{g}_u k} > \cos(\ )$. Consider a such that $< <$. Now, if at any time t, $\frac{v_u(t)}{kv_u(t)k}^{>} \frac{\hat{g}_u}{k\hat{g}_u k} < \cos(\ )$, then, similar to Equation (20), we can say

$$\frac{v_u(t+1)^{>} \hat{g}_u}{k\hat{g}_u k} \quad \frac{v_u(t)^{>} \hat{g}_u}{k\hat{g}_u k} + (\cos(\ ) \quad \cos(\ )) \quad (t)\frac{e^{\ u(t)}}{kv_u(t)k} kr_{\ w_u} L(w(t))k \Bigg)!$$

Using the upper bound on $kv_u(t+1)k$ from Equation (21), we can say

$$\frac{v_u(t+1)^{>} \hat{g}_u}{kv_u(t+1)kk\hat{g}_u k} \quad \frac{\frac{v_u(t)^{>} \hat{g}_u}{k\hat{g}_u k} + (\cos(\ ) \quad \cos(\ )) \quad (t)\frac{e^{\ u(t)}}{kv_u(t)k} kr_{\ w_u} L(w(t))k}{\sqrt{kv_u(t)k^2 + \quad (t)\frac{e^{\ u(t)}}{kv_u(t)k} kr_{\ w_u} L(w(t))k^{\ 2}}} \tag{23}$$

Let $(t)\frac{e^{\ u(t)}}{kv_u(t)k} kr_{\ w_u} L(w(t))k$ be denoted by (t). Then, the above equation can be rewritten as

$$\frac{v_u(t+1)^{>} \hat{g}_u}{kv_u(t+1)kk\hat{g}_u k} \quad \frac{v_u(t)^{>} \hat{g}_u}{kv_u(t)kk\hat{g}_u k} p\frac{kv_u(t)k}{\sqrt{kv_u(t)k^2 + \quad (t)^2}} + (\cos(\ ) \quad \cos(\ )) p\frac{(t)}{\sqrt{kv_u(t)k^2 + \quad (t)^2}}$$

Now, we are going to show that for a small enough (t), RHS is greater than $\frac{v_u(t)^{>} \hat{g}_u}{kv_u(t)kk\hat{g}_u k}$.

$$\frac{v_u(t)^{>} \hat{g}_u}{kv_u(t)kk\hat{g}_u k} p\frac{kv_u(t)k}{\sqrt{kv_u(t)k^2 + \quad (t)^2}} + (\cos(\ ) \quad \cos(\ )) p\frac{(t)}{\sqrt{kv_u(t)k^2 + \quad (t)^2}} > \frac{v_u(t)^{>} \hat{g}_u}{kv_u(t)kk\hat{g}_u k}\Bigg)!$$

$$=) \quad (\cos(\ ) \quad \cos(\ )) p\frac{(t)}{\sqrt{kv_u(t)k^2 + \quad (t)^2}} > \frac{v_u(t)^{>} \hat{g}_u}{kv_u(t)kk\hat{g}_u k} \quad 1 \quad p\frac{kv_u(t)k}{\sqrt{kv_u(t)k^2 + \quad (t)^2}}\Bigg)!$$

$$=) \quad (\cos(\ ) \quad \cos(\ )) > \frac{v_u(t)^{>} \hat{g}_u}{kv_u(t)kk\hat{g}_u k} \quad \frac{p\sqrt{kv_u(t)k^2 + \quad (t)^2} \quad k\ v_u(t)k}{(t)}$$

Clearly as $(t) ! \quad 0$, the RHS tends to 0, therefore the equation is satis ed. Thus for a small enough (t), RHS of Equation (23) is greater than $\frac{v_u(t)^{>} \hat{g}_u}{kv_u(t)kk\hat{g}_u k}$. As $kv_u(t)k$ keeps on increasing and

by Assumption (B3), $\lim_{t\to\infty} \eta(t)\|w_u(t)\|\|\nabla_{w_u} L(w(t))\| = 0$, we can say there exists a time t_5, such that for any $t > t_5$, $\frac{v_u(t)^\top g_u}{\|v_u(t)\|\|g_u\|}$ goes up whenever $\frac{v_u(t)}{\|v_u(t)\|}^\top \frac{g_u}{\|g_u\|} < \cos(\theta)$.

Also, by using Equation (19) and Assumption (B3), we can say, that there exists a time t_6, such that for $t > t_6$, $\frac{v_u(t)}{\|v_u(t)\|}^\top \frac{g_u}{\|g_u\|} > \cos(\theta) \implies \frac{v_u(t+1)}{\|v_u(t+1)\|}^\top \frac{g_u}{\|g_u\|} > \cos(\theta)$, as the RHS of Equation (19) goes to 0 norm in limit. Now, define $t_4 > \max(t_5, t_6)$ such that $\frac{v_u(t_4)}{\|v_u(t_4)\|}^\top \frac{g_u}{\|g_u\|} > \cos(\theta)$ (must exist from previous arguments). Then, as the dot product always goes up when between $\cos(\theta)$ and $\cos(\theta)$, and can't go in a single step from being greater than $\cos(\theta)$ to less than $\cos(\theta)$, therefore, for every $t > t_4$, $\frac{v_u(t)}{\|v_u(t)\|}^\top \frac{g_u}{\|g_u\|} > \cos(\theta)$.

Now as the above argument holds for any θ between 0 and $\frac{\pi}{2}$, and for any $\epsilon > 0$, we can say that $w_u(t)$ converges in direction of g_u. ■

C.2. Standard Weight Normalization

C.2.1. GRADIENT FLOW

Theorem Consider a node u in the network with $\|g_u\| > 0$ and $\lim_{t\to\infty} \|w_u(t)\| = 1$. Under assumptions (A1), (A2) for gradient flow, for SWN

(i) $\lim_{t\to\infty} \frac{w_u(t)}{\|w_u(t)\|} := w_u$ exists. (ii) $w_u = g_u$ for some $\alpha > 0$.

Update Equations:

$$\frac{d\eta(t)}{dt} = \eta(t)\frac{v_u(t)^\top \nabla_{w_u} L(w(t))}{\|v_u(t)\|} \tag{24}$$

$$\frac{dv_u(t)}{dt} = \eta(t)\frac{\eta(t)}{\|v_u(t)\|}\left(I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\right)\nabla_{w_u} L(w(t)) \tag{25}$$

Proof The proof will be given for $\eta(t) \to \infty$. The one for $\eta(t) \to 0$ can be handled similarly.

As $\|g_u\| > 0$, therefore $\nabla_{w_u} L(w(t))$ converges in direction. Therefore, for every θ satisfying $0 < \theta < \frac{\pi}{2}$, there exists a time $t_1(\theta)$, such that for $t > t_1(\theta)$, $\frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|}^\top \frac{g_u}{\|g_u\|} \geq \cos(\theta)$. Now, Let's assume that $v_u(t)$ does not converge in the direction of g_u. Then, there must exist a θ satisfying $0 < \theta < \frac{\pi}{2}$, such that for this θ, there exists a time $t_2 > t_1(\theta)$ satisfying $v_u(t_2)^\top \frac{g_u}{\|g_u\|} = \cos(\theta)$, where $\epsilon > 0$.

Now, we are going to show that for any θ satisfying $0 < \theta < \frac{\pi}{2}$, there exists a time $t_3 > t_2$ such that $v_u(t_3)^\top \frac{g_u}{\|g_u\|} > \cos(\theta)$. Let's say for a given θ, no such t_3 exists. Then, taking dot product with $\frac{g_u}{\|g_u\|}$ on both sides of Equation (25), we can say

$$\frac{g_u}{\|g_u\|}^\top \frac{dv_u(t)}{dt} = \eta(t)\,\eta(t)\|\nabla_{w_u} L(w(t))\|\,\frac{g_u}{\|g_u\|}^\top (I - v_u(t)v_u(t)^\top)\,\frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|}$$

Now, as $\frac{g_u}{\|g_u\|}^\top \frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|} \geq \cos(\theta)$ and $\frac{g_u}{\|g_u\|}^\top v_u \leq \cos(\theta)$, we can say

$$\frac{g_u}{\|g_u\|}^\top \frac{dv_u(t)}{dt} \geq \eta(t)\,\eta(t)\|\nabla_{w_u} L(w(t))\|(\cos(\theta) - \cos(\theta)) \tag{26}$$

Now, using the fact that $_u ! 1$ and using Equation (24), we can say

$$\int_{t=t_2}^{Z_{1}} (t) k r_{w_u} L(w(t)) k dt = 1$$

Using this fact and integrating the Equation (26) on both the sides from t_2 to 1, we get a contradiction as vectors on LHS have a nite norm while RHS tends to 1. Thus, for every between and , there must exist t_3, such that $v_u(t_3)^> \frac{g_u}{kg_uk} > \cos( )$.

Now, we are going to show for all t_3, $v_u(t)^> \frac{g_u}{kg_uk} > \cos( )$. Now, consider any such that < < . Using similar argument as in Equation (26), we can say, if for any $t_4 > t_3$, $v_u(t_4)^> \frac{g_u}{kg_uk} < \cos( )$, then

$$\frac{g_u}{kg_uk}^> \frac{dv_u(t_4)}{dt} \quad (t_4) \ _u(t_4) k r_{w_u} L(w(t_4)) k (\cos( ) \quad \cos( )) \tag{27}$$

This means that the dot product between $\frac{g_u}{kg_uk}$ and $v_u(t)$ goes up, whenever $\frac{g_u}{kg_uk}^> v_u(t) < \cos( )$. Therefore, its not possible that $v_u(t)^> \frac{g_u}{kg_uk} \cos( )$ for any $t > t_3$. As can be arbitrarily chosen between and , $w_u(t)$ converges in the direction of g_u ■

C.2.2. GRADIENT DESCENT

Theorem Consider a node u in the network with $kg_uk > 0$ and $\lim_{t!1} kw_u(t)k = 1$. Under assumptions (B1)-(B3) for gradient descent, for SWN

(i) $\lim_{t!1} \frac{w_u(t)}{kw_u(t)k} := w_u$ exists. (ii) $w_u = g_u$ for some > 0.

Update Equations:

$$_u(t+1) = \ _u(t) \quad (t) \frac{v_u(t)^> r_{w_u} L(w(t))}{kv_u(t)k} \tag{28}$$

$$v_u(t+1) = v_u(t) \quad (t) \frac{_u(t)}{kv_u(t)k} \ I \quad \frac{v_u(t)v_u(t)^>}{kv_u(t)k^2} \quad r_{w_u} L(w(t)) \tag{29}$$

Proof The proof will be given for $_u ! 1$. The one for $_u ! 1$ can be handled similarly.

As $kg_uk > 0$, therefore $r_{w_u} L(w(t))$ converges in direction. Therefore, for every satisfying $0 < < 2$, there exists a time $t_1( )$, such that for $t > t_1( )$, $\frac{r_{w_u} L(w(t))}{kr_{w_u} L(w(t))k}^> \frac{g_u}{kg_uk}$ cos( ). Now, Let's assume that $w_u(t)$ does not converge in the direction of g_u. Then, there must exist a satisfying $0 < < 2$, such that for this , there exists a time $t_2 > t_1( )$ satisfying $v_u(t_2)^> \frac{g_u}{kg_uk} = \cos( )$, where > .

