ACTIVATION FUNCTION DESIGN FOR DEEP NETWORKS: LINEARITY AND EFFECTIVE INITIALISATION

†‡ **M. Murray,** † **V. Abrol,** †‡ **J. Tanner**
† Mathematical Institute, University of Oxford, UK
‡ The Alan Turing Institute, London, UK
[murray,abrol,tanner]@maths.ox.ac.uk

ABSTRACT

The activation function deployed in a deep neural network has great influence on the performance of the network at initialisation, which in turn has implications for training. In this paper we study how to avoid two problems at initialisation identified in prior works: rapid convergence of pairwise input correlations, and vanishing and exploding gradients. We prove that both these problems can be avoided by choosing an activation function possessing a sufficiently large linear region around the origin, relative to the bias variance σ_b^2 of the network's random initialisation. We demonstrate empirically that using such activation functions leads to tangible benefits in practice, both in terms test and training accuracy as well as training time. Furthermore, we observe that the shape of the nonlinear activation outside the linear region appears to have a relatively limited impact on training. Finally, our results also allow us to train networks in a new hyperparameter regime, with a much larger bias variance than has previously been possible.

Keywords Activation Function Design, Deep Learning, Initialisation, Random Networks.

1 Introduction

The preactivations $\mathbf{h}^{(l)} \in \mathbb{R}^{N_l}$ and activations $\mathbf{z}^{(l)}(\mathbf{x}) \in \mathbb{R}^{N_l}$ of an input vector $\mathbf{x} \in \mathbb{R}^{N_0}$ at each layer $l \in [L]$ of a fully-connected, deep, feedforward neural network, with weight matrices $\mathbf{W}^{(l)} \in \mathbb{R}^{N_l \times N_{l-1}}$ and bias vectors $\mathbf{b}^{(l)} \in \mathbb{R}^{N_l}$, are computed by the following pair of recurrence relations,

$$\mathbf{z}^{(l)}(\mathbf{x}) = \mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l-1)}, \ \mathbf{h}^{(l)}(\mathbf{x}) = \phi(\mathbf{z}^{(l)}(\mathbf{x})), \tag{1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the pointwise nonlinearity, referred to as the activation function, deployed at each layer. The impact of the choice of activation function ϕ on a wide range of a properties, including for instance the trainability and generality of the network, is an active area of research [1, 2, 3, 4, 5, 6]. In this paper we study specifically how the choice of activation function impacts the performance of the network at initialisation, and the subsequent implications of this on training. We build on prior work [7, 8, 9, 10] concerning randomly initialised deep neural networks (DNNs), focusing on two objectives: first, ensuring deeper information propagation by slowing the rate of convergence of pairwise input correlations, and second, avoiding vanishing and exploding gradients [11] by at least approximately achieving dynamical isometry. In [7] estimates for how a) the length of the preactivations of an input and b) the correlation between the preactivations of two distinct inputs evolve with depth were derived. In both [7, 8] the authors investigated how the choice of the activation function ϕ and the variance of the random weights σ_w^2 and biases σ_b^2 effect the lengths of and correlations between preactivations. A key finding of [8] was the identification of a condition on (σ_w^2, σ_b^2), termed initialisation on the edge of chaos (EOC), which is necessary in order to avoid asymptotic exponentially fast convergence of both the lengths and correlations. Furthermore, [10] identified for a broad range of activation function that by letting $\sigma_b^2 \to 0$ then the rate of convergence in correlation can be reduced away from the fixed point, i.e., non-asymptotically. We review this line of research in detail in Section 2.1. A related line of

work [9, 12, 13] proposed that the problem of vanishing and exploding gradients can be mitigated by ensuring that the spectrum of the input-output Jacobian of the network is concentrated around one, a property termed dynamical isometry. In addition, it was shown that initialisation on the EOC implies that the mean of the spectrum of the input-output Jacobian is one. Therefore, dynamical isometry can at least approximately be achieved by making, through careful selection of $(\phi, \sigma_w^2, \sigma_b^2)$, the variance $\sigma_{JJ^T}^2$ of the spectrum close to zero. More specifically, the authors highlighted how this can be achieved for particular activation functions by using an orthogonal initialisation scheme for the weights and again by letting $\sigma_b^2 \to 0$. A number of algorithmic and architectural methods have been proposed and deployed in practice to overcome the problem of vanishing and exploding gradients across a variety of contexts, see for instance [1, 5, 14, 15, 16, 17, 18, 19]. For brevity we do not review these here, but remark that the line of work described above can be viewed as complimentary to these methods, focusing specifically on the network at initialisation.

Letting $\sigma_b^2 \to 0$ is unsatisfactory for a number of reasons, first, $\sigma_b^2 \to 0$ typically results in the length of the preactivations converging towards zero with depth. Second, and as demonstrated by our experiments in Section 4, more affine initialisations with a small but not overly small σ_b^2 appear to provide the best training outcomes in practice. Third and finally, this strategy does not leverage our freedom in designing the activation function deployed in order to achieve these two goals. In this paper we introduce and study a novel set of scaled, bounded activation functions which are linear in some neighbourhood of the origin. For functions in this set we derive upper bounds controlling both the rate of convergence of correlations between preactivations, as well as the variance of the spectrum of the input-output Jacobian. Notably these bounds depend on the ratio σ_b^2/a^2, where a denotes the size of the linear region of the activation function, and approach zero as $\sigma_b^2/a^2 \to 0$. The theory presented in this paper therefore provides a rigorous explanation for the observation that networks equipped with an activation function which approximates the identity at the origin often train well (see for example [20]). Furthermore, our experiments demonstrate the benefits of deploying these scaled, bounded activations in practice, in terms of train and test accuracy, training time and allowing us to train networks with a much larger bias variance than was previously possible.

The structure of this paper is as follows, in Section 2 we review the foundational results of prior works and summarise the contributions of this paper, in particular Theorem 2.1. In Section 3 we derive Theorem 2.1, then in Section 4 we investigate the implications of our theory for training deep networks in practice.

2   Principles for initialising deep networks

As referenced to in Section 1, the prior works on which our results are based can be divided into two distinct themes, one concerned with the dynamics of the preactivation correlations in the forward pass [7, 8, 10], and the other dynamical isometry [9, 12, 13]. We review these two themes in detail in Sections 2.1 and 2.2 respectively, then in Section 2.3 we introduce and present the key contribution of this paper, Theorem 2.1. Note that in what follows $Z, Z_1, Z_2 \sim \mathcal{N}(0, 1)$ are considered to be independent and identically distributed standard Gaussian random variables. Furthermore, the standard Gaussian measures in one and two dimensions are be denoted by γ and $\gamma^{(2)}$ respectively.

2.1   Dynamics of the preactivation correlations in the forward pass

In [21] it was established that the outputs of certain random, feedforward, single hidden layer neural networks converge in distribution to centred Gaussian processes as the width of the hidden layer goes to infinity. More recently in [22, 23], this result was extended to random multilayer neural networks, with the preactivations at each layer being found to converge in distribution to centred Gaussian processes in the large width limit. The behaviour of a centred Gaussian process is fully described by its covariance matrix. In particular, for networks whose forward pass is described by (1), with the weights and biases at each layer $l \in [L]$ being mutually independent, centred Gaussian random variables with variances σ_w^2/N_{l-1} and σ_b^2 respectively, and assuming that the same activation function ϕ is deployed at each neuron, then the kernel used to compute the entries of the covariance matrix of the Gaussian process at the lth layer is defined by the following recurrence relation,

$$\kappa^{(l)}(\mathbf{x}^{(\alpha)}, \mathbf{x}^{(\beta)}) = \sigma_w^2 \mathbb{E}[\phi(z_i^{(l)}(\mathbf{x}^{(\alpha)}))\phi(z_i^{(l)}(\mathbf{x}^{(\beta)}))] + \sigma_b^2. \tag{2}$$

Note here that $i \in [N_l]$ can be any entry of the preactivation. This recurrence relation has been highlighted a number of times in a variety of contexts, the one most relevant for this paper being a mean field approximation used to study signal propagation in the forward pass [7, 8, 10]. In this work we do not focus on the correspondence between wide, random neural networks and Gaussian processes as in [22, 23]. Rather, as in [7, 8, 10], we adopt the Gaussian process or mean field approximation in order to better understand and analyse key statistics of the network at initialisation. In particular, and as highlighted in [7], in the infinite width limit one can interpret the variance of the preactivations of an input at a given layer, corresponding to (2) with $\alpha \neq \beta$, as the expected euclidean length of said input at said layer. Adopting the notation of [7], then the sequence of variances, or expected lengths, $(q_\alpha^{(l)})_{l=1}^L$, associated with an arbitrary input $\mathbf{x}^{(\alpha)}$,

are generated by the following recurrence relation,

$$q_\alpha^{(l)} = V_\phi(q_\alpha^{(l-1)}) := \sigma_w^2 \int_{\mathbb{R}} \phi\left(\sqrt{q_\alpha^{(l-1)}}z\right)^2 d\gamma(z) + \sigma_b^2 = \sigma_w^2 \mathbb{E}\left[\phi\left(\sqrt{q_\alpha^{(l-1)}}Z\right)^2\right] + \sigma_b^2 \tag{3}$$

for $l \in [L]L$ with $q_\alpha^{(0)} := ||\mathbf{x}^{(\alpha)}||_2^2$. As demonstrated in Figure 1(a), analysis of the variance function $V_\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ provides insight into the impact of $(\phi, \sigma_w^2, \sigma_b^2)$ on the dynamics of the expected length of the preactivations with depth. Indeed, unless $(\phi, \sigma_w^2, \sigma_b^2)$ is chosen appropriately then $q_\alpha^{(l)}$ can either rapidly converge towards zero or diverge. This problem of vanishing or exploding activation lengths can be mitigated by choosing $(\phi, \sigma_w^2, \sigma_b^2)$ in order that there exists a stable fixed point $q^* = V_\phi(q^*)$ for which $q_\alpha^{(l)} \to q^*$ for all inputs $\mathbf{x}^{(\alpha)}$, see for example Tanh with $\sigma_b^2 = 0.1$ in Figure 1(a). We remark that this problem was also studied in [24] in the context of finite width deep ReLU networks.

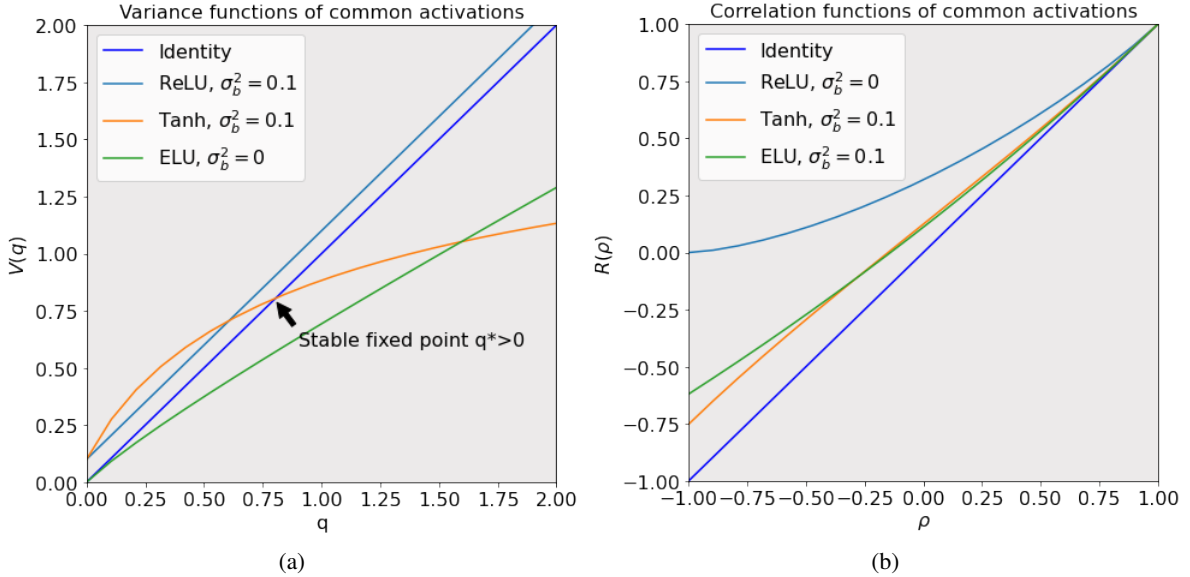(a)   (b)

Figure 1: (a) Variance functions for ReLU, Tanh and ELU with $\sigma_b^2 \in \{0, 0.1\}$. With $\sigma_b^2 = 0.1$ the variance function associated with ReLU has no fixed point and $q^{(l)} \to \infty$ from any initial variance $q^{(0)} \in \mathbb{R}_{\geq 0}$. With $\sigma_b^2 = 0$ the variance function associated with ELU has a unique and marginally stable point $q^* = 0$, therefore $q^{(l)} \to 0$ from any initial variance $q^{(0)} \in \mathbb{R}_{\geq 0}$. Finally, with $\sigma_b^2 = 0.1$, the variance function associated with Tanh has a unique and stable fixed point $q^* > 0$, therefore $q^{(l)} \to q^*$ from any initial variance $q^{(0)} \in \mathbb{R}_{\geq 0}$. (b) Correlation functions associated with Relu, Tanh and ELU with $\sigma_b^2 \in \{0, 0.1\}$. For ReLU, and as illustrated by (a), for there to exist a fixed point of the variance function it is necessary that $\sigma_b^2 = 0$. Note that without a fixed point q^* it is not possible to define a correlation function which is fixed with depth. Relative to the other activation functions considered, the correlation function of ReLU is further from the identity function, and as a result correlations converge faster with the depth. This can also be observed in Figure 10 in Appendix B.6. Note that in both (a) and (b) σ_w^2 is chosen so that $\chi_1 = 1$, see (7).

In [7] the covariance and correlation between preactivations of different inputs was also analysed and a recurrence relation for the covariance derived, which is equivalent to (2) with $\alpha \neq \beta$. In the infinite width limit the covariance can be interpreted as the inner product between the preactivations of two inputs, and the correlation as the angle. The evolution of these quantities with depth has important implications for the performance of the network at initialisation as well as subsequent training. Denoting the covariance between the preactivations of $\mathbf{x}^{(\alpha)}$ and $\mathbf{x}^{(\beta)}$ at the lth layer as $q_{\alpha\beta}^{(l)}$, with respective variances $q_\alpha^{(l)}$ and $q_\beta^{(l)}$, then it was shown in [7] that

$$q_{\alpha\beta}^{(l)} = \sigma_w^2 \int_{\mathbb{R}^2} \phi(u_1)\phi(u_2)d\gamma^{(2)}(z_1, z_2) + \sigma_b^2,$$
$$u_1 := \sqrt{q_\alpha^{(l-1)}}z_1, \; u_2 := \sqrt{q_\beta^{(l-1)}}(\rho_{\alpha\beta}^{(l-1)}z_1 + \sqrt{1 - (\rho_{\alpha\beta}^{(l-1)})^2}z_2), \tag{4}$$

where $\rho_{\alpha\beta}^{(l-1)} = q_{\alpha\beta}^{(l-1)}/(q_\alpha^{(l-1)}q_\beta^{(l-1)})$ is the correlation at the previous layer. If the variance function defined in (3) has a fixed point $q^* > 0$ then an additional key benefit beyond that already discussed is that the analysis of the evolution of

the covariance and correlation with depth can be simplified. In particular, assuming that the inputs are normalised so that $q_\alpha^{(0)} = q_\beta^{(0)} = q^* > 0$, then $q_\alpha^{(l)} = q_\beta^{(l)} = q^*$ for all $l \in [L]$. In this setting the sequence of correlations $(\rho_{\alpha\beta}^{(l)})_{l=1}^L$ are generated by the following recurrence relation,

$$\rho_{\alpha\beta}^{(l)} = R_{\phi,q^*}(\rho_{\alpha\beta}^{(l-1)}) := \frac{q_{\alpha\beta}^{(l)}}{q^*} = \frac{\sigma_w^2}{q^*} \int_{\mathbb{R}} \phi(u_1)\phi(u_2) d\gamma^{(2)}(z_1, z_2) + \frac{\sigma_b^2}{q^*} = \frac{\sigma_w^2}{q^*} \mathbb{E}[\phi(U_1)\phi(U_2)] + \frac{\sigma_b^2}{q^*}. \qquad (5)$$

Here we refer to $R_{\phi,q^*} : [-1, 1] \to [-1, 1]$ as the correlation function, and $U_1 := \sqrt{q^*}Z_1$, $U_2 := \sqrt{q^*}(\rho Z_1 + \sqrt{1 - \rho^2}Z_2)$, where ρ is the input argument $R_{\phi,q^*}(\rho)$, are dependent Gaussian random variables with $U_1, U_2 \sim \mathcal{N}(0, q^*)$.

Analysing the univariate recurrence relation $\rho^{(l)} = R_{\phi,q^*}(\rho^{(l-1)})$ allows for the identification of both depth limits, beyond which information cannot propagate, as well as issues around stability, in a manner analogous to that of the dynamical systems perspective given in [25]. In order to appreciate these points, observe, as in [7], that by construction as long as the correlation function is well defined on $[-1, 1]$ then one is always a fixed point,

$$R_{\phi,q^*}(1) = \frac{\sigma_w^2}{q^*}\mathbb{E}[\phi(U_1)^2] + \frac{\sigma_b^2}{q^*} = \frac{V(q^*)}{q^*} = 1.$$

The general shape of the correlation function, as well as the stability of the fixed point at one and the existence of other fixed points, is determined by the choice of $(\phi, \sigma_w^2, \sigma_b^2)$. This can be observed in Figure 1(b), in which the correlation functions of three commonly used activation functions are plotted side by side for comparison. Figure 2(a) shows the correlation function for three different values of σ_w^2 with $\sigma_b^2 = 0.1$ and $\phi(z) = \text{htanh}(z)$ fixed. For $\sigma_w^2 = 0.63$ and $\sigma_w^2 = 1.26$ it is clear from Figure 2(a) that $\rho^* = 1$ is the only positive fixed point. Furthermore, for these two choices of σ_w^2 it is also clear, by inspection, that $\rho^* = 1$ is a stable and marginally stable fixed point respectively, and that from any initial correlation the sequence of correlations $\rho^{(l)} \to 1$ with depth. As a result, in the infinite width limit any pair of inputs, no matter how large the angle between them is, are mapped to the same point by the network asymptotically with depth. However, for $\sigma_w^2 = 0.63$ a new, stable fixed point $\rho^* < 1$ is introduced. In this case, then for any initial correlation the sequence $\rho^{(l)}$ converges to the limit $\rho^* < 1$. Therefore, in the infinite width limit any pair of inputs, no matter how small the initial angle between them is, are mapped to different points by the network asymptotically with depth. Figure 2(a) therefore illustrates the existence of two very distinct regimes for a given choice of activation function, one ordered, in which all inputs are mapped asymptotically to the same output, and the other chaotic, in which even arbitrarily small perturbations of an input lead to a different output. Figure 2(b) illustrates the limiting ρ^* throughout the (σ_w, σ_b) plane for $\phi = \text{htanh}(\cdot)$.

In order to understand and separate these two regions [7, 8] studied the slope of $R_{\phi,q^*}(\rho)$ at $\rho = 1$: for a sufficiently smooth activation function ϕ the derivative is given by

$$R'_{\phi,q^*}(\rho) = \sigma_w^2 \mathbb{E}[\phi'(U_1)\phi'(U_2)], \qquad (6)$$

the slope at $\rho = 1$ is therefore

$$\chi_1 := R'_{\phi,q^*}(1) = \sigma_w^2 \mathbb{E}[\phi'(U_1)^2]. \qquad (7)$$

The set of pairs $(\sigma_w^2, \sigma_b^2) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ such that $\chi_1 = 1$ is colloquially referred to in the literature as the edge of chaos (EOC) and in Figure 2(c) this set, or curve, is plotted for three different activation functions. In [8] it was proved that if (σ_w^2, σ_b^2) does not lie on the EOC then the sequences generated by the variance and correlation functions, (3) and (5) respectively, approach their respective fixed points asymptotically exponentially fast. As a result, the EOC curve can be interpreted as a transition boundary between order, corresponding to $\chi_1 < 1$, in which pairwise correlations converge asymptotically exponentially fast to 1 with depth, and chaos, corresponding to $\chi_1 > 1$, in which even pairwise input correlations that are arbitrarily close to 1 diverge asymptotically exponentially fast with depth. The dynamics of the preactivation correlations can also be understood and interpreted in terms of a network's sensitivity to perturbations of the input, with networks initialised in the chaotic regime carrying the risk of being overly sensitive, and those initialised in the ordered regime being insensitive. These observations are also of practical relevance for training, indeed experimental evidence [8] illustrates that initialising closer to the EOC consistently results in reduced training times.