Now, we are going to show that for any satisfying < < , there exists a time $t_3 > t_2$ such that $\frac{v_u(t_3)}{kv_u(t_3)k}^> \frac{g_u}{kg_uk} > \cos( )$. Let's say for a given , no such t_3 exists. Then, taking

dot product with $\frac{g_u}{\|g_u\|}$ on both sides of Equation (29), we can say

$$\frac{v_u(t+1)^\top g_u}{\|g_u\|} = \frac{v_u(t)^\top g_u}{\|g_u\|} + (t)\frac{_u(t)}{\|v_u(t)\|}\nabla_{w_u}L(w(t))\| \frac{g_u}{\|g_u\|}^\top I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\frac{\nabla_{w_u}L(w(t))}{\|\nabla_{w_u}L(w(t))\|}$$

Now, as $\frac{g_u}{\|g_u\|}^\top \frac{\nabla_{w_u}L(w(t))}{\|\nabla_{w_u}L(w(t))\|} \cos(\ )$ and $\frac{g_u}{\|g_u\|}^\top \frac{v_u(t)}{\|v_u(t)\|} \cos(\ )$, we can say

$$\frac{v_u(t+1)^\top g_u}{\|g_u\|} \frac{v_u(t)^\top g_u}{\|g_u\|} + (\cos(\ ) \cos(\ ))(t)\frac{_u(t)}{\|v_u(t)\|}\nabla_{w_u}L(w(t))\| \tag{30}$$

However, in this case, $\|v_u(t)\|$ doesn't stay constant and thus increase in dot product doesn't directly correspond to an increase in angle. Now, using Equation (29), we can say

$$\|v_u(t+1)\|^2 \ \|v_u(t)\|^2 + (t)\frac{_u(t)}{\|v_u(t)\|}\nabla_{w_u}L(w(t))\|^2 \tag{31}$$

Using the above two equations, we can say, for time $t > t_2$,

$$\frac{v_u(t+1)^\top g_u}{\|v_u(t+1)\|\|g_u\|} \frac{\frac{v_u(t)^\top g_u}{\|g_u\|} + (t)\frac{_u(t)}{\|v_u(t)\|}\nabla_{w_u}L(w(t))\|}{\sqrt{\|v_u(t)\|^2 + (t)\frac{_u(t)}{\|v_u(t)\|}\nabla_{w_u}L(w(t))\|^2}}$$

Unrolling the equation above, we get

$$\frac{v_u(t+1)^\top g_u}{\|v_u(t+1)\|\|g_u\|} \frac{\frac{v_u(t_2)^\top g_u}{\|g_u\|} + \sum_{k=t_2}^{k=t}(k)\frac{_u(k)}{\|v_u(k)\|}\nabla_{w_u}L(w(k))\|}{\sqrt{\|v_u(t_2)\|^2 + \sum_{k=t_2}^{k=t}(k)\frac{_u(k)}{\|v_u(k)\|}\nabla_{w_u}L(w(k))\|^2}} \tag{32}$$

Now, as $_u(t) ! 1$, therefore, using Equation (29), we can say

$$\sum_{k=t_2}^{k=1}(k)\nabla_{w_u}L(w(k))\| = 1$$

Now, using this identity, along with the assumption (A5), Equation (31) and Lemma 10, we can say

$$\sum_{k=t_2}^{1}(k)\frac{_u(k)}{\|v_u(k)\|}\nabla_{w_u}L(w(k))\| = 1$$

Using this along with Equation (32) and Lemma 9, we can say

$$\lim_{t!1}\frac{v_u(t+1)^\top g_u}{\|v_u(t+1)\|\|g_u\|} \ 1$$

However, this is not possible as the vectors on LHS have bounded norm. This contradicts. Thus there must exist a t_3 such that $\frac{v_u(t_3)}{\|v_u(t_3)\|} \cdot \frac{g_u}{\|g_u\|} > \cos(\theta)$.

Now, we are going to show that there exists a $t_4 > t_3$, such that for all $t > t_4$, $\frac{v_u(t)}{\|v_u(t)\|} \cdot \frac{g_u}{\|g_u\|} > \cos(\theta)$. Consider a ϕ such that $0 < \phi < \theta$. Now, if at any time t, $\frac{v_u(t)}{\|v_u(t)\|} \cdot \frac{g_u}{\|g_u\|} < \cos(\theta)$, then, similar to Equation (30), we can say

$$\frac{v_u(t+1) \cdot g_u}{\|g_u\|} \geq \frac{v_u(t) \cdot g_u}{\|g_u\|} + (\cos(\theta - \phi) - \cos(\theta)) \frac{\eta(t)}{\|v_u(t)\|} \|\nabla_{w_u} L(w(t))\|$$

Using the upper bound on $\|v_u(t+1)\|$ from Equation (31), we can say

$$\frac{v_u(t+1) \cdot g_u}{\|v_u(t+1)\| \|g_u\|} \geq \frac{\frac{v_u(t) \cdot g_u}{\|g_u\|} + (\cos(\theta - \phi) - \cos(\theta)) \frac{\eta(t)}{\|v_u(t)\|} \|\nabla_{w_u} L(w(t))\|}{\sqrt{\|v_u(t)\|^2 + \left(\frac{\eta(t)}{\|v_u(t)\|} \|\nabla_{w_u} L(w(t))\|\right)^2}} \tag{33}$$

Let $\frac{\eta(t)}{\|v_u(t)\|} \|\nabla_{w_u} L(w(t))\|$ be denoted by $\alpha(t)$. Then, the above equation can be rewritten as

$$\frac{v_u(t+1) \cdot g_u}{\|v_u(t+1)\| \|g_u\|} \geq \frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|} \frac{\|v_u(t)\|}{\sqrt{\|v_u(t)\|^2 + \alpha(t)^2}} + (\cos(\theta - \phi) - \cos(\theta)) \frac{\alpha(t)}{\sqrt{\|v_u(t)\|^2 + \alpha(t)^2}}$$

Now, we are going to show that for a small enough $\alpha(t)$, RHS is greater than $\frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|}$.

$$\frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|} \frac{\|v_u(t)\|}{\sqrt{\|v_u(t)\|^2 + \alpha(t)^2}} + (\cos(\theta - \phi) - \cos(\theta)) \frac{\alpha(t)}{\sqrt{\|v_u(t)\|^2 + \alpha(t)^2}} > \frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|}$$

$$\Leftarrow (\cos(\theta - \phi) - \cos(\theta)) \frac{\alpha(t)}{\sqrt{\|v_u(t)\|^2 + \alpha(t)^2}} > \frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|} \left(1 - \frac{\|v_u(t)\|}{\sqrt{\|v_u(t)\|^2 + \alpha(t)^2}}\right)$$

$$\Leftarrow (\cos(\theta - \phi) - \cos(\theta)) > \frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|} \frac{\sqrt{\|v_u(t)\|^2 + \alpha(t)^2} - \|v_u(t)\|}{\alpha(t)}$$

Clearly as $\alpha(t) \to 0$, the RHS tends to 0, therefore the equation is satisfied. Thus for a small enough $\alpha(t)$, RHS of Equation (33) is greater than $\frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|}$. As $\|v_u(t)\|$ keeps on increasing and by Assumption (A5), $\lim_{t \to \infty} \alpha(t) = \frac{\eta(t)}{\|v_u(t)\|} \|\nabla_{w_u} L(w(t))\| = 0$, we can say there exists a time t_5, such that for any $t > t_5$, $\frac{v_u(t) \cdot g_u}{\|v_u(t)\| \|g_u\|}$ goes up whenever $\frac{v_u(t)}{\|v_u(t)\|} \cdot \frac{g_u}{\|g_u\|} < \cos(\theta)$.

Also, by using Equation (29) and Assumption (A5), we can say, that there exists a time t_6, such that for $t > t_6$, $\frac{v_u(t)}{\|v_u(t)\|} \cdot \frac{g_u}{\|g_u\|} > \cos(\theta) \Rightarrow \frac{v_u(t+1)}{\|v_u(t+1)\|} \cdot \frac{g_u}{\|g_u\|} > \cos(\theta)$, as the RHS of Equation (29) goes to 0 norm in limit. Now, define $t_4 > \max(t_5, t_6)$ such that $\frac{v_u(t_4)}{\|v_u(t_4)\|} \cdot \frac{g_u}{\|g_u\|} > \cos(\theta)$ (must exist from previous arguments). Then, as the dot product always goes up when between $\cos(\theta)$ and $\cos(\theta - \phi)$, and can't go in a single step from being greater than $\cos(\theta)$ to less than $\cos(\theta - \phi)$, therefore, for every $t > t_4$, $\frac{v_u(t)}{\|v_u(t)\|} \cdot \frac{g_u}{\|g_u\|} > \cos(\theta)$.

Now as the above argument holds for any θ between α and β, and for any $\epsilon > 0$, we can say that $w_u(t)$ converges in direction of g_u. ■

Appendix D. Proof of Theorem 4

Theorem Consider two nodes u and v in the network with $\|g_u\| \neq \|g_v\| > 0; \lim_{t\to\infty} \|w_u(t)\| = \infty$ and $\lim_{t\to\infty} \|w_v(t)\| = \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c. Under assumptions (A1), (A2) for gradient flow and (B1)-(B3) for gradient descent,

(i) for SWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

(ii) for EWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|}$ is either $0; \infty$ or $\frac{1}{c}$

Proof for different cases will be split into different subsections and the corresponding case will be restated there for ease of the reader.

D.1. Exponential Weight Normalization

D.1.1. GRADIENT FLOW

Theorem Consider two nodes u and v in the network with $\|g_u\| \neq \|g_v\| > 0; \lim_{t\to\infty} \|w_u(t)\| = \infty$ and $\lim_{t\to\infty} \|w_v(t)\| = \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c. Under assumptions (A1), (A2) for gradient flow, for EWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|}$ is either $0; \infty$ or $\frac{1}{c}$

Proof Using Theorem 3 and the fact that $\lim_{t\to\infty} \frac{\|r_{w_u} L(w(t))\|}{\|r_{w_v} L(w(t))\|} = c$, for any $0 < \epsilon < c$ and $0 < \theta < \frac{\pi}{2}$, there exists a time t_1, such that for $t > t_1$, the following hold

(i) $\frac{\|r_{w_u} L(w(t))\|}{\|r_{w_v} L(w(t))\|} \in [c - \epsilon, c + \epsilon]$ (ii) $\frac{w_u(t)}{\|w_u(t)\|} \cdot \frac{r_{w_u} L(w(t))}{\|r_{w_u} L(w(t))\|} > \cos(\theta)$

(iii) $\frac{w_v(t)}{\|w_v(t)\|} \cdot \frac{r_{w_v} L(w(t))}{\|r_{w_v} L(w(t))\|} > \cos(\theta)$.

Now, we will provide the proof of part (iii) of Proposition 5, i.e, for EWN, if at some time $t_2 > t_1$,

(a) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} > \frac{1}{(c-\epsilon)\cos(\theta)} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 1$

(b) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} < \frac{\cos(\theta)}{c+\epsilon} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 0$

(a) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} > \frac{1}{(c-\epsilon)\cos(\theta)} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 1$

Using Equation (14),

$$\frac{d\|w_u(t)\|}{dt} = \frac{de^{\eta_u(t)}}{dt} = \dot{\eta}_u(t)\|w_u(t)\|^2 (v_u(t) \cdot r_{w_u} L(w(t))) \tag{34}$$

Using the equation above, we can say for $t > t_1$,

$$\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} = \frac{\|w_v(t)\|\frac{d\|w_u(t)\|}{dt} - \|w_u(t)\|\frac{d\|w_v(t)\|}{dt}}{\|w_v(t)\|^2}$$

$$(t)\frac{\|w_u(t)\|}{\|w_v(t)\|}(\|w_u(t)\|\|r_{w_u}L(w(t))\|\cos(\ ) - \|w_v(t)\|\|r_{w_v}L(w(t))\|)$$

$$(t)\|w_u(t)\|\|r_{w_u}L(w(t))\| - \frac{\|w_u(t)\|}{\|w_v(t)\|}\cos(\ ) - \frac{1}{c} \tag{35}$$

In this case, using Equation (35), we can see $\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} > 0$ at t_2. Thus, $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ always remains greater than $\frac{1}{(c\ )\cos(\ )}$ and keeps on increasing. Let's denote $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|}$ by $\ $. Then, for $t > t_2$,

$$\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} \quad \cos(\ ) - \frac{1}{c} \quad (t)\|w_u(t)\|\|r_{w_u}L(w(t))\|$$

As $_u \to 1$, therefore using Equation (14), we can say $\int_{t_2}^{R_1} (t)\|w_u(t)\|\|r_{w_u}L(w(t))\|dt \to 1$. Thus, integrating both the sides of the equation above from t_2 to 1, we get

$$\int_{t_2}^{1} \frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt}dt \quad 1$$

Thus $\lim_{t\to 1} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 1$.

(b) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} < \frac{\cos(\ )}{c+} \Rightarrow \lim_{t\to 1} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 0$.

Using Equation (34), we can say for $t > t_1$,

$$\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} = \frac{\|w_v(t)\|\frac{d\|w_u(t)\|}{dt} - \|w_u(t)\|\frac{d\|w_v(t)\|}{dt}}{\|w_v(t)\|^2}$$

$$(t)\frac{\|w_u(t)\|}{\|w_v(t)\|}(\|w_u(t)\|\|r_{w_u}L(w(t))\| - \|w_v(t)\|\|r_{w_v}L(w(t))\|\cos(\ ))$$

$$= (\ (t)\|w_v(t)\|\|r_{w_v}L(w(t))\|)\frac{\|w_u(t)\|}{\|w_v(t)\|} - \frac{\|w_u\|\|r_{w_u}L(w(t))\|}{\|w_v\|\|r_{w_v}L(w(t))\|}\cos(\ )$$

$$(\ (t)\|w_v(t)\|\|r_{w_v}L(w(t))\|)\frac{\|w_u(t)\|}{\|w_v(t)\|} - \frac{\|w_u\|}{\|w_v\|}(c+\ ) \quad \cos(\ ) \tag{36}$$

In this case, using Equation (36), we can see $\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} < 0$ at t_2. Thus, $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ always remains smaller than $\frac{\cos(\ )}{c+}$ and keeps on decreasing. Now, lets say $\lim_{t\to 1} \frac{\|w_u(t)\|}{\|w_v(t)\|} > 0$. This means that $\frac{\|w_u(t)\|}{\|w_v(t)\|} > \ $, for some $\ > 0$. Also, let's denote $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|}$ by $\ $. Then we can say

$$\frac{d\frac{\|w_u(t)\|}{\|w_v(t)\|}}{dt} \quad (\cos(\ ) \quad (c+\ ))\ (t)\|w_v(t)\|\|r_{w_v}L(w(t))\|$$

As $_v ! 1$, therefore using Equation (14), we can say $\int_{t_2}^{1} (t)kw_v(t)kkr_{w_v}L(w(t))kdt = 1$. Thus, integrating both the sides of the equation above from t_2 to 1 , we get

$$\int_{t_2}^{1} \frac{d\frac{kw_u(t)k}{kw_v(t)k}}{dt}dt \quad 1$$

This is not possible as $\frac{kw_u(t)k}{kw_v(t)k}$ is lower bounded by 0. Thus $\lim_{t!1} \frac{kw_u(t)k}{kw_v(t)k} = 0$.

Now, as and tend to 0, the length of the interval of stability $[\frac{\cos( )}{c+}; \frac{1}{(c)\cos( )}]$ shrinks to zero, around the point $\frac{1}{c}$. Thus, either $\frac{kw_u(t)k}{kw_v(t)k}$ moves out of the interval of stability and converges to either 0 or 1 , or it always remains within the interval of stability and converges to $\frac{1}{c}$. ■

D.1.2. GRADIENT DESCENT

Theorem Consider two nodes u and v in the network with $kg_uk \quad k g_vk > 0; \lim_{t!1} kw_u(t)k = 1$ and $\lim_{t!1} kw_v(t)k = 1$. Let $\frac{kg_uk}{kg_vk}$ be denoted by . Under assumptions (B1)-(B3) for gradient descent, for EWN, $\lim_{t!1} \frac{kw_u(t)k}{kw_v(t)k}$ is either $0; 1$ or $\frac{1}{c}$

Proof Using Theorem 3 and the fact that $\lim_{t!1} \frac{kr_{w_u}L(w(t))k}{kr_{w_v}L(w(t))k} = c$, for any $0 < < c$ and $0 < < 2$, there exists a time t_1, such that for $t > t_1$, the following hold

(i) $\frac{kr_{w_u}L(w(t))k}{kr_{w_v}L(w(t))k} 2 [c \quad ; c + \quad]$ (ii) $\frac{w_u(t)}{kw_u(t)k} > \frac{r_{w_u}L(w(t))}{kr_{w_u}L(w(t))k} \quad \cos( )$

(iii) $\frac{w_v(t)}{kw_v(t)k} > \frac{r_{w_v}L(w(t))}{kr_{w_v}L(w(t))k} \quad \cos( )$.

Now, we will provide the proof of part (iii) of Proposition 5, i.e, for EWN, if at some time $t_2 > t_1$,

(a) $\frac{kw_u(t_2)k}{kw_v(t_2)k} > \frac{1}{(c)\cos( )} \implies \lim_{t!1} \frac{kw_u(t)k}{kw_v(t)k} = 1$

(b) $\frac{kw_u(t_2)k}{kw_v(t_2)k} < \frac{\cos( )}{c+} \implies \lim_{t!1} \frac{kw_u(t)k}{kw_v(t)k} = 0$

(a) $\frac{kw_u(t_2)k}{kw_v(t_2)k} > \frac{1}{(c)\cos( )} \implies \lim_{t!1} \frac{kw_u(t)k}{kw_v(t)k} = 1$

Using Equation (18) and part 1 of the Proposition, we can say

$$\frac{kw_u(t_2+1)k}{kw_v(t_2+1)k} \quad \frac{kw_u(t_2)k + \quad (t_2)\cos( )kw_u(t_2)k^2kr_{w_u}L(w(t_2))k}{kw_v(t_2)k + \quad (t_2)kw_v(t_2)k^2kr_{w_v}L(w(t_2))k}$$

$$= \frac{kw_u(t_2)k}{kw_v(t_2)k} \quad \frac{1 + \cos( ) \quad (t_2)kw_u(t_2)kkr_{w_u}L(w(t_2))k}{1 + \quad (t_2)kw_v(t_2)kkr_{w_v}L(w(t_2))k}$$

$$\frac{kw_u(t_2)k}{kw_v(t_2)k}$$

Thus, $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ keeps on increasing for $t > t_2$. It can either diverge to infinity or converge to a finite value. If it converges to a finite value, then by Stolz Cesaro theorem,

$$\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = \lim_{t\to\infty} \frac{\|w_u(t)\|^2 \|\nabla_{w_u} L(w(t))\|}{\|w_v(t)\|^2 \|\nabla_{w_v} L(w(t))\|}$$

However, this is not possible as $\frac{\|w_u(t)\|}{\|w_v(t)\|} > \frac{1}{c}$ for every $t > t_2$. Thus, $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ diverges to infinity.

(b) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} < \frac{\cos(\theta)}{c+\epsilon} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 0$

Using Equation (18) and part 1 of the Proposition, we can say

$$\frac{\|w_u(t_2+1)\|}{\|w_v(t_2+1)\|} = \frac{\|w_u(t_2) + \eta(t_2)\|w_u(t_2)\|^2 \nabla_{w_u} L(w(t_2))\|}{\|w_v(t_2) + \eta(t_2)\cos(\theta)\|w_v(t_2)\|^2 \nabla_{w_v} L(w(t_2))\|} = \frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} \frac{1 + \eta(t_2)\|w_u(t_2)\| \|\nabla_{w_u} L(w(t_2))\|}{1 + \eta(t_2)\cos(\theta)\|w_v(t_2)\| \|\nabla_{w_v} L(w(t_2))\|}$$

Thus, $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ keeps on decreasing for $t > t_2$. As it is always greater than zero, it must converge. Therefore, by Stolz Cesaro Theorem,

$$\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = \lim_{t\to\infty} \frac{\|w_u(t)\|^2 \|\nabla_{w_u} L(w(t))\|}{\|w_v(t)\|^2 \|\nabla_{w_v} L(w(t))\|}$$

For $\frac{\|w_u(t)\|}{\|w_v(t)\|} < \frac{1}{c}$, this can only be satisfied when $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 0$.

Now, as ϵ and θ tend to 0, the length of the interval of stability $[\frac{\cos(\theta)}{c+\epsilon}, \frac{1}{(c-\epsilon)\cos(\theta)}]$ shrinks to zero, around the point $\frac{1}{c}$. Thus, either $\frac{\|w_u(t)\|}{\|w_v(t)\|}$ moves out of the interval of stability and converges to either 0 or ∞, or it always remains within the interval of stability and converges to $\frac{1}{c}$. ■

D.2. Standard Weight Normalization

D.2.1. GRADIENT FLOW

Theorem Consider two nodes u and v in the network with $\|g_u\| \neq \|g_v\| > 0$; $\lim_{t\to\infty} \|w_u(t)\| = \infty$ and $\lim_{t\to\infty} \|w_v(t)\| = \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c. Under assumptions (A1), (A2) for gradient flow, for SWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

Proof From Theorem 3, we can say, for both u and v, weights and gradients converge in opposite directions.

Consider a time t_1, such that for any $t > t_1$,

- $\nabla_{w_u} L(w(t))$ and $w_u(t)$ almost make an angle θ with each other
- $\nabla_{w_v} L(w(t))$ and $w_v(t)$ almost make an angle θ with each other

Then, using Equation (24), we can say for any time $t \geq t_1$,

$$\|w_u(t)\| \geq \|w_u(t_2)\| + \cos(\theta) \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_u} L(w(k))\| dk$$

$$\|w_u(t)\| \leq \|w_u(t_2)\| + \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_u} L(w(k))\| dk$$

$$\|w_v(t)\| \leq \|w_v(t_2)\| + \cos(\theta) \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_v} L(w(k))\| dk$$

$$\|w_v(t)\| \geq \|w_v(t_2)\| + \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_v} L(w(k))\| dk$$

Using the above equations, we can say, for time $t \geq t_2$,

$$\frac{\|w_u(t)\|}{\|w_v(t)\|} \geq \frac{\|w_u(t_2)\| + \cos(\theta) \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_u} L(w(k))\| dk}{\|w_v(t_2)\| + \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_v} L(w(k))\| dk}$$

$$\frac{\|w_u(t)\|}{\|w_v(t)\|} \leq \frac{\|w_u(t_2)\| + \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_u} L(w(k))\| dk}{\|w_v(t_2)\| + \cos(\theta) \int_{k=t_2}^{t} \rho(k) \|\nabla_{w_v} L(w(k))\| dk}$$

We know that both integrals diverge as $\|w_u(t)\|$ and $\|w_v(t)\| \to \infty$, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|}$ and $\lim_{t\to\infty} \frac{\|\nabla_{w_u} L(w(t))\|}{\|\nabla_{w_v} L(w(t))\|}$ exist. Taking limit $t \to \infty$ on both the equations and using the Integral form of Stolz-Cesaro theorem, we get

$$\liminf_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} \geq \frac{\cos(\theta)\|g_u\|}{\|g_v\|}$$

$$\limsup_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} \leq \frac{\|g_u\|}{\cos(\theta)\|g_v\|}$$

As this holds for any $\epsilon > 0$, therefore

$$\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$$

■

D.2.2. GRADIENT DESCENT

Theorem Consider two nodes u and v in the network with $\|g_u\| \neq \|g_v\| > 0$; $\lim_{t\to\infty} \|w_u(t)\| = \infty$ and $\lim_{t\to\infty} \|w_v(t)\| = \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c. Under assumptions (B1) - (B3) for gradient descent, for SWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

Proof For ease of notation, we will denote the two nodes u and v by u and s. From Theorem 3, we can say, for both u and s, weights and gradients converge in opposite directions. Now from Equation (28), we can say

$$w_u(t) = w_u(0) \sum_{k=0}^{t-1} \rho(k) \frac{v_u(k)^\top \nabla_{w_u} L(w(k))}{\|v_u(k)\|}$$

$$\rho_s(t) = \rho_s(0) \sum_{k=0}^{k=t-1} \eta(k) \frac{v_s(k)^\top \nabla_{w_s} L(w(k))}{\|v_s(k)\|}$$

Now, $\rho_s(t)$ either diverges to $-\infty$ or $+\infty$. In both the cases, it is a strictly monotonic sequence for large enough t. Also $\lim_{t\to\infty} \frac{\rho_u(t+1)-\rho_u(t)}{\rho_s(t+1)-\rho_s(t)}$ exists. Therefore, using Stolz-Cesaro Theorem, we can say

$$\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_s(t)\|} = c$$

■

Appendix E. Proof of Proposition 5

Proposition Consider two nodes u and v in the network such that $\|\nabla_{w_v}\| - \|\nabla_{w_u}\| > 0$ and $\|w_u(t)\|, \|w_v(t)\| \to \infty$. Let $\frac{\|\nabla_{w_u}\|}{\|\nabla_{w_v}\|}$ be denoted by c. Consider any ϵ, δ such that $0 < \epsilon < c$ and $0 < \delta < \frac{\pi}{2}$. Then, the following holds:

(i) There exists a time t_1, such that for all $t > t_1$ both SWN and EWN trajectories have the following properties:

(a) $\frac{\|\nabla_{w_u} L(w(t))\|}{\|\nabla_{w_v} L(w(t))\|} \in [c-\epsilon, c+\epsilon]$ (b) $\frac{w_u(t)}{\|w_u(t)\|}^\top \frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|} \geq \cos(\delta)$

(c) $\frac{w_v(t)}{\|w_v(t)\|}^\top \frac{\nabla_{w_v} L(w(t))}{\|\nabla_{w_v} L(w(t))\|} \geq \cos(\delta)$.