To recap the discussion presented so far, rapid convergence in correlations to a fixed point appears to result in poor training outcomes, with the network being either highly sensitive or insensitive to perturbations of the input. To avoid asymptotically exponentially fast convergence it is necessary to choose (σ_w^2, σ_b^2) so that $\chi_1 = 1$. The specific asymptotic rate however depends on the activation function deployed. In [10], and under the assumption that initialisation is on the EOC, the asymptotic convergence rate of the sequence of correlations for different families of activation functions was analysed. A key finding of this work is that while ReLU like activation functions have asymptotic convergence $|1 - \rho^{(l)}| = \mathcal{O}(l^{-2})$, a broad class of smooth activation functions, including, tanh, elu and swish, have convergence
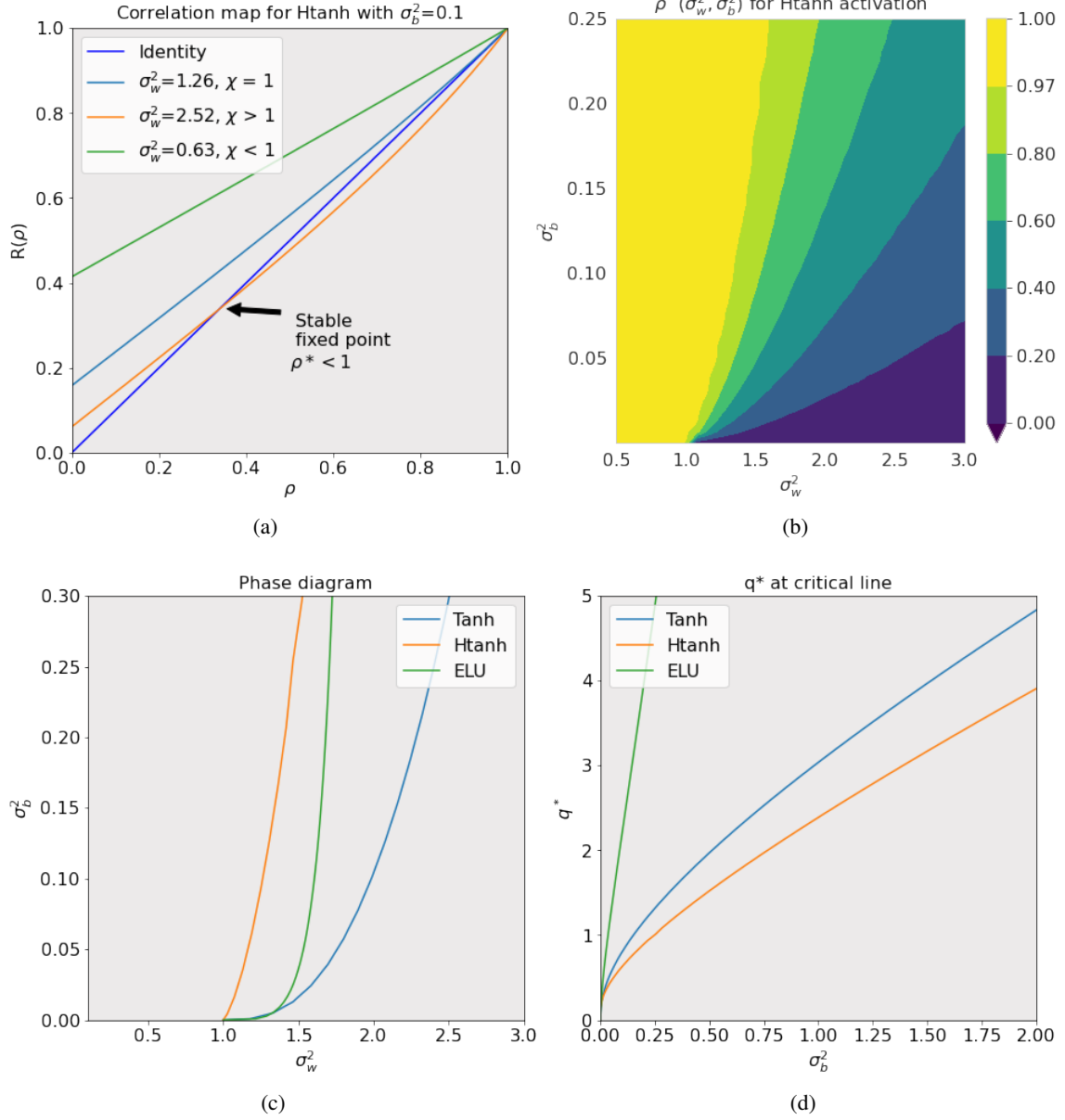
4

Figure 2: (a) Correlation map for htanh activation function, plotted on the interval $[0, 1]$ to better display fixed point behaviour (there are no fixed points in the interval $[-1, 0]$). When $\chi \leq 1$ then one is the unique fixed point of the correlation function. Furthermore, for $\chi < 1$ then one is a stable fixed point, and for $\chi = 1$ then one is a marginally stable fixed point. When $\chi > 1$ then although one is still a fixed point it is unstable, and a new stable fixed point less than one is introduced. (b) Stable fixed point ρ^* as a function of (σ_w^2, σ_b^2) for the tanh activation function. (c) The EOC critical line $\chi = 1$ for different activation functions. (d) σ_w^2 as a function of q^*, as $\sigma_b^2 \to 0$ then $q^* \to 0$.

$|1 - \rho^{(l)}| = \mathcal{O}(l^{-1})$. However, it seems highly reasonable that avoiding fast non-asymptotic, i.e., away for $\rho^* = 1$, convergence of correlations is also potentially of value. Figure 1(b) illustrates, assuming initialisation on the EOC, that certain combinations of (ϕ, σ_b^2) can result in a correlation function which is closer to the identity function. This in turn implies a slower convergence of the correlation sequence throughout the layers, rather than just at the depth limit. This idea was promoted in [10, Appendix B.2], where it was suggested that one should choose a small σ_b^2 and an activation function which is approximately linear, for example, one which is the weighted sum of a linear and nonlinear activation function.

2.2   Avoiding vanishing and exploding gradients via dynamical isometry

A key ingredient behind the success of neural networks is that their modular, compositional structure enables easy computation of the gradients of the network output with respect to the model parameters. The computation of these gradients can efficiently be performed using the backpropagation algorithm [26]. Backpropagation, also referred to as the backward pass of the network, generates a sequence of vectors $(\delta^{(l)}(\mathbf{x}))_{l=1}^{L}$ at each layer with respect to an input $\mathbf{x} \in \mathbb{R}^N$, which measure the error associated with minimising the loss $\mathcal{L}(\theta, x)$. Here θ denotes the trainable network parameters, $\theta := \bigcup_{l=1}^{L}\{\mathbf{W}^{(l)}\} \cup \{\mathbf{b}^{(l)}\}$. For a suitably smooth loss function $\mathcal{L}(\theta, x)$, the backpropagation error vector $\delta^{(l)}(\mathbf{x})$, and the gradient of $\mathcal{L}(\theta, x)$ with respect to biases $b_j^{(l)}$ and weights $w_{j,i}^{(l)}$, at each layer $l \in [L]$ can be computed using the following recurrence equations,

$$
\begin{aligned}
\delta^{(L)}(\mathbf{x}) &= \mathbf{D}^{(L)}(\mathbf{x})\nabla_{\mathbf{h}^{(L)}}\mathcal{L}(\theta, \mathbf{x}), \\
\delta^{(l)}(\mathbf{x}) &= \left(\mathbf{D}^{(l)}(\mathbf{x})\mathbf{W}^{(l+1)}\right)^{T}\delta^{(l+1)}(\mathbf{x}), \\
\frac{\partial \mathcal{L}(\theta, \mathbf{x})}{\partial b_j^{(l)}} &= \delta_j^{(l)}(\mathbf{x}), \\
\frac{\partial \mathcal{L}(\theta, \mathbf{x})}{\partial w_{j,i}^{(l)}} &= h_i^{(l-1)}(\mathbf{x})\delta_j^{(l)}(\mathbf{x}).
\end{aligned}
\tag{8}
$$

Here, for each $l \in [L]$, $\mathbf{D}^{(l)}(\mathbf{x})$ is a diagonal matrix with $D_{ii}^{(l)}(\mathbf{x}) := \phi'(z_i^{(l)}(\mathbf{x}))$ for all $i \in [N_l]$. Consequently, the weights and biases of the network can be iteratively updated via gradient descent type algorithms, or some variant thereof, in order that the forward pass improves its performance, as measured typically by averaging $\mathcal{L}(\theta, \mathbf{x})$ over batches of the training data, at executing the task in hand. We emphasise that in practice, instead of updating the parameters with respect to a single data point, gradients are calculated and averaged over batches of the training data. It is important therefore not to confuse the input-output Jacobian of the network, which we will discuss presently, with the Jacobian of the loss surface over the training data, or batches of the training data.

We now turn our attention to reviewing the findings of the literature concerning the propagation of $\delta^{(l)}$ in the backwards pass. Analogous to the recurrence relation for the variance function (3) in the forward pass, in [8] a mean field approximation for $\delta^{(l)}$ was developed. Specifically, the recurrence relation for the variance, $\tilde{q}^l := \mathbb{E}[(\delta_i^l)^2]$, of entries of the error vectors δ_i^l was shown to be

$$
\tilde{q}^l = \frac{N^{l+1}}{N^l}\tilde{q}^{l+1}\chi_1
\tag{9}
$$

with χ_1 defined as in (7). Although certain assumptions needed for this approximation are undesirable, notably the weights used during forward propagation are drawn independently from the weights used in backpropagation, empirical evidence indicates that it is nonetheless a useful model [8]. In particular, unlike the analysis of the gradients used to update the network parameters. The reappearance of χ_1 in (9) shows the benefit of selecting $(\phi, \sigma_w^2, \sigma_b^2)$ such that $\chi = 1$, as this ensures, at least in expectation, that the magnitudes of the error vectors are stable during the backward pass, i.e., they neither converge to zero or diverge.