(ii) for SWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

(iii) for EWN, if at some time $t_2 > t_1$,

(a) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} > \frac{1}{(c-\epsilon)\cos(\delta)} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 1$

(b) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} < \frac{\cos(\delta)}{c+\epsilon} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 0$

Proof (i) It follows from Theorem 3 and the fact that $\lim_{t\to\infty} \frac{\|\nabla_{w_u} L(w(t))\|}{\|\nabla_{w_v} L(w(t))\|} = c$.

(ii) Proof provided in Appendix D.2

(iii) Proof provided in Appendix D.1 ■

Appendix F. Proof of Proposition 6

Proposition Consider a linear model over $\mathbb{R}^d$ given by $f(x) = w^\top x$, where each w_i is further reparameterized as w_i. Consider a dataset consisting of a single data point $(0, 0)$, that is labelled as $+1$. According to the initialization of , de ne a relation R on $\{1, \ldots, d\}$, given by $i \quad j$ if $w_i(0)z_i = w_j(0)z_j$. Then, R is an equivalence relation on $\{1, \ldots, d\}$. Let these equivalent sets be denoted by $I_1, I_2, \ldots, I_k$. De ne a total order on these sets given by $I_a > I_b$ if $\exists i \in I_a, j \in I_b$ such that $w_i(0)z_i > w_j(0)z_j$. Let the maximum set according to this order be denoted by . Then, for gradient ow on exponential loss, the following holds

(i) For any $i \in I$, $\lim_{t\to\infty} w_i(t) = \infty$

(ii) For $i, j \in I$, $\frac{w_i(t)}{w_j(t)} = \frac{z_j}{z_i}$.

(iii) For any $i \in I$, $\lim_{t \to \infty} w_i(t) = \left(\frac{1}{w_i(0)} \sum_{j \in I} \frac{z_i}{w_j(0) z_j} \right)^{-1}$, where j is any element in I.

Proof In this case, the loss is given by

$$L(w) = e^{-w^\top z}$$

As each w_i is further exponentially reparameterized, therefore, using Theorem 1,

$$\frac{dw_i}{dt} = w_i^2 z_i e^{-w^\top z} \tag{37}$$

Using this equation, we can say, for any pair of indices i, j,

$$\frac{d\left(\frac{z_i}{w_j} - \frac{z_j}{w_i}\right)}{dt} = 0$$

Thus, for any pair of indices i, j

$$\frac{z_i}{w_j(t)} - \frac{z_j}{w_i(t)} = \frac{z_i}{w_j(0)} - \frac{z_j}{w_i(0)} \tag{38}$$

Thus, we can say for every equivalent set I_a, if $i, j \in I_a$, then as $\frac{z_j}{w_i(0)} - \frac{z_i}{w_j(0)} = 0$, therefore, at any time t,

$$\frac{w_i(t)}{w_j(t)} = \frac{z_j}{z_i}$$

Now, clearly the gradient flow stops only when $w^\top z \to \infty$. This means atleast one of the w_i must tend to ∞ (Notice from Equation (37) that w_i always goes up). For two equivalents sets I_a and I_b, such that $I_a > I_b$, we can say, for $i \in I_a$ and $j \in I_b$, the RHS of Equation (38) is positive. Thus, it is not possible for w_j to go to ∞, otherwise the quantity on the LHS will be negative when $t \to \infty$. Using this argument, we can say, only for $i \in I$, w_i tend to ∞. Then, for $i \in I$, considering $j \in I$ and using Equation (38),

$$\lim_{t \to \infty} w_i(t) = \left(\frac{1}{w_i(0)} \sum_{j \in I} \frac{z_i}{w_j(0) z_j} \right)^{-1}$$

■

Appendix G. Proof of Theorem 7

Theorem For Exponential Weight Normalization, under assumption (A1), the following hold for $t > t_0$ in case of gradient flow

(i) $\|w(t)\|$ grows with t as $\Omega((\log t)^{\frac{1}{L}})$ (ii) $L(t)$ goes down with t as $O\left(\frac{1}{t}\right)$

Proof Using Equation (13), we can say

$$\frac{1}{L(t)} \quad \frac{1}{L(t_0)} + \frac{L^2 \ ^2}{k}(t \quad t_0)$$

Thus $L(t)$ goes down with t as $O\left(\frac{1}{t}\right)$. Now, we know, $L(w) = \sum_i e^{\ y_i\ (\ w; x_i)}$. Clearly, as $(\ w; x_i)$ is smooth, therefore $L(w)$ attains a minima over the compact set $kwk = 1$. Using homogeneity of and the fact that L goes down as $O\left(\frac{1}{t}\right)$, we can say $kw(t)k$ grows with t as $((\log t)^{\frac{1}{L}})$. ■

Appendix H. Proof of Theorem 8

Theorem For Exponential Weight Normalization, under Assumptions (B1)-(B4), $_0$, $(t) = O\left(\log\frac{1}{L}^{\ c}\right)$ for $c < 1$ and $\lim_{t!1} \frac{kr(t+1)}{g(t+1)} \quad \frac{r(t)k}{g(t)} = 0$, the following hold

(i) $kw(t)k$ asymptotically grows with t as $(\log d(t))^{\frac{1}{L}}$

(ii) $L(w(t))$ asymptotically goes down with t as $\frac{1}{d(t)(\log d(t))^2}$.

First, we will establish rates for gradient ow to elucidate the proof technique and then go to the case of gradient descent.

H.1. Gradient Flow

Although the asymptotic convergence rates for smooth homogeneous neural nets have been established in Lyu and Li (2020), the proof technique becomes easier to understand for smooth homogeneous nets, without weight normalization. In this case, we will use an assumption (A3) that $\lim_{t!1} \frac{w(t)}{kw(t)k}$ exists.

H.1.1. UNNORMALIZED NETWORK

Theorem For Unnorm, under Assumptions (A1)-(A3) for gradient ow, $_0$ and $\lim_{t!1} \frac{k\frac{dr(t)}{dt}k}{g'(t)} = 0$, the following hold

1. $kw(t)k$ asymptotically grows at $(\log t)^{\frac{1}{L}}$

2. $L(w(t))$ asymptotically goes down at the rate of $\frac{1}{t(\log t)^{2\ \frac{2}{L}}}$.

Proof Consider $w = g(t)w + r(t)$, where $\lim_{t!1} \frac{kr(t)k}{g(t)} = 0$ and $r(t)^> w = 0$. Now, we make an additional assumption that $\lim_{t!1} \frac{k\frac{dr(t)}{d(t)}k}{g'(t)} = 0$. This basically avoids any oscillations in $r(t)$ for large t, where it can have a higher derivative, but the value may be bounded. Now, we know

$$\frac{dw(t)}{dt} = \sum_{i=1}^{n} e^{\ y_i\ (\ w(t); x_i)} y_i r_{\ w}\ (\ w(t); x_i) \tag{39}$$

Now, we know $\left\|\frac{dw(t)}{dt}\right\| \neq 0$ for any finite t, otherwise w won't change and L can't converge to 0. Thus, for all t, we can say

$$\frac{\left\|\frac{dw(t)}{dt}\right\|}{\left\|\sum_{i=1}^{m} e^{-y_i\, w(t)^\top x_i}\, y_i\, r_w(w; x_i)\right\|} = 1$$

Taking limit $t \to \infty$ on both the sides, we get

$$\lim_{t\to\infty} \frac{\left\|\frac{dw(t)}{dt}\right\|}{\left\|\sum_{i=1}^{m} e^{-y_i\, w(t)^\top x_i}\, y_i\, r_w(w; x_i)\right\|} = 1$$

Now, we know

$$\left\|\frac{dw(t)}{dt}\right\| = \left\|g'(t)\hat{w} + \frac{dr(t)}{dt}\right\|$$

$$\left\|\sum_{i=1}^{m} e^{-y_i\, w(t)^\top x_i}\, y_i\, r_w(w; x_i)\right\| = \left\|\sum_{i=1}^{m} e^{-y_i g(t) \hat{w}^\top \frac{r(t)}{g(t)} x_i}\, (y_i g(t)^{L-1})\, r_w\left(\hat{w} + \frac{r(t)}{g(t)}; x_i\right)\right\|$$

Thus, we can say

$$\left\|\sum_{i=1}^{m} e^{-y_i\, w(t)^\top x_i}\, y_i\, r_w(w; x_i)\right\| = L(w(t))\, g(t)^{L-1} \left\|\sum_{i=1}^{m} \frac{e^{-y_i\, w(t)^\top x_i}}{L(w(t))}\, y_i\, r_w\left(\hat{w} + \frac{r(t)}{g(t)}; x_i\right)\right\|$$

Now, for large enough t, $\left\|\sum_{i=1}^{m} \frac{e^{-y_i\, w(t)^\top x_i}}{L(w(t))}\, y_i\, r_w\left(\hat{w} + \frac{r(t)}{g(t)}; x_i\right)\right\|$ is bounded as, using Euler's homogeneous theorem, we can say

$$\lim_{t\to\infty} \hat{w}^\top \left(\sum_{i=1}^{m} \frac{e^{-y_i\, w(t)^\top x_i}}{L(w(t))}\, y_i\, r_w\left(\hat{w} + \frac{r(t)}{g(t)}; x_i\right)\right) = \lim_{t\to\infty} L\left(\sum_{i=1}^{m} \frac{e^{-y_i\, w(t)^\top x_i}}{L(w(t))}\, (\hat{w}^\top x_i)\right)$$

Thus, its a convex combination of positive de ned terms and hence bounded. Thus, we can say

$$k_1 \leq \lim_{t\to\infty} \frac{\left\|\frac{dw(t)}{dt}\right\|}{L(w(t))\, g(t)^{L-1}} \leq k_2 \tag{40}$$

where k_1 and k_2 are some constants. Also, by the assumption

$$\lim_{t\to\infty} \frac{\left\|\frac{dw(t)}{dt}\right\|}{g'(t)} = 1$$

Thus, we can say

$$k_1 \leq \lim_{t\to\infty} \frac{g'(t)}{L(w(t))\, g(t)^{L-1}} \leq k_2$$

Now, as for large enough t, for all i that satisfy $(\hat{w}^\top x_i) = \gamma$ and any ϵ satisfying $0 < \epsilon < \gamma$, we can say

$$\hat{w}^\top \left(\hat{w} + \frac{r(t)}{g(t)}; x_i\right) \geq \gamma +$$

therefore, for large enough t, we can say

$$c_1 e^{-g(t)^{L-(\alpha+\beta)}} \leq \mathcal{L}(w(t)) \leq c_2 e^{-g(t)^{L-(\alpha-\beta)}}$$

where c_1 and c_2 are some constants. Using the above equations, we can say

$$\lim_{t\to\infty} \frac{g'(t)}{e^{-g(t)^{L-(\alpha+\beta)}} g(t)^{L-1-\beta}} \geq k_1 c_1 \tag{41}$$

$$\lim_{t\to\infty} \frac{g'(t)}{e^{-g(t)^{L-(\alpha-\beta)}} g(t)^{L-1-\beta}} \leq k_2 c_2 \tag{42}$$

Now, in Equation 41, multiplying numerator and denominator by $g(t)^{L-1-(\alpha+\beta)}$ and denoting $g(t)^{L-(\alpha+\beta)}$ by $h_1(t)$, we get

$$\lim_{t\to\infty} \frac{h_1'(t)}{e^{-h_1(t)} h_1(t)^{2-\frac{2}{L}}}$$

where α is some constant. This leads us to conclude

$$\lim_{t\to\infty} \frac{h_1(t)}{\log(t)} \geq 1 \implies \lim_{t\to\infty} \frac{g(t)^{L-(\alpha+\beta)}}{\log(t)} \geq 1$$

Similarly, in Equation 42, multiplying numerator and denominator by $g(t)^{L-1-(\alpha-\beta)}$ and denoting $g(t)^{L-(\alpha-\beta)}$ by $h_2(t)$, we get

$$\lim_{t\to\infty} \frac{h_2'(t)}{e^{-h_2(t)} h_2(t)^{2-\frac{2}{L}}}$$

where α is some constant. This leads us to conclude

$$\lim_{t\to\infty} \frac{h_2(t)}{\log(t)} \leq 1 \implies \lim_{t\to\infty} \frac{g(t)^{L-(\alpha-\beta)}}{\log(t)} \leq 1$$

As α can be chosen to be arbitrarily small, we can conclude

$$\lim_{t\to\infty} \frac{g(t)^L}{\log(t)} = 1$$

Substituting this in Equation 40, we get that loss asymptotically goes down at $\frac{1}{t(\log t)^{2-\frac{2}{L}}}$. ■

H.1.2. EXPONENTIAL WEIGHT NORMALIZATION

Theorem For EWN, under Assumptions (A1)-(A3) for gradient flow and $\lim_{t\to\infty} \frac{k\frac{dr(t)}{dt}k}{g'(t)} = 0$, the following hold

1. $\|w(t)\|$ asymptotically grows at $\log(t)^{\frac{1}{L}}$
2. $\mathcal{L}(w(t))$ asymptotically goes down at the rate of $\frac{1}{t(\log t)^2}$.

Proof Consider $w = g(t)\hat{w} + r(t)$, where $\lim_{t\to\infty} \frac{\|r(t)\|}{g(t)} = 0$ and $r(t)^\top \hat{w} = 0$. Now, we make an additional assumption that $\lim_{t\to\infty} \frac{\|\frac{dr(t)}{d(t)}\|}{g'(t)} = 0$.

In this case,

$$\frac{dL(w(t))}{dw} = \sum_{i=1}^{m} e^{-y_i (w(t)^\top x_i)} y_i r_w(w(t); x_i)$$

However, in this case, for a node u,

$$\frac{dw_u(t)}{dt} = \|w_u(t)\|^2 \frac{dL(w(t))}{dw_u}$$

Consider a vector $a(t)$ of equal dimension as w, and its components corresponding to a node u is given by $a_u(t) = \|w_u(t)\|^2 \frac{dL(w(t))}{dw_u}$. Now as we know w converges in direction to $\hat{w}$, therefore, using the update equation above, we can say

$$\lim_{t\to\infty} \frac{\|\frac{dw(t)}{dt}\|}{g(t)^2 \|a(t)\|} = 1$$

Using the update equation for $\frac{dL(w(t))}{dw}$, which is the same as Equation 39, and using the same arguments as for Unnorm in Appendix H.1.1, we can say

$$k_1 \lim_{t\to\infty} \frac{\|\frac{dL(w(t))}{dw}\|}{L(w(t))\, g(t)^{L-1}} k_2 \tag{43}$$

where k_1 and k_2 are some constants. Now, using the expression of $a(t)$, we can say

$$k_3 \lim_{t\to\infty} \frac{\|a(t)\|}{L(w(t))\, g(t)^{L-1}} k_4$$

where k_3 and k_4 are some constants. Now, from the assumption, we know

$$\lim_{t\to\infty} \frac{\|\frac{dw(t)}{dt}\|}{g'(t)} = 1$$

Using the equations above, we can say

$$k_3 \frac{g'(t)}{L(w(t))\, g(t)^{L+1}} k_4 \tag{44}$$

Using similar reasoning as for Unnorm in Appendix H.1.1, we can say, for a large enough t,

$$c_1 e^{-g(t)^L(\hat{\gamma}+\epsilon)} \le L(w(t)) \le c_2 e^{-g(t)^L(\hat{\gamma}-\epsilon)}$$

where c_1 and c_2 are some constants. Substituting this in equation above, we get

$$\lim_{t\to\infty} \frac{g'(t)}{e^{-g(t)^L(\hat{\gamma}+\epsilon)} g(t)^{L+1}} k_3 c_1 \tag{45}$$

$$\lim_{t\to\infty} \frac{g'(t)}{e^{-g(t)^L(\hat{\gamma}-\epsilon)} g(t)^{L+1}} k_4 c_2 \tag{46}$$

Now, using similar arguments as for Unnorm in Appendix H.1.1, we can say

$$\lim_{t\to\infty} \frac{g(t)^L}{\log(t)} = 1$$

Substituting this in Equation 44, we get that loss asymptotically goes down at $\frac{1}{t(\log t)^2}$. ■

H.2. Gradient Descent

Theorem For Exponential Weight Normalization, under Assumptions (B1)-(B4), $\eta(t) = O\left(\log\frac{1}{L}^{c}\right)$ for $c < 1$ and $\lim_{t\to\infty} \frac{kr(t+1)}{g(t+1)} - \frac{r(t)k}{g(t)} = 0$, the following hold

(i) $\|w(t)\|$ asymptotically grows with t as $(\log d(t))^{\frac{1}{L}}$

(ii) $L(w(t))$ asymptotically goes down with t as $\frac{1}{d(t)(\log d(t))^2}$.

Proof Consider $w = g(t)\hat{w} + r(t)$, where $\lim_{t\to\infty} \frac{kr(t)k}{g(t)} = 0$ and $r(t)^\top \hat{w} = 0$. Now, we make additional assumptions that $\lim_{t\to\infty} \frac{kr(t+1)}{g(t+1)} - \frac{r(t)k}{g(t)} = 0$.

Consider a node u in the network that has $\|w_u\| > 0$. The update equations for $v_u(t)$ and $g_u(t)$ are given by

$$g_u(t+1) = g_u(t) - \eta(t)e^{g_u(t)} \frac{v_u(t)^\top \nabla_{w_u} L(w(t))}{\|v_u(t)\|}$$

$$v_u(t+1) = v_u(t) - \eta(t)\frac{e^{g_u(t)}}{\|v_u(t)\|}\left(I - \frac{v_u(t)v_u(t)^\top}{\|v_u(t)\|^2}\right)\nabla_{w_u} L(w(t))$$

Now, we will first estimate $\left\|e^{g_u(t+1)}\frac{v_u(t+1)}{\|v_u(t+1)\|} - e^{g_u(t)}\frac{v_u(t)}{\|v_u(t)\|}\right\|$. Let $\rho_u(t)$ denote $\eta(t)e^{g_u(t)}\|\nabla_{w_u} L(w(t))\|$ and $\theta_u(t)$ denote the angle between $v_u(t)$ and $-\nabla_{w_u} L(w(t))$. We know $\lim_{t\to\infty} \rho_u(t) = 0$ and $\lim_{t\to\infty} \theta_u(t) = 0$. Now, rewriting update equations in terms of these symbols, we get

$$e^{g_u(t+1)} = e^{g_u(t)} - e^{g_u(t)}\cos(\theta_u(t))$$

$$v_u(t+1) = v_u(t) + \rho_u(t)\sin(\theta_u(t))\frac{-\nabla_{w_u} L(w(t))_{v_u(t)^\perp}}{\|-\nabla_{w_u} L(w(t))_{v_u(t)^\perp}\|}$$

where $-\nabla_{w_u} L(w(t))_{v_u(t)^\perp}$ denotes the component of $-\nabla_{w_u} L(w(t))$ perpendicular to $v_u(t)$. Now, using these equations we can say

$$\begin{aligned}\left\|e^{g_u(t+1)}\frac{v_u(t+1)}{\|v_u(t+1)\|} - e^{g_u(t)}\frac{v_u(t)}{\|v_u(t)\|}\right\| &= \left\|e^{g_u(t)} - e^{g_u(t)}\cos(\theta_u(t))\frac{kv_u(t)k}{kv_u(t+1)k}\left(1 - \frac{v_u(t)}{\|v_u(t)\|}\right)\right. \\ &\quad \left. + \frac{e^{g_u(t+1)}\rho_u(t)\sin(\theta_u(t))}{\|v_u(t)\|\|v_u(t+1)\|}\frac{-\nabla_{w_u} L(w(t))_{v_u(t)^\perp}}{\|-\nabla_{w_u} L(w(t))_{v_u(t)^\perp}\|}\right\|\end{aligned} \tag{47}$$

Now as $\lim_{t\to\infty} \frac{\|v_u(t)\|}{\|v_u(t+1)\|} = 1$, therefore we can say

$$\lim_{t\to\infty} \frac{e^{\eta_u(t)\cos(\theta_u(t))}\frac{\|v_u(t)\|}{\|v_u(t+1)\|} - 1}{\eta_u(t)\cos(\theta_u(t))} = 1$$

Now, as $\|v_u(t)\|$ keeps on increasing during the gradient descent trajectory, therefore we can say $\frac{1}{\|v_u(t)\|\|v_u(t+1)\|} \geq k$, where $k > 0$ is some constant. Now dividing both sides of Equation (47) by $e^{\eta_u(t)}\eta_u(t)\cos(\theta_u(t))$ and analyzing the coefficient of the second term on RHS, we get

$$\lim_{t\to\infty} \frac{e^{\eta_u(t)\cos(\theta_u(t))}\sin(\theta_u(t))}{\|v_u(t)\|\|v_u(t+1)\|\eta_u(t)\cos(\theta_u(t))} \geq 0$$

Taking norm on both sides of Equation (47), using Pythagoras theorem and the limits established above, we can say

$$\lim_{t\to\infty} \frac{\left\| e^{\eta_u(t+1)}\frac{v_u(t+1)}{\|v_u(t+1)\|} - e^{\eta_u(t)}\frac{v_u(t)}{\|v_u(t)\|}\right\|}{e^{\eta_u(t)}\eta_u(t)} = 1$$

Now, we also know

$$\lim_{t\to\infty} \frac{\left\| e^{\eta_u(t+1)}\frac{v_u(t+1)}{\|v_u(t+1)\|} - e^{\eta_u(t)}\frac{v_u(t)}{\|v_u(t)\|}\right\|}{g(t+1) - g(t)} = \|w_u\|$$

Now, using equations above and Equation (43), we can say

$$k_1 \leq \lim_{t\to\infty} \frac{g(t+1) - g(t)}{\eta(t)L(w(t))g(t)^{L+1}} \leq k_2$$

where k_1 and k_2 are some constants. Using similar reasoning as for Unnorm in Appendix H.1.1, we can say

$$c_1 e^{-g(t)^L(\alpha+\beta)} \leq L(w(t)) \leq c_2 e^{-g(t)^L(\alpha-\beta)}$$

where c_1 and c_2 are some constants. Substituting in the equation above, we get

$$\lim_{t\to\infty} \frac{g(t+1) - g(t)}{\eta(t)e^{-g(t)^L(\alpha+\beta)}g(t)^{L+1}} \geq k_1c_1$$

$$\lim_{t\to\infty} \frac{g(t+1) - g(t)}{\eta(t)e^{-g(t)^L(\alpha-\beta)}g(t)^{L+1}} \leq k_2c_2$$

These equations govern the rate of $g(t)$ for any $\eta(t)$ that satisfies assumption (A4). Now, to obtain a better closed form, we will need the new assumption on $\eta(t)$, i.e, $\eta(t) = O\left(\log^{\frac{1}{L}-c}\right)$, where $c < 1$.