Unfortunately, initialising on the EOC suffices only to ensure that vanishing and exploding gradients are avoided on average, and in practice does not guarantee good training performance. In particular, if, as it propagates backwards, an error vector is projected onto a smaller and smaller subspace then the directions in which the parameters in the lower layers can be updated becomes limited, thereby resulting in model degeneracy [12, 13]. In order to improve training, in [12, 13] the authors proposed that the product of Jacobians at each layer should act as an isometry on as large a subspace as possible, implying that the singular values of the input-output Jacobian at any layer should concentrate around one. As in [9, 13], we denote the input-output Jacobian of the network as $\mathbf{J} : \mathbb{R}^{N_0} \to \in \mathbb{R}^{N_0 \times N_L}$,

$$
\mathbf{J}(\mathbf{x}) = \prod_{l=1}^{L}\mathbf{D}^{(l)}(\mathbf{x})\mathbf{W}^{(l)}.
\tag{10}
$$

Comparing (10) with (8), then the Jacobian of $\mathbf{x}$ can be viewed as the transpose of the product of linear backpropagation operators used to compute the error vectors associated with $\mathbf{x}$ at each layer. In [9, 13] the authors considered two initialisation schemes, in both the biases $\mathbf{b}_i^{(l)}$ at each layer are mutually independent and identically distributed Gaussians with mean 0 and variance σ_b^2. The weight matrices at each layer however either have mutually independent identically distributed Gaussian entries with mean 0 and variance σ_w^2/N_{l-1}, or are drawn from a uniform distribution over scaled orthogonal matrices such that $(\mathbf{W}^{(l)})^T\mathbf{W}^{(l)} = \sigma_w^2\mathbf{I}$. To analyse the Jacobian of the network at an arbitrary point $\mathbf{x}$, the

authors considered again the large width setting, in which, as per the discussion in Section 2.1, the preactivations at layer l layer can be modelled as mutually independent, identically distributed, centred Gaussian random variables with variance $q^{(l)}$ [1]. For both of the initialisation schemes described, the limiting spectral density of $\mathbf{J}(\mathbf{x})$, in terms of its moment generating transform

$$M_{JJ^T}(z) = \sum_{k=1}^{\infty} \frac{m_k}{z^k}, \tag{11}$$

was computed and analysed using tools from free probability [9, 13, 27]. We refer the reader to Appendix 6 of [9] for further details. In the context of ensuring dynamical isometry, then of particular interest are the first and second moments, which can be expressed in terms of the moments μ_k of $\mathbf{D}^2$ and s_k of $(\mathbf{W}^T\mathbf{W})$. Here we drop the dependence on the layer index l by assuming that the condition $q^{(0)} = q^*$ holds true for all inputs $\mathbf{x}$. Adopting again the notation used in [9],

$$\mu_k := \int \phi'(\sqrt{q^* z})^{2k} d\gamma(z) = \mathbb{E}[\phi'(\sqrt{q^*})Z)^{2k}],$$
$$m_1 := (\sigma_w^2 \mu_1)^L = (\chi_1)^L, \tag{12}$$
$$m_2 := (\chi_1)^{2L} L\Big(\frac{\mu_1}{\mu_2} + \frac{1}{L} - 1 - s_1\Big).$$

It is evident from (12) that the mean squared singular value m_1 of the Jacobian either exponentially explodes or vanishes unless the network is initialised on the EOC. However, although initialisation on the EOC makes m_1 independent of depth, the variance

$$\sigma_{JJ^T}^2 = m_2 - m_1^2 = L\Big(\frac{\mu_2}{\mu_1^2} - 1 - s_1\Big) \tag{13}$$

grows linearly with depth. As a result, under the limiting width assumption then for an arbitrary, normalised input vector initialisation on the EOC may still result in an increasingly ill-conditioned Jacobian with depth L. Indeed, we require that $m_1 = 1$ and $\sigma_{JJ^T}^2 \approx 0$ so as to at least approximately achieve dynamical isometry. Equation (13) shows that using a Gaussian initialisation scheme, corresponding to $s_1 = -1$, results in a linear growth in $\sigma_{JJ^T}^2$ with depth regardless of the activation function used. As a result, deep, dense feed-forward networks cannot achieve dynamical isometry using Gaussian initialisation. In contrast, orthogonal initialisation, corresponding to $s_1 = 0$, can in theory be used to ensure that $\sigma_{JJ^T}^2$ is arbitrarily close to one by controlling the moment ratio μ_2/μ_1^2. However, the ability to control the moment ratio depends on the activation function deployed: in the case of ReLU for instance, whose EOC is a singleton, the moment ratio is a constant and hence once again dynamical isometry cannot be achieved. As highlighted in [9], for certain activation functions, e.g., erf and tanh, which have a non-singleton EOC as illustrated by Figure 2(c), then the moment ratio can be reduced by shrinking q^*, which in turn can be achieved by reducing σ_b^2, see Figure 2 (d). Finally we remark that similar analyses were conducted in [28] in the context of finite width deep ReLU networks.

2.3   Contribution: activations which approximately preserve correlations and achieve dynamical isometry

To recap, in Section 2.1 we discussed how, in order to achieve deep information propagation and avoid the network being either highly sensitive or insensitive to perturbations of the input, it is important to avoid a rapid rate of convergence of the correlation. To this end it is necessary to choose $(\phi, \sigma_w^2, \sigma_b^2)$ so that $\chi_1 = 1$ and the associated correlation function is close to identity. In Section 2.2 we discussed the problem of vanishing and exploding gradients. Prior works suggest that this problem can be avoided if the singular values of the input-output Jacobian at each layer concentrate around 1. The condition $\chi_1 = 1$ is equivalent to ensuring that the mean of the input-output Jacobian's spectrum at any layer is one. Therefore, if the variance $\sigma_{JJ^T}^2$ of the spectrum, defined in (13), is close to zero, then the spectrum of the input-output Jacobian is guaranteed to concentrate around one. Furthermore, with orthogonal initialisation on the EOC then so long as $\mu_2/\mu_1^2 \approx 1$ then $\sigma_{JJ^T}^2 \approx 0$.

In this paper we present principles for choosing the activation function ϕ, see Definition 2.1, and an additional linear scaling parameter which simultaneously enables uniform convergence of the correlation function $R_{\phi,q^*}(\rho)$ to the identity map, and convergence of the variance $\sigma_{JJ^T}^2$ of the input-output Jacobian's spectrum to zero, without requiring σ_b^2 to shrink towards zero. For further details we refer the reader to Theorem 2.1. From prior work it was unclear how one could achieve this. For example, initialising on the EOC, using orthogonal initialisation and deploying the modulus activation function achieves perfect dynamical isometry, but results in rapid convergence of correlations in the forward pass, leading to poor training outcomes. Treated separately, a key theme that emerges in the prior works towards achieving both goals is that of reducing σ_b^2 in order to make q^* small. As mentioned briefly in section 2.1, in

---

[1] Note that (3) was derived in the context of Gaussian initialisation, for now we also assume it holds true also for the Orthogonal case.

[10] the authors highlight that, for a smooth class of activation functions, taking this action can reduce the gap between the correlation map and the identity. The authors also provide a rule for selecting σ_b^2 just small enough so that pairwise correlations avoid being within some $\epsilon > 0$ of one at the output layer. Furthermore, experiments conducted in [10] indicate that selecting the variance hyperparameters in this manner accelerates training in practice. However, the authors also highlight that as the depth of the network increases this rule still requires $\sigma_b^2 \to 0$. Likewise, and as mentioned in section 2.2, in [9] the authors show, for the htanh and erf activation functions, that reducing q^* by shrinking σ_b^2 results in the moment ratio converging towards one. However, this solution is not entirely satisfactory, first if the expected length of the activations goes to zero then asymptotically the network will on average annihilate inputs and fail to pass information to the output. Second, experimental evidence indicates that a small, but not overly small σ_b^2, gives the best results in practice, which we demonstrate in Section 4. Third, this solution relies only on the choice of σ_b^2 and ignores the role and design of ϕ. Our work seeks to address this issue by achieving both goals through the design of ϕ, even when σ_b^2 and q^* are fixed away from zero.

In light of the above, and assuming σ_b^2 is fixed and that σ_w^2 is chosen in order that $\chi_1 = 1$, or goal is to design or choose ϕ in order that both $\max_{\rho \in [-1,1]} |R_{\phi,q^*}(\rho) - \rho| \approx 0$ and $|\mu_2/\mu_1^2 - 1| \approx 0$. As highlighted in [10], linear activation functions are advantageous from the perspective of slowing the convergence of the pairwise correlations in the forward pass. Likewise, [27, 12] highlight that linear networks with orthogonal initialisation achieve perfect dynamical isometry. However, linear activations are not a suitable option as a linear network can only represent linear functions. To preserve the rich approximation capabilities of nonlinear networks we analyse a set of activation functions characterised by being odd, bounded, Lipschitz continuous and linear around the origin. We refer to activation functions of this type, defined in Definition 2.1, as *scaled-bounded activations*.

Definition 2.1 (**scaled-bounded activations**). *We refer to the set of activation functions $\phi : \mathbb{R} \to \mathbb{R}$ which satisfy the following properties as scaled-bounded activations.*

1. *Continuous.*

2. *Odd, meaning that $\phi(z) = -\phi(-z)$ for all $z \in \mathbb{R}$.*

3. *Linear around the origin and bounded: in particular there exists $a, k \in \mathbb{R}_{>0}$ such that $\phi(z) = kz$ for all $z \in [-a, a]$ and $\phi(z) \leq ak$ for all $z \in \mathbb{R}$.*

4. *Twice differentiable at all points $z \in \mathbb{R}\backslash\mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}$ is a finite set. Furthermore $|\phi'(z)| \leq k$ for all $z \in \mathbb{R}\backslash\mathcal{D}$.*

As illustrated in Figure 3, the structure of scaled-bounded activation functions outside of their linear region $[-a, a]$ can vary substantially. We note that for any scaled-bounded activation ϕ there exists $a, k \in \mathbb{R}_{>0}$ such that $|\phi(z)| \leq |\text{Shtanh}_{a,k}(z)|$, where

$$\text{Shtanh}_{a,k}(z) = \begin{cases} zk, & |z| < a \\ ak, & |z| \geq a. \end{cases} \tag{14}$$

Here Shtanh stands for scaled-bounded hard tanh, and we further note that $\text{Shtanh}_{a,k}(\cdot)$ is a scaled-bounded activation for any $a, k \in \mathbb{R}_{>0}$. Our main contribution is Theorem 2.1, which asserts, for scaled-bounded activations, that the correlation function R_{ϕ,q^*} can be made arbitrarily close to the identity on $[0, 1]$, and the moment ratio μ_2/μ_1^2 arbitrarily close to one, by choosing σ_b^2/a^2 to be small.

Theorem 2.1. *Let ϕ be a scaled-bounded activation, see Definition 2.1, $\sigma_b^2 > 0$ and suppose that*

$$\chi_1 := \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q^*}Z)^2] = 1,$$

where $q^* > 0$ is a fixed point of the associated variance function V_ϕ. In addition, assume that all inputs $\mathbf{x}$ are normalised so that $||\mathbf{x}||_2^2 = q^*$. With $y := \frac{\sigma_b^2}{a^2}$ and Λ defined as in Lemma 3.4, then

$$\max_{\rho \in [0,1]} |R_\phi(\rho) - \rho| < \left(\frac{8}{\pi}\right)^{1/3} y^{1/3} \tag{15}$$

and

$$\left|\frac{\mu_2}{\mu_1^2} - 1\right| \leq erf\left(\frac{\Lambda(y)}{\sqrt{2}}\right)^{-2} - 1. \tag{16}$$

We emphasise that as $y := \sigma_b^2/a^2 \to 0$ then both

$$\max_{\rho \in [0,1]} |R_{\phi,q^*}(\rho) - \rho|, \left|\mu_2/\mu_1^2 - 1\right| \to 0.$$
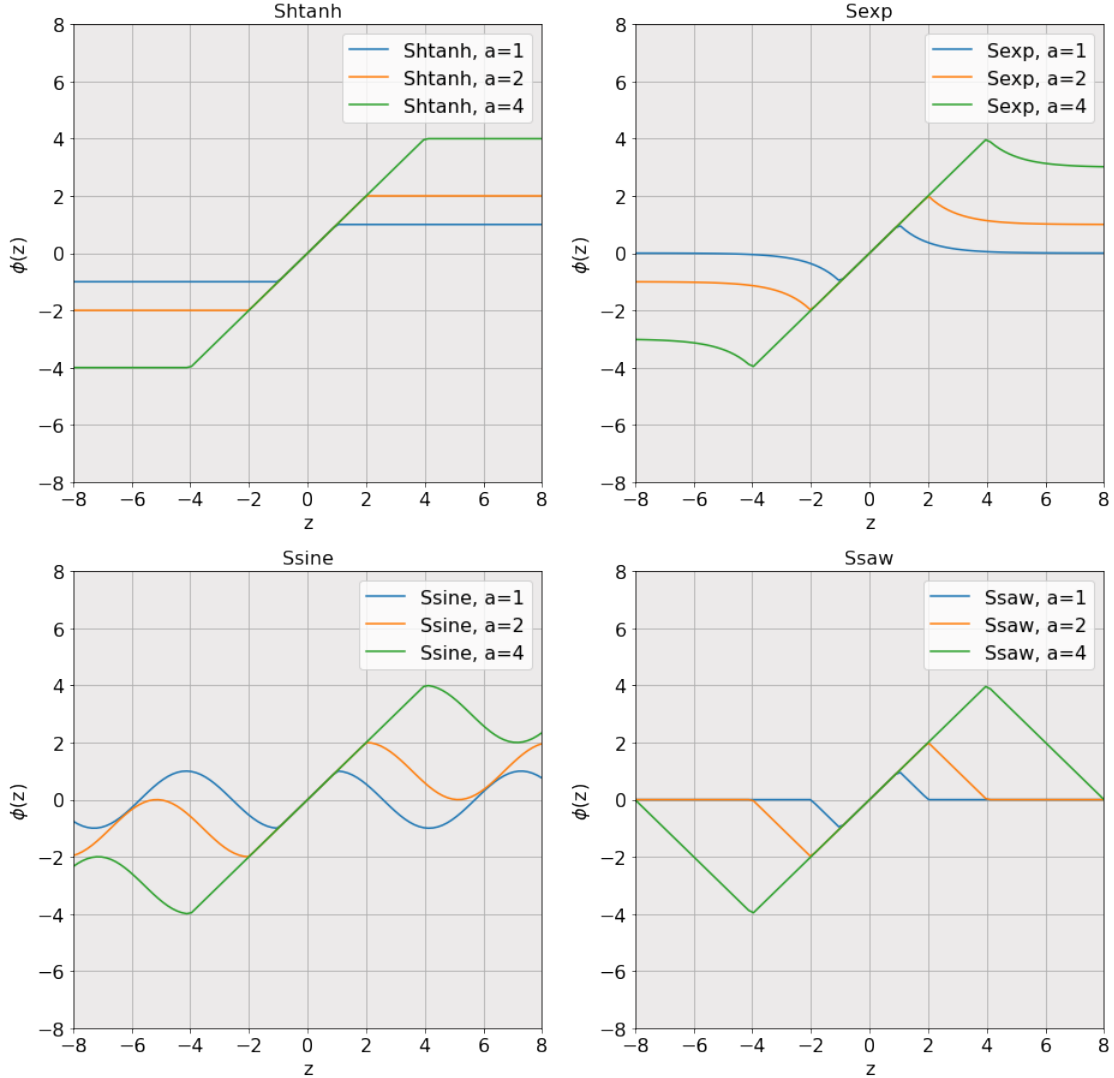
Figure 3: examples of scaled-bounded activation functions with $k = 1$ and $a \in \{1, 2, 4\}$. The prefix S refers to the scaling of the linear region via the parameter a.

We first remark that the uniform bound on the correlation provided by Theorem 2.1 is only for nonnegative correlations. However, Figure 6 indicates that letting $\sigma_b^2/a^2 \to 0$ results $\max_{\rho \in [-1,1]} |R_{\phi,q^*}(\rho) - \rho| \to 0$. The key takeaway of Theorem 2.1 for practitioners is that an improved initialisation can be achieved by using an activation function which has a sufficiently large linear region spanning either side of the origin. The size of this linear region, a, needs to be selected on the basis of both the bias variance hyperparameter σ_b^2 and the depth of the network. The deeper the network or the larger σ_b^2 is, the larger a needs to be in order to avoid both convergence of the correlations and achieve approximate dynamical isometry. In practice, as per Figure 5, modest increases in a typically suffice to capture many of potential benefits at initialisation. We emphasise that there is a tension between the linear region being large enough to achieve a good initialisation, while being small enough so that the expressivity of the network is not reduced. In particular, if the network is initialised so that it acts as linear transform on the input data, and if the optimiser becomes trapped in a local minimum close to the initialisation point, then the network may continue to act as a linear transform throughout training.

9

3 Analysis of scaled-bounded activation functions

3.1 Derivation of Theorem 2.1

We start our analysis by revisiting what it means to initialise on the EOC: the condition in (7) presupposes the existence of a fixed point q^* of V_ϕ. To resolve this matter we introduce and study the fixed points of a related function, W. Before presenting this analysis we introduce the following slight abuse of notation,

$$\phi'(z) := \frac{1}{2}\left(\frac{d\phi}{dz}(z^-) + \frac{d\phi}{dz}(z^+)\right) \tag{17}$$

where $\frac{d\phi}{dz}(z^-)$ and $\frac{d\phi}{dz}(z^+)$ are the left and right derivatives of ϕ at $z \in \mathbb{R}$ respectively. This is convenient for what follows as it allows us to define a notion of a derivative of a scaled-bounded activation ϕ over the whole of the domain: we remark, by Definition 2.1, that the true derivative of ϕ, and $\phi'(\cdot)$, as defined above, are equal almost everywhere.

Lemma 3.1. *Let ϕ be a scaled-bounded activation, see Definition 2.1, and $\sigma_b^2 > 0$. Define*

$$W_\phi(q) := \frac{\mathbb{E}[\phi(\sqrt{q}Z)^2]}{\mathbb{E}[\phi'(\sqrt{q}Z)^2]} + \sigma_b^2 \tag{18}$$

for all $q \in \mathbb{R}_{\geq 0}$. Then $W_\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and W has a fixed point $q^* > 0$.

A proof of Lemma 3.1 is provided in Appendix B.1. As a consequence of Lemma 3.1, we are able to make the following claims concerning the existence of fixed points of the variance function V_ϕ for any scaled-bounded activation ϕ.

Corollary 3.1.1. *Let ϕ be a scaled-bounded activation, see Definition 2.1, $\sigma_b^2 > 0$ and suppose*

$$\chi_1 := \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q^*}Z)^2] = 1, \tag{19}$$

where $q^* > 0$ is a fixed point of W_ϕ, defined in (18). Then q^* is a fixed point of the associated variance function V_ϕ, defined in (3).

A proof of corollary 3.1.1 is included in Appendix B.2 for completeness. The key takeaway of Corollary 3.1.1 is that for any scaled-bounded activation there exists a fixed point q^* of V_ϕ satisfying $q^* > 0$. In fact, as $V_\phi(0) = \sigma_b^2 > 0$ then any fixed point of V_ϕ is greater than 0. We emphasise that such analysis is necessary as in general care is required when making assumptions concerning the existence of fixed points. For example, the variance function for ReLU like functions do not have any fixed points unless $\sigma_b^2 = 0$.

To simplify the analysis we will now proceed under the assumption that the input data is normalised to have euclidean length q^* and that $q^{(l)} = q^*$ for all $l \in [L]$. If one where interested in initialising in this manner in practice, it might also be of interest to explore guarantees concerning the ease with which a stable fixed point of V_ϕ can be computed: in particular if the fixed point is not attractive then numerical precision issues might mean that the sequence $q^{(l)}$ diverges from q^*. We do not pursue this line of inquiry and instead leave it as potential future work. Now that we have identified the existence of fixed points of the variance functions of scaled-bounded activations, it is possible to study the associated correlation functions.