Now, define a map $d : \mathbb{N} \to \mathbb{R}$, given by $d(t) = \sum_{\tau=0}^{t-1} \eta(\tau)$ and a real analytic function $f(t)$ satisfying $f(d(t)) = g(t)$ for all $t \in \mathbb{N}$ and $\lim_{t\to\infty} \frac{f(d(t+1)) - f(d(t))}{\eta(t)f'(d(t))} = 1$. We will later verify that the $f(t)$ obtained indeed satisfies this for the given $\eta(t)$. Substituting in the equations above,

$$k_1 \leq \lim_{t\to\infty} \frac{f'(d(t))}{L(w(t))f(d(t))^{L+1}} \leq k_2 \tag{48}$$

$$\lim_{t\to\infty} \frac{f'(d(t))}{e^{f(d(t))^{L}(\quad+\quad)} f(d(t))^{L+1}} \quad k_1 c_1$$

$$\lim_{t\to\infty} \frac{f'(d(t))}{e^{f(d(t))^{L}(\quad\quad)} f(d(t))^{L+1}} \quad k_2 c_2$$

As $t \to \infty$, $d(t) \to \infty$, therefore

$$\lim_{t\to\infty} \frac{f'(t)}{e^{f(t)^{L}(\quad+\quad)} f(t)^{L+1}} \quad k_1 c_1$$

$$\lim_{t\to\infty} \frac{f'(t)}{e^{f(t)^{L}(\quad\quad)} f(t)^{L+1}} \quad k_2 c_2$$

Now, using similar arguments as in Appendix H.1.1,

$$\lim_{t\to\infty} \frac{f(t)^{L}}{\log(t)} = 1$$

Substituting in the Equation 48, we get that the loss goes down at $\frac{1}{d(t)(\log d(t))^2}$.

We also verify that $\lim_{t\to\infty} \frac{f(d(t+1)) \quad f(d(t))}{(t) f'(d(t))} = 1$ for $(t) = O(\log^{\frac{1}{L}\quad c})$, where $c < 1$. This can be easily verified by using mean value theorem, and simply verifying $\lim_{t\to\infty} \frac{(t) f''(d(t))}{f'(d(t))} = 0$. Obtaining the expressions for $f'(d(t))$ and $f''(d(t))$, we get

$$\lim_{t\to\infty} \frac{(t) f''(d(t))}{f'(d(t))} = \lim_{t\to\infty} \frac{(t) \quad d(t)(\frac{1}{L} \quad 1) \quad \log d(t)}{d(t)\log(d(t))}$$

As loss goes down at $\frac{1}{d(t)(\log d(t))^2}$, therefore if $(t) = O(\log^{\frac{1}{L}\quad c})$ for $c < 1$, the above limit tends to 0 as $d(t) \to \infty$. ■

Appendix I. Cross-Entropy Loss

In this section, we will provide the corresponding assumptions and theorems, along with their proofs, for cross-entropy loss.

I.1. Notations

Let k denote the total number of classes. $f(w; x_i)$ is a multidimensional function for multi-class classification, let's denote the j^{th} component of the output by $f_j(w; x_i)$. Also, denote the asymptotic normalized margin for i^{th} class corresponding to j^{th} data point ($j \neq y_i$) by $_{i;j}$, i.e, $_{i;j} = \quad_{y_i}(w; x_i) \quad \quad_j(w; x_i)$. Margin for a data point i is defined as $_i = \min_{j \neq y_i} \quad _{i;j}$. The margin for the entire network is defined as $= \min_i \quad _i$.

I.2. Assumptions

The assumptions can be broadly divided into loss function/architecture based assumptions and trajectory based assumptions. The loss functions/architecture based assumptions are shared across both gradient flow and gradient descent.

Loss function/Architecture based assumptions

1 $\ell(y_i; \Phi(w; x_i)) = \log\left(1 + \sum_{j \neq y_i} e^{-(\Phi_{y_i}(w;x_i) - \Phi_j(w;x_i))}\right)$

2 $\Phi(\cdot; x)$ is a C^1 function, for a fixed x

3 $\Phi(\alpha w; x) = \alpha^L \Phi(w; x)$, for some $\alpha > 0$ and $L > 0$

Gradient flow. For gradient flow, we make the following trajectory based assumptions

(A1) There exists a time t_0 such that $\mathcal{L}(w(t_0)) < \log 2$.

(A2) $\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|} := \rho$.

Gradient Descent For gradient descent, we require the learning rate $\eta(t)$ to not grow too fast, and a slightly stronger assumption on loss.

(B1) $\lim_{t\to\infty} \mathcal{L}(w(t)) = 0$ (B2) $\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|} := \rho$

(B3) $\lim_{t\to\infty} \eta(t)\|w_u(t)\|\nabla_{w_u}\mathcal{L}(w(t))\| = 0$ for all u in the network.

The assumption (B3) is mild, as the norm of the gradient of cross-entropy loss goes down exponentially fast as compared to norm of the weights.

I.3. Asymptotic relations between weights and gradients

This section contains the main theorems that establish asymptotic relations between weights and gradients for SWN and EWN. First, we will state a common proposition for both SWN and EWN.

Proposition 11 Under assumption (A1) for gradient flow, for both SWN and EWN, $\lim_{t\to\infty} \mathcal{L}(w(t)) = 0$.

Proof First of all, for cross-entropy loss

$$\frac{d\mathcal{L}(t)}{dw} = \sum_i \frac{\sum_{j \neq y_i} e^{-(\Phi_{y_i}(w;x_i) - \Phi_j(w;x_i))}(\nabla_w \Phi_{y_i}(w;x_i) - \nabla_w \Phi_j(w;x_i))}{1 + \sum_{j \neq y_i} e^{-(\Phi_{y_i}(w;x_i) - \Phi_j(w;x_i))}} \tag{49}$$

Now, using Theorem 1,

$$\frac{d\mathcal{L}(t)}{dt} = \frac{d\mathcal{L}(t)}{dw}^\top \frac{dw(t)}{dt} = \sum_u \|w_u(t)\|^2 \left\|\frac{d\mathcal{L}(t)}{dw_u}\right\|^2$$

Let k be the total number of neurons in the network. Then using the elementary inequality $(\sum_{i=1}^n a_i)^2 \le n \sum_{i=1}^n a_i^2$, we get

$$\frac{d\mathcal{L}(t)}{dt} \ge \frac{1}{k}\left(\sum_u \|w_u(t)\| \left\|\frac{d\mathcal{L}(t)}{dw_u}\right\|\right)^2$$

Again using the fact that $w(t)^\top \frac{d\mathcal{L}(t)}{dw} \le \sum_u \|w_u(t)\| \left\|\frac{d\mathcal{L}(t)}{dw_u}\right\|$, we get

$$\frac{d\mathcal{L}(t)}{dt} \ge \frac{1}{k}\left(w(t)^\top \frac{d\mathcal{L}(t)}{dw}\right)^2 \tag{50}$$

Taking the dot product with w on both sides of Equation (49) and using $w^\top r_w(w; x_i) = L(w; x_i)$ (Euler's homogeneity theorem), we get

$$w(t)^\top \frac{dL(t)}{dw} = -\sum_i \frac{\sum_{j \neq y_i} e^{-(\langle w, x_{y_i} \rangle - \langle w, x_j \rangle)} (\langle w, x_{y_i} \rangle - \langle w, x_j \rangle)}{1 + \sum_{j \neq y_i} e^{-(\langle w, x_{y_i} \rangle - \langle w, x_j \rangle)}}$$

Now, using the fact, that at time t_0, $L(t_0) < \log 2$, which means $\min_i \min_{j \neq y_i} (\langle w, x_{y_i} \rangle - \langle w, x_j \rangle) = \gamma > 0$. Also, as we know, for gradient flow, the loss cannot go up, therefore, for any time $t > t_0$, $\min_i \min_{j \neq y_i} (\langle w, x_{y_i} \rangle - \langle w, x_j \rangle) > \gamma > 0$. Using this, we can say, for any $t > t_0$,

$$w(t)^\top \frac{dL(t)}{dw} \leq -\gamma \sum_i \frac{\sum_{j \neq y_i} e^{-(\langle w, x_{y_i} \rangle - \langle w, x_j \rangle)}}{1 + \sum_{j \neq y_i} e^{-(\langle w, x_{y_i} \rangle - \langle w, x_j \rangle)}}$$

Using the fact that $\ln(1 + t) > \frac{t}{1+t}$ for $t \in (0, 1)$, therefore,

$$w(t)^\top \frac{dL(t)}{dw} \leq -\gamma L \leq -\gamma L(t)$$

Substituting this in Equation (50), we get

$$\frac{dL(t)}{dt} \leq -\gamma^2 \eta^2 L(t)^2$$

Integrating this equation from t_0 to t, we get

$$\frac{1}{L(t)} - \frac{1}{L(t_0)} \geq \gamma^2 \eta^2 (t - t_0) \tag{51}$$

Clearly as t tends to ∞, RHS tends to ∞ and thus L tends to 0. ■

Now, we provide one of our main theorem that establishes gradient convergence implies weight convergence.

Theorem 12 Consider a node u in the network with $\|g_u\| > 0$ and $\lim_{t \to \infty} \|w_u(t)\| = \infty$. Under assumptions (A1), (A2) for gradient flow and (B1)-(B3) for gradient descent, for both SWN and EWN

(i) $\lim_{t \to \infty} \frac{w_u(t)}{\|w_u(t)\|} := w_\infty$ exists. (ii) $w_\infty = g_u$ for some $\alpha > 0$.

Proof Same as in Appendix C. ■

Now, we provide the main theorem that distinguishes SWN and EWN.

Theorem 13 Consider two nodes u and v in the network with $\|g_u\| \neq \|g_v\| > 0$, $\lim_{t \to \infty} \|w_u(t)\| = \infty$ and $\lim_{t \to \infty} \|w_v(t)\| = \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c. Under assumptions (A1), (A2) for gradient flow and (B1)-(B3) for gradient descent,

(i) for SWN, $\lim_{t \to \infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

(ii) for EWN, $\lim_{t \to \infty} \frac{\|w_u(t)\|}{\|w_v(t)\|}$ is either $0, 1$ or $\frac{1}{c}$

Proof Same as in Appendix D ■

I.4. Sparsity Inductive Bias for Exponential Weight Normalisation

The inverse relation between $\|w_u(t)\|$ and $\|\nabla_{w_u} L(w(t))\|$ in the EWN trajectory results in an interesting inductive bias that favours movement along sparse directions.

Proposition 14 Consider two nodes u and v in the network such that $\|g_v\| - \|g_u\| > 0$ and $\|w_u(t)\|, \|w_v(t)\| \to \infty$. Let $\frac{\|g_u\|}{\|g_v\|}$ be denoted by c. Consider any ϵ such that $0 < \epsilon < c$ and $0 < \theta < \frac{\pi}{2}$. Then, the following holds:

(i) There exists a time t_1, such that for all $t > t_1$ both SWN and EWN trajectories have the following properties:

(a) $\frac{\|\nabla_{w_u} L(w(t))\|}{\|\nabla_{w_v} L(w(t))\|} \in [c - \epsilon, c + \epsilon]$ (b) $\frac{w_u(t)}{\|w_u(t)\|} \cdot \frac{\nabla_{w_u} L(w(t))}{\|\nabla_{w_u} L(w(t))\|} \geq \cos(\theta)$

(c) $\frac{w_v(t)}{\|w_v(t)\|} \cdot \frac{\nabla_{w_v} L(w(t))}{\|\nabla_{w_v} L(w(t))\|} \geq \cos(\theta)$.

(ii) for SWN, $\lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = c$

(iii) for EWN, if at some time $t_2 > t_1$,

(a) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} > \frac{1}{(c-\epsilon)\cos(\theta)} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 1$

(b) $\frac{\|w_u(t_2)\|}{\|w_v(t_2)\|} < \frac{\cos(\theta)}{c+\epsilon} \implies \lim_{t\to\infty} \frac{\|w_u(t)\|}{\|w_v(t)\|} = 0$

Proof The proof follows from Appendix E. ■

I.5. Convergence rates

In this section, we provide convergence rate of loss for EWN.

Gradient Flow: We provide a finite-time convergence rate of loss for gradient flow in case of EWN.

Theorem 15 For Exponential Weight Normalization, under assumption (A1), the following hold for $t > t_0$ in case of gradient flow

(i) $\|w(t)\|$ grows with t as $o((\log t)^{\frac{1}{L}})$ (ii) $L(t)$ goes down with t as $O\left(\frac{1}{t}\right)$

Proof Follow from Equation (51) ■

Gradient Descent:

Theorem 16 For Exponential Weight Normalization, under Assumptions (B1)-(B4), $\eta(t) = O\left(\log\frac{1}{L}\right)^{-c}$ for $c < 1$ and $\lim_{t\to\infty} \frac{\|r(t+1)\| - \|r(t)\|}{\|g(t+1)\| - \|g(t)\|} = 0$, the following hold

(i) $\|w(t)\|$ asymptotically grows with t as $(\log d(t))^{\frac{1}{L}}$

(ii) $L(w(t))$ asymptotically goes down with t as $\frac{1}{d(t)(\log d(t))^2}$.