Lemma 3.2. *Let ϕ be a scaled-bounded activation, see Definition 2.1, $\sigma_b^2 > 0$ and suppose*

$$\chi_1 := \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q^*}Z)^2] = 1, \tag{20}$$

where $q^* > 0$ is a fixed point of W_ϕ. In addition, assume that all inputs $\mathbf{x}$ are normalised so that $||\mathbf{x}||_2^2 = q^*$. Then the associated correlation map $R_{\phi,q*}$, defined in (5), is fixed at each layer $l \in [L]$, satisfies $R_{\phi,q*} : [-1, 1] \to [-1, 1]$ and is differentiable with

$$R'_{\phi,q^*}(\rho) = \sigma_w^2 \mathbb{E}[\phi'(U_1)\phi'(U_2)] \tag{21}$$

for all input correlations $\rho \in (-1, 1)$.

A proof of Lemma 3.2 is given in Appendix B.3. This expression for the correlation is equivalent to that given in [7], however we emphasise, due to the fact that ϕ is not necessarily continuously differentiable, that significantly more care is required to derive it. Indeed, if ϕ were not odd then this equivalence would not hold and additional terms would appear, potentially complicating the analysis downstream. We are now ready to present a key lemma, which provides a tight uniform bound on the interval $[0, 1]$ between the correlation functions associated with scaled-bounded activations and the identity function.

Lemma 3.3. *Under the same conditions and assumptions as in Lemma 3.2, it holds that*

$$\max_{\rho \in [0,1]} |R_{\phi,q^*}(\rho) - \rho| = \frac{\sigma_b^2}{q^*}. \tag{22}$$

A proof of Lemma 3.3 is provided in Appendix B.4. The final lemma we present before proving Theorem 2.1 concerns the relationship between a, the size of the linear region, and q^*, the fixed point of the variance function. This lemma provides lower and upper bounds on the ratio $\frac{a}{\sqrt{q^*}}$ for all scaled-bounded activations as a function of a and σ_b^2. The fact that this is possible indicates that the particular shape of the tails of a scaled-bounded activation do not play a key role in achieving a good initialisation. In Figure 4 we plot these bounds, as well as example $a/\sqrt{q^*}$ ratios, for a number of activation functions as a function of a.
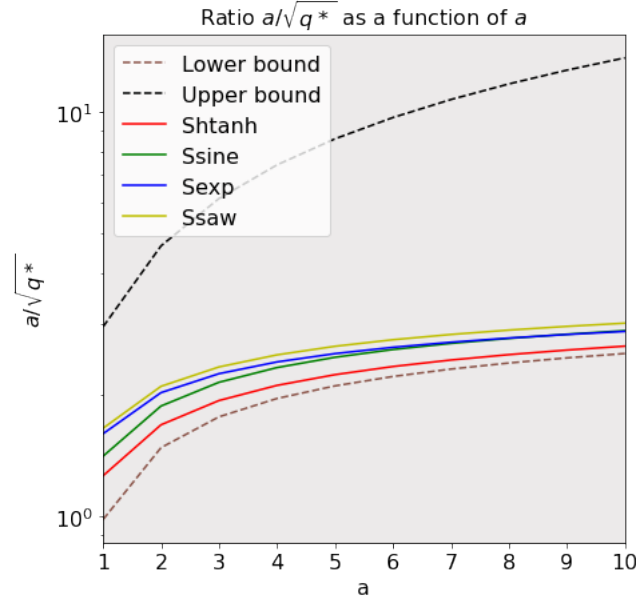

Figure 4: Bounds on $\frac{a}{\sqrt{q^*}}$ given in (23) vs. $\frac{a}{\sqrt{q^*}}$, computed numerically for the same scaled-bounded activations visualised in Figure 3.

Lemma 3.4. *Under the same conditions and assumptions as in Lemma 3.2, and defining $y := \frac{\sigma_b^2}{a^2}$, then*

$$\Lambda(y) < \frac{a}{\sqrt{q^*}} < \left(\frac{8}{\pi}\right)^{1/6} y^{-1/3}, \tag{23}$$

where $\Lambda(y)$ is defined as

$$\left(W_0\left(\frac{2}{\pi}\left(\sqrt{\frac{8}{\pi}}\exp\left(-\frac{\left(\sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}\right)^2}{2}\right)\left(\frac{1}{\left(\sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}\right)} + \left(\sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}\right)\right)\right)^{-2}\right)\right)^{1/2}$$

and W_0 denotes the principal branch of the Lambert W function.

A proof of Lemma 3.4 is given in Appendix B.5. We note that while the upper bound in Lemma 3.4 is easy to interpret, the lower bound is not immediately interpretable. However, this lower bound still allows us to compute a numerical lower bound for any scaled-bounded activation as per Figure 4. In terms of asymptotic behaviour it is easy to check that as $y \to 0$ then $y^{-1/3}, \Lambda(y) \to \infty$, and as $y \to \infty$ then $y^{-1/3}, \Lambda(y) \to 0$. We also observe from Figure 4 that, at least empirically, the lower bound seems to be far tighter: we leave it as potential future work to investigate deriving a tighter upper bound. We are now ready to prove Theorem 2.1.

Proof. To derive the inequality concerning the correlation function given (15) we use the upper bound

$$\frac{a}{\sqrt{q^*}} < \left(\frac{8}{\pi}\right)^{1/6} \frac{a^{2/3}}{\sigma_b^{2/3}}$$

derived in Lemma 3.4. Squaring both sides, dividing by a^2 and multiplying by σ_b^2 then

$$\frac{\sigma_b^2}{q^*} < \left(\frac{8}{\pi}\right)^{1/3} \frac{\sigma_b^{2/3}}{a^{2/3}}.$$

To conclude we apply Lemma 3.3. We now turn our attention to proving the inequality concerning the moment ratio provided in equation 16. Analysing the lth moment,

$$\mu_l := \mathbb{E}[\phi'(\sqrt{q^*}Z)^{2l}]$$
$$= 2\left(k^{2l}\int_0^{a/\sqrt{q^*}} d\gamma(z) + \int_{a/\sqrt{q^*}}^\infty \phi'(\sqrt{q^*}z)^{2l}d\gamma(z)\right).$$

By assumption $|\phi'(z)| \leq k$, hence we can bound this quantity as

$$k^{2l}\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) \leq \mu_l \leq k^{2l},$$

where the lower bound arises simply by zeroing the second integral term. It therefore follows that

$$\frac{k^4\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)}{k^4} \leq \frac{\mu_2}{\mu_1^2} \leq \frac{k^4}{k^4\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^2},$$

$$\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) \leq \frac{\mu_2}{\mu_1^2} \leq \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^{-2}$$

Observe that $\mu_2/\mu_1^2 - 1 \leq \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^{-2} - 1$ and $1 - \mu_2/\mu_1^2 \leq 1 - \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)$. We now prove, by contradiction, that $\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^{-2} - 1 \geq 1 - \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)$. Indeed, assume that $\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^{-2} - 1 < 1 - \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)$, then

$$1 - \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^2 < \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)\left(\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) - \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^2\right)$$

$$\leq \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)\left(1 - \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^2\right).$$

As $\mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) < 1$ for $\frac{a}{\sqrt{2q^*}} < \infty$ then this is a contradiction. Therefore

$$\left|\frac{\mu_2}{\mu_1^2} - 1\right| \leq \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)^{-2} - 1.$$

As erf is monotonically increasing it suffices to lower bound the quantity $\frac{a}{\sqrt{q^*}}$. The result claimed is then recovered by applying the lower bound on $\frac{a}{\sqrt{q^*}}$ derived in Lemma 3.4. □

3.2 Discussion and practical takeaways

Equation (15) implies, by choosing a sufficiently large relative to σ_b^2, that problems arising as a result of limits on the depth of information propagation as well as network sensitivity can be mitigated without the need for $q^*, \sigma_b^2 \to 0$. Equation (16) likewise implies, as long as orthogonal initialisation is used, that increasing a also mitigates the problem of model degeneracy, again without the need for $q^*, \sigma_b^2 \to 0$. Our numerics support these conclusions: in Figure 5 the bounds in (15) and (16) are plotted numerically with σ_b^2 fixed, and converge to 0 as a increases. Likewise, Figure 6 shows that increasing a moves R_{ϕ,q^*} closer to the identity. We remark that (3) and (5) where derived in the context of
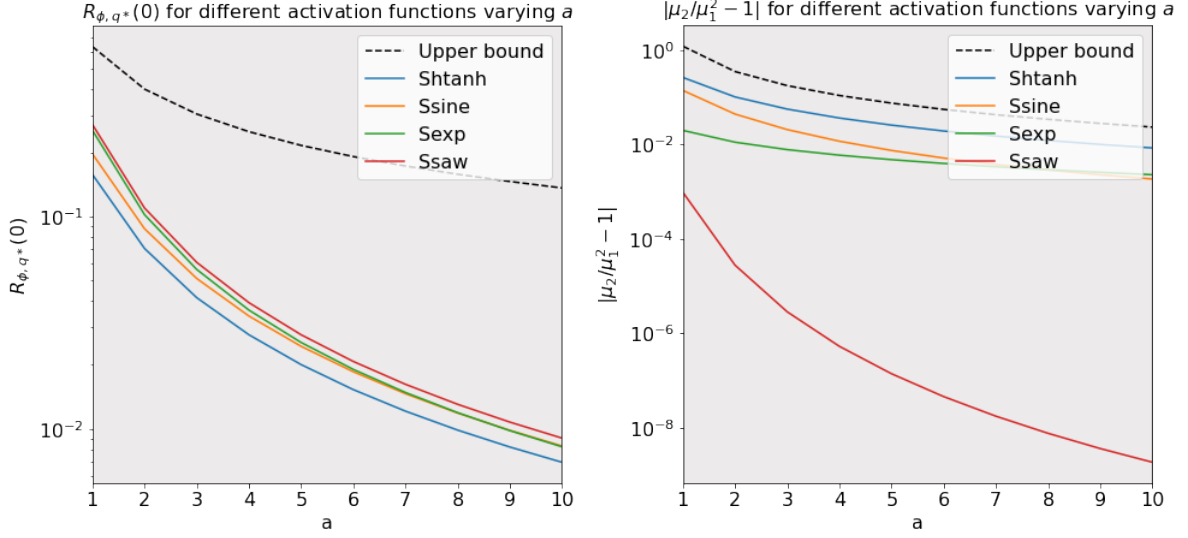
Figure 5: The left and right plots display the correlation and moment ratio bounds, given in Equations (15) and (16) respectively, vs. the equivalent numerically computed quantities for a variety of different activation functions $\phi \in \Omega$, as shown in Figure 3. We note that the right hand curve for hard saw is accurate only up to an a of 6 due to numerical precision issues arising from numerical integration steps.
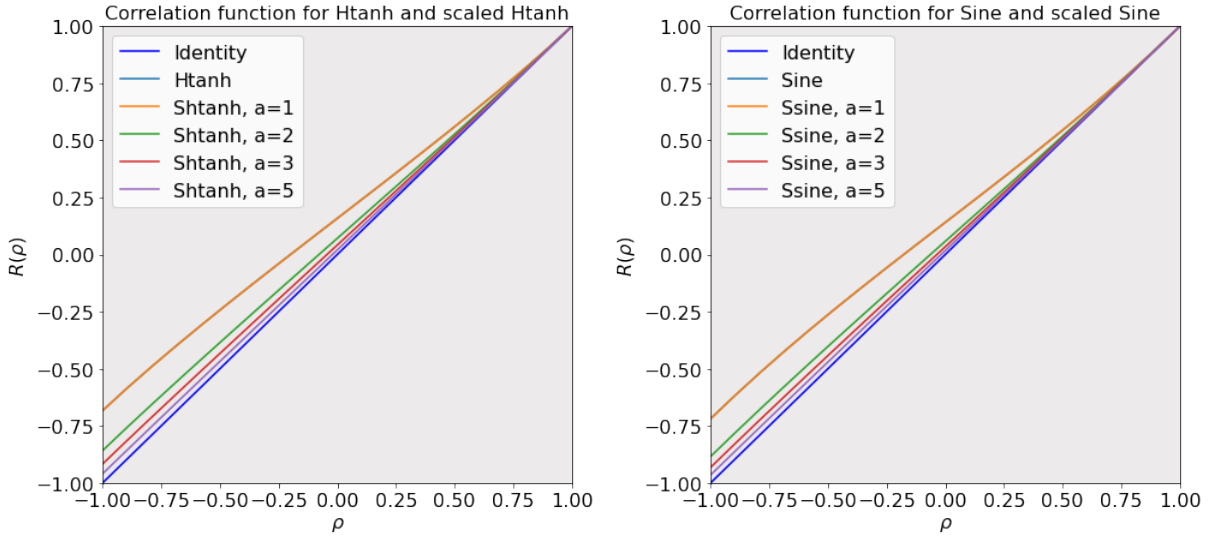
Figure 6: impact of scale parameter a on correlation function of scaled (or adapted) htanh (left-hand plot) and sinusoid (right-hand plot).

Gaussian initialisation. As a result one might question whether (15) is immediately applicable to orthogonally initialised networks. Indeed, although referenced to in [9, 13], the correspondence between orthogonally initialised networks and Gaussian processes is to our knowledge yet to be rigorously established. However, it seems highly likely that the same correspondence holds, an assertion supported both by empirical observation and the fact that large random orthogonal matrices are well approximated by Gaussian matrices (see e.g., [29]). We defer a detailed study of this correspondence to later work. We also note that the uniform bound we provide holds for non-negative correlations only, this is due to the fact that the upper bound on this interval, σ_b^2/q^* as per Lemma 3.3, is relatively easy to analyse. However, we suspect for scaled-bounded activations that letting $\sigma_b^2/a^2 \to 0$ results in a uniform convergence of the correlation function to the identity for $\rho \in [-1, 1]$. This hypothesis is supported by Figure 6.

We conclude this section by considering the relevance and importance of each of the four properties of scaled-bounded activation functions in regard to achieving a good initialisation in practice. Property 1), ϕ being continuous, does not seem controversial, indeed most of the commonly used activation functions are continuous. Furthermore, it seems reasonable that continuous activation functions result in a loss landscape that is easier for an optimiser to navigate, compared with that induced by non-continuous ones. Property 2), ϕ being odd, is beneficial in theory as it removes additional constant terms from (22) in Lemma 3.3. However, as per the proof of Lemma 3.3, it can be observed that these terms will decay exponentially fast as a grows due to the fact that the locations of the points in $\mathcal{D}$ all have magnitude at least a. As for property 3), while the existence of a linear region around the origin is critical to our results, we suspect that boundedness is more an artefact of our proof than a necessity in practice. Assuming that ϕ is bounded simplifies certain pieces of analysis, and crucially allows us to formulate upper and lower bounds on $V_\phi(q)$, needed for the proof of Lemma 3.4. We hypothesise that this condition could be relaxed to $|\phi(z)| < |z|$. In regard to property 4) and contrary to the conclusion one might be inclined to draw from [10], ϕ being non-differentiable demonstrates that smoothness is not necessary for a good initialisation. Finally, the bound on the derivative of ϕ was introduced to allow us to to derive bounds on V_ϕ.

4 Experiments

4.1 Experimental setup

Across all experiments we train networks with depths $L \in \{20, 50, 100, 200\}$, with each layer having a fixed width $N = 400$, for 100 epochs on CIFAR-10. The results are averaged over 10 trials. Variance hyperparameters (σ_w^2, σ_b^2) are selected to lie on the EOC, with $\sigma_b^2 \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and σ_w^2 computed in in order that $\chi_1 = 1$, see (7). The parameters of the network are initialised according to the following two schemes.

- **Gaussian initialisation:** all parameters are drawn mutually independent of one another. The weights in each layer are identically distributed with $w_{i,j}^{(l)} \sim \mathcal{N}(0, \sigma_w^2/N)$. The biases are all identically distributed with $b_{i,j}^{(l)} \sim \mathcal{N}(0, \sigma_b^2)$.

- **Orthogonal initialisation:** all weight matrices and bias parameters are drawn mutually independent of one another. The weight matrix at each layer is drawn according to the Haar measure, i.e., uniformly, over the orthogonal group of $N \times N$ matrices such that $(\mathbf{W}^{(l)})^T \mathbf{W}^{(l)} = \sigma_w^2 \mathbf{I}_N$. The biases are all identically distributed with $b_{i,j}^{(l)} \sim \mathcal{N}(0, \sigma_b^2)$.
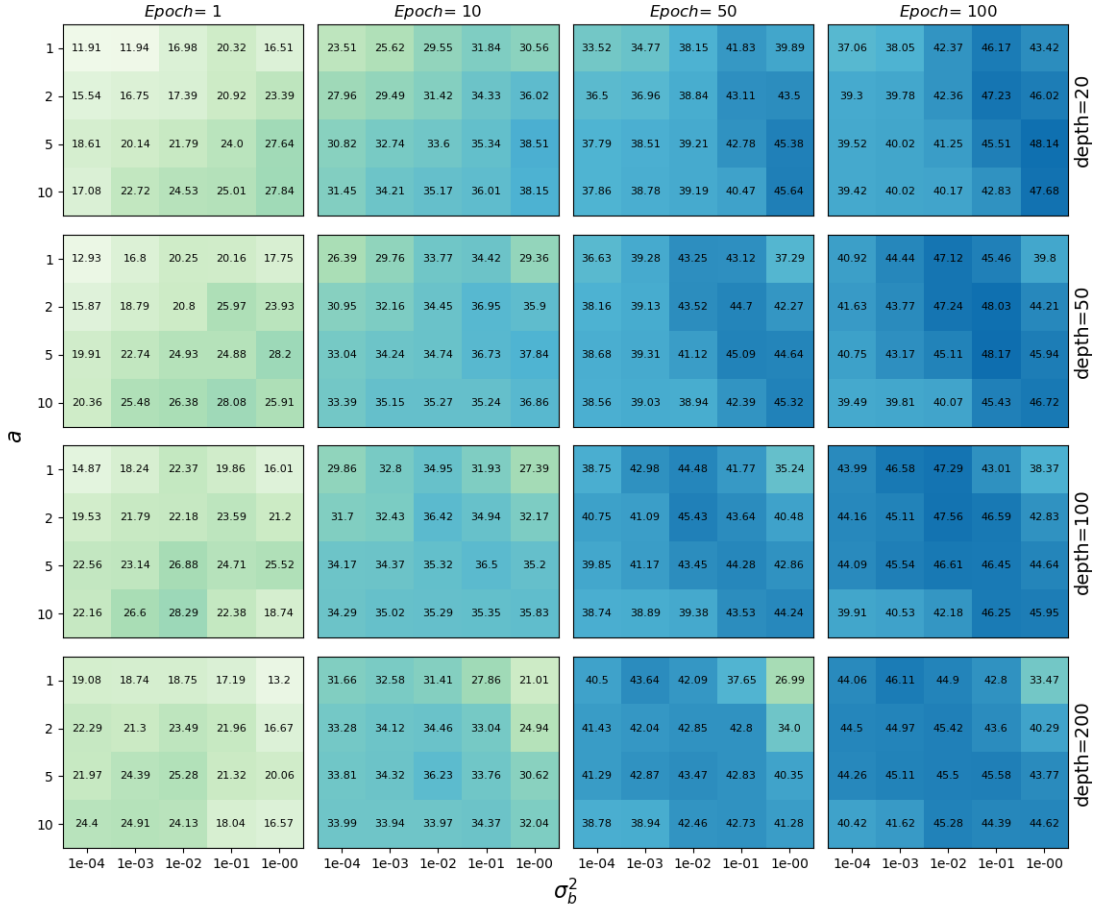
Finally, optimisation is performed using SGD with a batch size of 64. A learning rate of 10^{-4} was found to be provide good results across all experiments.