Proof The proof follows Appendix H.2, the only difference is in the gradient update. Let w be represented as $w = g(t)\hat{w} + r(t)$, where $\lim_{t\to\infty} \frac{\|r(t)\|}{g(t)} = 0$. Using Equation (49), we can say

$$k_1 \lim_{t\to\infty} \frac{\left\|\frac{dL(w(t))}{dw}\right\|}{L(w(t))\, g(t)^{L-1}} k_2 \tag{52}$$

where k_1 and k_2 are some constants. As the order remains the same as in the proof for exponential loss, the proof follows from Appendix H.2. ■

Appendix J. Lemma Proofs

Lemma Consider sequence a satisfying the following properties

1. $a_k > 0$
2. $\sum_{k=0}^{\infty} a_k = 1$
3. $\lim_{k\to\infty} a_k = 0$

Then $\sum_{k=0}^{\infty} \frac{a_k}{\sqrt{\sum_{j=0}^{k} a_j^2}} = \infty$

Proof If $\sum_{k=0}^{\infty} a_k^2$ is bounded, then the statement is obvious. Let's consider the case when $\sum_{k=0}^{\infty} a_k^2$ diverges. As $\lim_{k\to\infty} a_k = 0$, therefore there must be an index k_1, such that for $k > k_1$, $a_k < 1$. Now, as $a_k < 1$, therefore $a_k^2 < a_k$. Now, as $\sum_{k=0}^{\infty} a_k^2$ diverges, therefore, there must be an index $k_2 > k_1$, such that for any $k > k_2$, $\sum_{j=k_1}^{k} a_j^2 > \sum_{j=0}^{k_1-1} a_j^2$. Now, for $k > k_2$, we can say

$$\begin{aligned}
\sum_{j=k_1}^{k} \frac{a_j}{\sqrt{\sum_{l=0}^{j} a_l^2}} &> \frac{1}{\sqrt{2}} \sum_{j=k_1}^{k} \frac{a_j}{\sqrt{\sum_{l=k_1}^{j} a_l^2}} \\
&> \frac{1}{\sqrt{2}} \sum_{j=k_1}^{k} \frac{a_j}{\sqrt{\sum_{l=k_1}^{j} a_l}} \\
&> \frac{1}{\sqrt{2}} \sum_{j=k_1}^{k} \frac{a_j}{\sqrt{\sum_{l=k_1}^{k} a_l}} \\
&= \frac{1}{\sqrt{2}} \sqrt{\sum_{j=k_1}^{k} a_j}
\end{aligned}$$

As $\sum_{k=0}^{\infty} a_k$ diverges, therefore $\sum_{k=0}^{\infty} \frac{a_k}{\sqrt{\sum_{j=0}^{k} a_j^2}}$ diverges as well. ■

Lemma Consider two sequences a and b satisfying the following properties

1. $a_k > 0$, $\sum_{k=0}^{\infty} a_k = 1$ and $\lim_{k\to\infty} a_k = 0$

2. $b_0 > 0$, b is increasing and $b_{k+1}^2 \leq b_k^2 + \left(\frac{a_k}{b_k}\right)^2$

Then $\sum_{k=0}^{\infty} \frac{a_k}{b_k} = \infty$.

Proof As we know b is increasing and $b_{k+1}^2 \leq b_k^2 + (\frac{a_k}{b_k})^2$, we get

$$b_k \leq \sqrt{b_0^2 + \sum_{j=0}^{k-1} \left(\frac{a_j}{b_j}\right)^2} \leq \sqrt{b_0^2 + \frac{1}{b_0^2} \sum_{j=0}^{k-1} a_j^2}$$

Using this, we can say

$$\sum_{j=0}^{k} \frac{a_j}{b_j} \geq \sum_{j=0}^{k} \frac{a_j}{\sqrt{b_0^2 + \frac{1}{b_0^2} \sum_{l=0}^{j-1} a_l^2}} \geq \sum_{j=0}^{k} \frac{a_j}{\sqrt{b_0^2 + \frac{1}{b_0^2} \sum_{l=0}^{k-1} a_l^2}}$$

Now, if $\sum_{k=0}^{\infty} a_k^2$ does not diverge to infinity, then b remains bounded using the bound above and then its trivial to establish that $\sum_{k=0}^{\infty} \frac{a_k}{b_k}$ diverges. In case, $\sum_{k=0}^{\infty} a_k^2$ diverges to infinity, then there must be an index k_1 such that for any $k > k_1$, we can say $\sum_{j=0}^{k-1} a_j^2 \geq b_0^4$. So, for $k > k_1$, we can say

$$\sum_{j=0}^{k} \frac{a_j}{b_j} \geq \sum_{j=0}^{k} \frac{b_0}{2} \frac{a_j}{\sqrt{\sum_{l=0}^{k-1} a_l^2}}$$

Now, as we have assumed a_k tends to zero, so there must be an index k_2 such that for any $k > k_2$, $a_k \leq 1$. Also, as we have assumed $\sum_{j=0}^{\infty} a_j^2$ diverges, therefore there must be an index $k_3 > k_2$, such that for $k > k_3$, $\sum_{j=k_2}^{k} a_j^2 \geq \sum_{j=0}^{k_2} a_j^2$. Using these things and that if $a_j \leq 1$, then $a_j^2 \leq a_j$, we can say for $k > k_3$,

$$\sum_{j=k_3}^{k} \frac{a_j}{b_j} \geq \sum_{j=k_3}^{k} \frac{b_0}{2} \frac{a_j}{\sqrt{\sum_{l=k_3}^{k-1} a_l}} \geq \frac{b_0}{2} \sqrt{\sum_{j=k_3}^{k-1} a_j}$$

Now, as $\sum_{k=0}^{\infty} a_k$ diverges, thus $\sum_{k=0}^{\infty} \frac{a_k}{b_k}$ diverges as well. ■

Appendix K. Integral Form of Stolz-Cesaro Theorem

We first state the Stolz-Cesaro Theorem.

Theorem (Muresan, 2015) Assume that $\{a_k\}_{k=1}^{\infty}$ and $\{b_k\}_{k=1}^{\infty}$ are two sequences of real numbers such that $\{b_k\}_{k=1}^{\infty}$ is strictly monotonic and diverging. Additionally, $\lim_{k\to\infty} \frac{a_{k+1} - a_k}{b_{k+1} - b_k} = L$ exists, then $\lim_{k\to\infty} \frac{a_k}{b_k}$ exists and is equal to L.

Now, we state and prove the Integral Form of Stolz-Cesaro Theorem.

Theorem Consider two functions $f(t)$ and $g(t)$ greater than zero satisfying $\int_a^b f(t)dt < \infty$ and $\int_a^b g(t)dt < \infty$ for every finite a, b. For any time t, its known that $\int_t^{\infty} f(t)dt = \infty$ and $\int_t^{\infty} g(t)dt = \infty$. If $\lim_{t\to\infty} \frac{f(t)}{g(t)}$ exist and is equal to L, then $\lim_{t\to\infty} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt}$ exists for any c and is equal to L.

Proof Case 1:$L = 0$ or 1 :

We will prove for $L = 1$. The case for 0 can be handled similarly. For any $M > 0$, there must exist a time $t_1 > c$, such that $\frac{f(t)}{g(t)} > M$, for $t > t_1$. Thus we can say for $t > t_1$,

$$\int_c^t f(t)dt > \int_c^{t_1} f(t)dt + M \int_{t_1}^t g(t)dt$$

Adding $M \int_c^{t_1} g(t)dt$ on both the sides, we get

$$\int_c^t f(t)dt + M \int_c^{t_1} g(t)dt > \int_c^{t_1} f(t)dt + M \int_c^t g(t)dt$$

Dividing both sides by $\int_c^t g(t)dt$ and taking limsup $t \to 1$ (using also the fact that $\int_a^b f(t)dt < 1$ and $\int_a^b g(t)dt < 1$ for every finite a, b), we get

$$\limsup_{t \to 1} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt} > M$$

Similarly the equation holds for liminf as well. Thus, both liminf and limsup are greater than M for any M . Hence $\lim_{t \to 1} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt} = 1$.

Case 2: L is finite

In this case, there must exist some time $t_1 > c$, such that $L - < \frac{f(t)}{g(t)} < L + $. Thus, we can say for $t > t_1$,

$$\int_c^{t_1} f(t)dt + (L -) \int_{t_1}^t g(t)dt < \int_c^t f(t)dt < \int_c^{t_1} f(t)dt + (L +) \int_{t_1}^t g(t)dt$$

Taking the left inequality, adding $(L -) \int_c^{t_1} g(t)dt$ on both the sides, dividing both the sides by $\int_c^t g(t)dt$ and taking $\liminf_{t \to 1}$, we get

$$L - \liminf_{t \to 1} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt}$$

Similarly, taking the right inequality, adding $(L +) \int_c^{t_1} g(t)dt$ on both the sides, dividing both the sides by $\int_c^t g(t)dt$ and taking $\limsup_{t \to 1}$, we get

$$\limsup_{t \to 1} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt} \quad L +$$

Using the two inequalities, we get, for any > 0,

$$\limsup_{t \to 1} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt} \quad \liminf_{t \to 1} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt} \quad 2$$

Thus, $\lim_{t \to 1} \frac{\int_c^t f(t)dt}{\int_c^t g(t)dt}$ exists and is equal to . ■

Appendix L. Standard Weight Normalization is not Locally Lipschitz in its parameters

In this section, we will denote w by θ so as to be consistent with the notaion in Lyu and Li (2020). SWN(in its parameters and v) is also a homogeneous network. Therefore, results from Lyu and Li (2020) should directly apply to the case of SWN as well. However, a crucial point to be noted is that it is not even locally Lipschitz around $\|v_u\| = 0$. Therefore, the assumptions from Lyu and Li (2020) do not hold.

However, during gradient descent or gradient flow, if started from a point $\|v_u\| > 0$, for all u, then during the entire trajectory $\|v_u\|$ cannot go down. Therefore, the network is still locally Lipschitz along the trajectory it takes. Examining the proofs from Lyu and Li (2020), its clear that the proof regarding monotonicity of margin and convergence rates are just dependent on the path that gradient descent/ flow takes and thus the proofs hold.

However, the result regarding the limit points of $\frac{\theta_k}{\rho_k}$ do not hold. One of the crucial theorems the proof relies on is stated below

Theorem Let $\{x_k \in \mathbb{R}^d : k \in \mathbb{N}\}$ be a sequence of feasible points of an optimization problem (P), $\{\eta_k > 0 : k \in \mathbb{N}\}$ and $\{\epsilon_k > 0 : k \in \mathbb{N}\}$ be two sequences x_k is an (η_k, ϵ_k)-KKT point for every k and $\eta_k \to 0, \epsilon_k \to 0$. If $x_k \to x$ as $k \to \infty$ and MFCQ holds at x, then x is a KKT point of (P)

The above statement requires MFCQ to be satisfied, that was shown in Lyu and Li (2020) assuming local lipschitzness/smoothness as true however, in this case, for gradient flow, as $\|v_u\|$ does not grow, while $|\rho_u| \to \infty$, therefore the convergent point of $\frac{\theta_k}{\rho_k}$ will always have the component corresponding to v_u as 0. Thus, the network is not locally lipschitz at x and the proof that MFCQ holds is violated. Similarly, for gradient descent as well, it can't be said that v_u has a non-zero component in $\frac{\theta_k}{\rho_k}$. Thus, the proof does not hold.

Appendix M. Experiment Details

In all the experiments, techniques for handling numerical underflow were used as described in Lyu and Li (2020). However, the learning rate they used was $\mathcal{O}\left(\frac{1}{L}\right)$, but in our case, we generally modify it to be $\mathcal{O}\left(\frac{1}{L^c}\right)$, where $c < 1$.

M.1. Lin-Sep

The learning rate used was $\frac{\eta(t)}{L^{0.97}}$, so that it speeds up at the beginning of training, but slows down as loss approaches e^{-300}. The constant $\eta(t)$ was initialized at 0.01, and was increased by a factor of 1.1 every time loss went down and decreased by a factor of 1.1 every time loss went up after a gradient step. Its value was capped at 0.01 for EWN and SWN.

M.2. Simple-Traj

The learning rate used was $\frac{\eta(t)}{L^{0.9}}$, so that it speeds up at the beginning of training, but slows down as loss approaches e^{-50}. The constant $\eta(t)$ was initialized at 0.01, and was increased by a factor of 1.1 every time loss went down and decreased by a factor of 1.1 every time loss went up after a gradient step. Its value was capped at 0.1 for EWN and Unnorm.

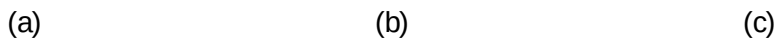


Figure 7: Demonstration of Results for SWN in *Lin-Sep* experiment: (a) Evolution of $\|w_u\|$ (b) Cosine between weights and gradients for weights 5, 6 and 8. (c) Weight and gradient norms for weights 5, 6 and 8.

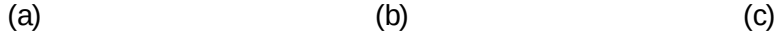


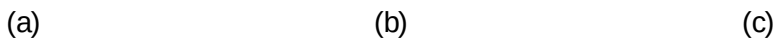
Figure 8: Demonstration of Results for EWN in *XOR* experiment with ReLU-square activation: (a) Evolution of $\|w_u\|$ (b) Cosine between weights and gradients for weights 0, 1, 13 and 17. (c) Weight and gradient norms for weights 0, 1, 13 and 17.

M.3. XOR

The learning rate used was $\frac{k(t)}{L^{0.93}}$ for SWN and Unnorm, while $\frac{k(t)}{L^{0.8}}$ for EWN, so that it speeds up at the beginning of training, but slows down as loss approaches 0. The constant $k(t)$ was initialized at 0.01, and was increased by a factor of 1.1 every time loss went down and decreased by a factor of 1.1 every time loss went up after a gradient step. Its value was capped at 0.1 for EWN and Unnorm and 0.01 for SWN.

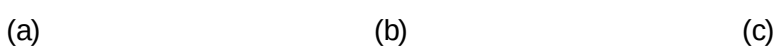
M.4. Convergence rate experiment

For all SWN, EWN and Unnorm, the learning rate was constant $= 0.001$ and they were trained for 5000 steps. All the networks were explicitly initialized to the same point in function space.



(a) (b) (c)

Figure 9: Demonstration of Results for SWN in XOR experiment with ReLU-square activation: (a) Evolution of $\|w_u\|$ (b) Cosine between weights and gradients for weights 0, 1, 13 and 17. (c) Weight and gradient norms for weights 0, 1, 13 and 17.



(a) (b) (c)

Figure 10: Demonstration of Results for EWN on MNIST dataset with 2-class classification: (a) Evolution of $\|w_u\|$ (b) Cosine between weights and gradients for weights 96 and 105. (c) Weight and gradient norms for weights 96 and 105.

M.5. Pruning Experiments

The learning rate used was $\frac{k(t)}{L}$. The constant $k(t)$ was initialized at 0.01, and was increased by a factor of 1.1 every time loss went down and decreased by a factor of 1.1 every time loss went up after an epoch.

Appendix N. Demonstration of Theorems on various datasets

In this section, we demonstrate Theorem 3 and 4 on various datasets: Lin-Sep , XOR and MNIST (2-class classification).

N.1. Lin-Sep

We demonstrate Theorem 3 and Theorem 4 for EWN and SWN on a linearly separable dataset (Lin-Sep) in Figure 1 and 7 respectively. As can be seen in Figure 7, for weights 5, 6 and 8,

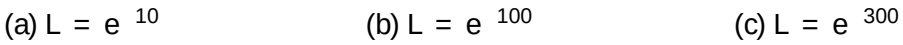


Figure 11: Convergence rate of EWN, SWN and Unnorm on the MNIST dataset for seed - 2356 at different loss values

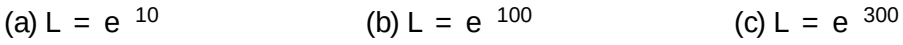


Figure 12: Convergence rate of EWN, SWN and Unnorm on the MNIST dataset for seed - 3576 at different loss values

whose norms keep on growing, weights and gradients eventually become oppositely aligned, and their norms are directly proportional to each other.

N.2. XOR

In this experiment, we train a 2-layer network with 20 hidden neurons and ReLU-square activation on XOR dataset, till a loss value of e^{-40}. We demonstrate Theorems 3 and 4 for EWN and SWN in Figure 8 and Figure 9 respectively.

As can be seen in Figure 8, for weights 0, 1, 13 and 17, whose norms keep on growing, weights and gradients eventually become oppositely aligned, and their norms are inversely proportional to each other.

Similarly, in Figure 9, for weights 0, 1, 13 and 17, whose norms keep on growing, weights and gradients eventually become oppositely aligned, and their norms are directly proportional to each other.

N.3. MNIST

In this experiment, we train a 2-layer network with 128 hidden neurons and ReLU-square activation on MNIST dataset with 2 classes $\{0, 1\}$. Since even after considering just 2 classes, this dataset is huge, therefore training takes longer. Therefore, we only consider exponential weight normalized network in this case.

We demonstrate Theorems 3 and 4 for MNIST dataset for EWN in Figure 10. As can be seen, for weights 96 and 105, whose norms keep on growing, weights and gradients eventually become oppositely aligned, and their norms are inversely proportional to each other.

Appendix O. Convergence rate plots for pruning experiment

In this section, we provide convergence rate plots for the pruning experiments.

The convergence rate for 2 different seeds at various loss levels are shown in Figure 11 and 12. As can be seen, initially the convergence rate of all the normalizations are comparable. But at extremely low loss values, Unnorm becomes slightly faster as compared to SWN or EWN. Note that the results regarding asymptotic convergence rate do not apply in this case, as we are training at extremely high learning rates of $\frac{1}{L}$.

Appendix P. Verification of Assumptions for various datasets

In this section, we verify the assumptions on three datasets: Lin-Sep, XOR and MNIST (2-class classification). Note that in Figures 13, 14, 15, 16 and 17, plots demonstrating the components of unit gradient vector, each line corresponds to a single parameter of the network, while in the plots demonstrating the evolution of $\langle \hat{w}_u(t), -\nabla_{w_u} \mathcal{L}(w(t)) \rangle$, each line corresponds to a neuron of the network.

P.1. Lin-sep

We verify assumptions (B1)-(B3) for the Lin-Sep experiment for EWN and SWN in Figure 13 and Figure 14 respectively. As can be seen, for both the cases, the components of unit gradient vector become constant as training proceeds. Another thing to note, is that even for an aggressive learning rate schedule of the form $\frac{1}{L^{0.97}}$, $\langle \hat{w}_u(t), -\nabla_{w_u} \mathcal{L}(w(t)) \rangle$ still goes down to 0.

P.2. XOR

In this experiment, we train a 2-layer network with 20 hidden neurons and ReLU-square activation on XOR dataset, till a loss value of 10^{-40}. We verify assumptions (B1)-(B3) for the XOR experiment for EWN and SWN in Figure 15 and Figure 16 respectively. As can be seen, for both the cases, the components of unit gradient vector become constant as training proceeds. Another thing to note, is that even for an aggressive learning rate schedule of the form $\frac{1}{L^{0.93}}$, $\langle \hat{w}_u(t), -\nabla_{w_u} \mathcal{L}(w(t)) \rangle$ still goes down to 0.

P.3. MNIST

In this experiment, we train a 2-layer network with 128 hidden neurons and ReLU-square activation on MNIST dataset with 2 classes $\{0, 1\}$. Since even after considering just 2 classes, this dataset is

(a) (b)

Figure 13: Verification of Assumptions for EWN in Lin-Sep experiment: (a) Evolution of $\frac{\nabla_w L(t)}{\|\nabla_w L(t)\|}$ (b) Evolution of $\|w_u(t)\|\|\nabla_{w_u} L(t)\|$

(a) (b)

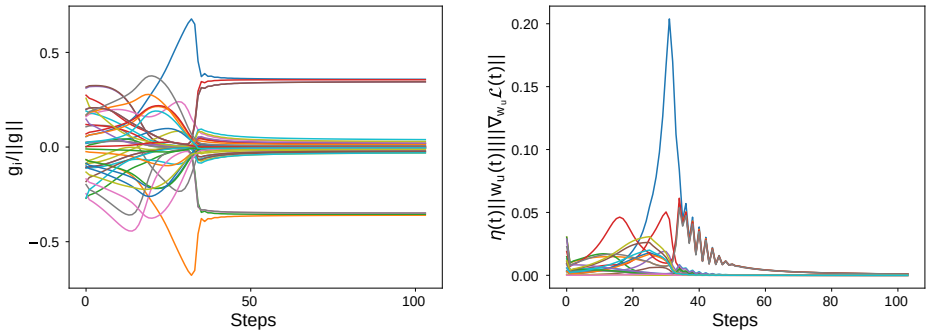
Figure 14: Verification of Assumptions for SWN in Lin-Sep experiment: (a) Evolution of $\frac{\nabla_w L(t)}{\|\nabla_w L(t)\|}$ (b) Evolution of $\|w_u(t)\|\|\nabla_{w_u} L(t)\|$

huge, therefore training takes longer. Therefore, we only consider exponential weight normalized network in this case.

We verify assumptions (B1)-(B3) for MNIST dataset for EWN in Figure 17. As can be seen, the components of unit gradient vector become constant as training proceeds. Another thing to note, is that even for an aggressive learning rate schedule of the form $\frac{1}{L^{0.9}t}$, $\|w_u(t)\|\|\nabla_{w_u} L(w(t))\|$ still goes down to 0.

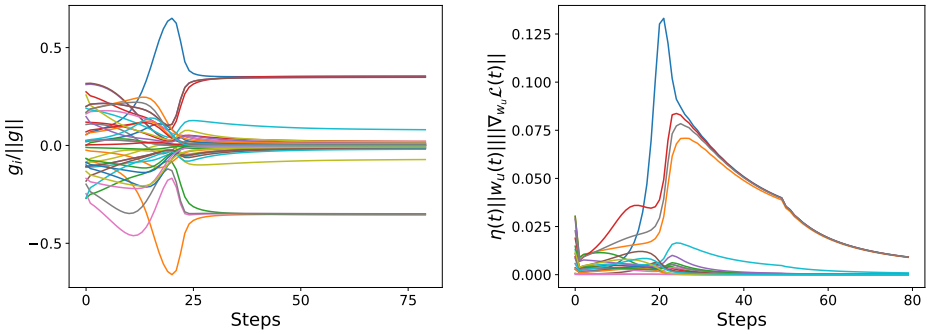
Appendix Q. Pruning algorithm

We will explain the pruning algorithm used in Figure 5b. The same algorithm is used for SWN, EWN as well as the unnormalized net.



(*a*) (*b*)

Figure 15: **Verification of Assumptions for EWN in XOR experiment with ReLU-square activation:** (a) Evolution of $\frac{r_{\mathbf{w}}L(t)}{k\,r_{\mathbf{w}}L(t)k}$ (b) Evolution of $\eta(t)k\mathbf{w}_u(t)kk\,r_{\mathbf{w}_u}L(t)k$



(*a*) (*b*)

Figure 16: **Verification of Assumptions for SWN in XOR experiment with ReLU-square activation:** (a) Evolution of $\frac{r_{\mathbf{w}}L(t)}{k\,r_{\mathbf{w}}L(t)k}$ (b) Evolution of $\eta(t)k\mathbf{w}_u(t)kk\,r_{\mathbf{w}_u}L(t)k$

Let t_1 and t_2 denote the optimization iteration indices when log-loss has a value of 10 and 100 respectively. Consider three pruning strategies available at the time instant given by t_2:

- Prune weights on the basis of $k\mathbf{w}_u(t_2)k$
- Prune weights on the basis of $k\mathbf{w}_u(t_2)$ $\mathbf{w}_u(0)k$
- Prune weights on the basis of $k\mathbf{w}_u(t_2)$ $\mathbf{w}_u(t_1)k$

For a given level of pruning, we try all the 3 strategies, and then pick the one with the best test performance.

Variants of option (a) are the most prevalent pruning algorithms. The non-standard options of (b) and (c) represent the intuition of pruning neurons whose weight has not moved in the recent