4.2 Experimental validation of the results of Section 3

In order to test and validate the results of Section 3, we deploy the scaled Shtanh activation function, defined in (14), with $a \in \{1, 2, 5, 10\}$ and $k = 1$, and measure the test accuracy of the networks described in Section 4.1 at various stages of training. The results of these experiments are summarised in the heatmaps provided in Figure 7. As per the implications of Theorem 2.1, it is clear from Figure 7 that as σ_b^2 increases a larger but not necessarily maximal value of a gives the best results. This is illustrated by the ridges, indicating the highest test accuracy, running from top left to bottom right in the heatmap subplots. This therefore highlights the trade-off between activation functions which are linear enough to allow for a good initialisation, while not being overly linear that the network loses approximation power. We emphasise that these results are not unique to Shtanh and that the same conclusions can be drawn for other scaled-bounded activation functions [2]. As established in prior works, we also observe an advantage in using orthogonal over Gaussian initialisation. However, we also note that this performance gap is relatively small. We leave it to future work to investigate the impact of optimising over, or regularising with respect to, the set of orthogonal weight matrices so that orthogonality is preserved, at least approximately, throughout training.

We also provide some very preliminary results concerning a new training protocol, in which a scaled-bounded activation function is deployed and the value of a is decreased slowly over the first few epochs. To motivate this, observe in Figure 8 that, as per our theory, a large value of a always gives the best performance at initialisation. However, for reasons already discussed, an overly large value of a has a negative impact on training long term. This strategy can therefore be interpreted as computing a good initialisation for the final network. The purple curve in Figure 8 displays the outcome of this strategy, which appears promising over a wide range of depths and hyperparameter choices.

---

[2]Implementation and additional results can be found at `https://github.com/Cross-Caps/AFLI`

Epoch = 1, depth = 20

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	11.91	11.94	16.98	20.32	16.51
2	15.54	16.75	17.39	20.92	23.39
5	18.61	20.14	21.79	24.0	27.64
10	17.08	22.72	24.53	25.01	27.84

Epoch = 10, depth = 20

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	23.51	25.62	29.55	31.84	30.56
2	27.96	29.49	31.42	34.33	36.02
5	30.82	32.74	33.6	35.34	38.51
10	31.45	34.21	35.17	36.01	38.15

Epoch = 50, depth = 20

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	33.52	34.77	38.15	41.83	39.89
2	36.5	36.96	38.84	43.11	43.5
5	37.79	38.51	39.21	42.78	45.38
10	37.86	38.78	39.19	40.47	45.64

Epoch = 100, depth = 20

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	37.06	38.05	42.37	46.17	43.42
2	39.3	39.78	42.36	47.23	46.02
5	39.52	40.02	41.25	45.51	48.14
10	39.42	40.02	40.17	42.83	47.68

Epoch = 1, depth = 50

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	12.93	16.8	20.25	20.16	17.75
2	15.87	18.79	20.8	25.97	23.93
5	19.91	22.74	24.93	24.88	28.2
10	20.36	25.48	26.38	28.08	25.91

Epoch = 10, depth = 50

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	26.39	29.76	33.77	34.42	29.36
2	30.95	32.16	34.45	36.95	35.9
5	33.04	34.24	34.74	36.73	37.84
10	33.39	35.15	35.27	35.24	36.86

Epoch = 50, depth = 50

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	36.63	39.28	43.25	43.12	37.29
2	38.16	39.13	43.52	44.7	42.27
5	38.68	39.31	41.12	45.09	44.64
10	38.56	39.03	38.94	42.39	45.32

Epoch = 100, depth = 50

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	40.92	44.44	47.12	45.46	39.8
2	41.63	43.77	47.24	48.03	44.21
5	40.75	43.17	45.11	48.17	45.94
10	39.49	39.81	40.07	45.43	46.72

Epoch = 1, depth = 100

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	14.87	18.24	22.37	19.86	16.01
2	19.53	21.79	22.18	23.59	21.2
5	22.56	23.14	26.88	24.71	25.52
10	22.16	26.6	28.29	22.38	18.74

Epoch = 10, depth = 100

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	29.86	32.8	34.95	31.93	27.39
2	31.7	32.43	36.42	34.94	32.17
5	34.17	34.37	35.32	36.5	35.2
10	34.29	35.02	35.29	35.35	35.83

Epoch = 50, depth = 100

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	38.75	42.98	44.48	41.77	35.24
2	40.75	41.09	45.43	43.64	40.48
5	39.85	41.17	43.45	44.28	42.86
10	38.74	38.89	39.38	43.53	44.24

Epoch = 100, depth = 100

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	43.99	46.58	47.29	43.01	38.37
2	44.16	45.11	47.56	46.59	42.83
5	44.09	45.54	46.61	46.45	44.64
10	39.91	40.53	42.18	46.25	45.95

Epoch = 1, depth = 200

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	19.08	18.74	18.75	17.19	13.2
2	22.29	21.3	23.49	21.96	16.67
5	21.97	24.39	25.28	21.32	20.06
10	24.4	24.91	24.13	18.04	16.57

Epoch = 10, depth = 200

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	31.66	32.58	31.41	27.86	21.01
2	33.28	34.12	34.46	33.04	24.94
5	33.81	34.32	36.23	33.76	30.62
10	33.99	33.94	33.97	34.37	32.04

Epoch = 50, depth = 200

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	40.5	43.64	42.09	37.65	26.99
2	41.43	42.04	42.85	42.8	34.0
5	41.29	42.87	43.47	42.83	40.35
10	38.78	38.94	42.46	42.73	41.28

Epoch = 100, depth = 200

a	1e-04	1e-03	1e-02	1e-01	1e-00
1	44.06	46.11	44.9	42.8	33.47
2	44.5	44.97	45.42	43.6	40.29
5	44.26	45.11	45.5	45.58	43.77
10	40.42	41.62	45.28	44.39	44.62

The horizontal axis is σ_b^2 and the vertical axis is a.

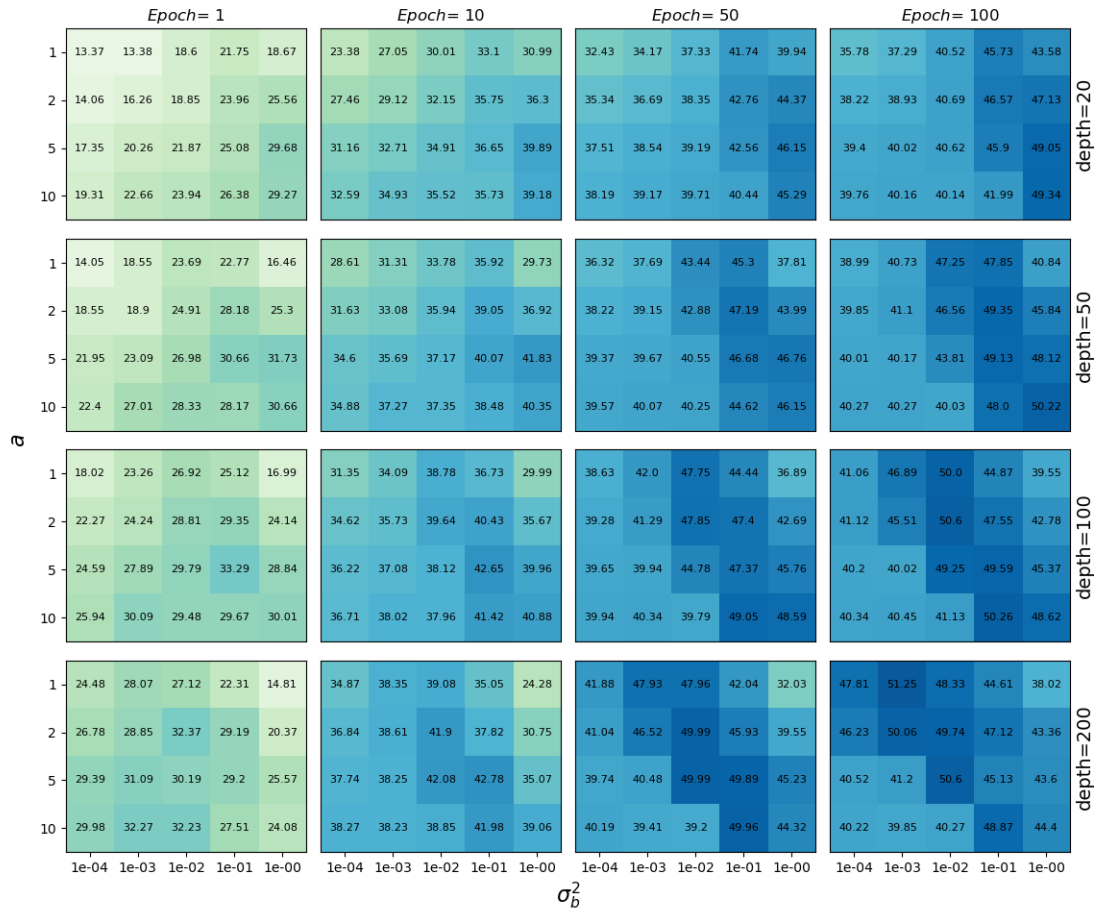(a) Shtanh with Gaussian initialisation

5   Conclusion and avenues for future work

In this paper we considered the role of the activation function in avoiding certain problems at initialisation, namely limited information propagation with depth, high sensitivity or insensitivity to perturbations of the input, and vanishing and exploding gradients. The first two of these we investigated by studying the dynamics of the preactivation correlations, and the third by analysing the spectrum of the input-output Jacobian of the network. Previously, these problems have been ameliorated by shrinking σ_b^2 in relation to the depth of the network. This is unsatisfactory for two reasons: firstly it results in the expected euclidean length of the activations shrinking towards 0 with depth, and second it places constraints on the initialisation regime, which can result in suboptimal training outcomes. Our theory and experiments clarify that shrinking σ_b^2 is not necessary, instead, it is possible to avoid these problems at initialisation by ensuring that the activation function deployed has a sufficiently large linear region around the origin. This work therefore provides a rigorous explanation for the observation that activation functions which approximate the identity near the origin perform well, particularly at initialisation.

Avenues for future work include characterising more precisely how large the linear region needs to be for a given depth, a more comprehensive investigation as to the potential benefits of a more affine initialisation, and an analysis of the approximation capabilities of a neural network whose parameters are constrained to lie in some neighbourhood of a given initialisation point. In addition, a quantitative description of how the preactivation correlation dynamics impact training is desirable.
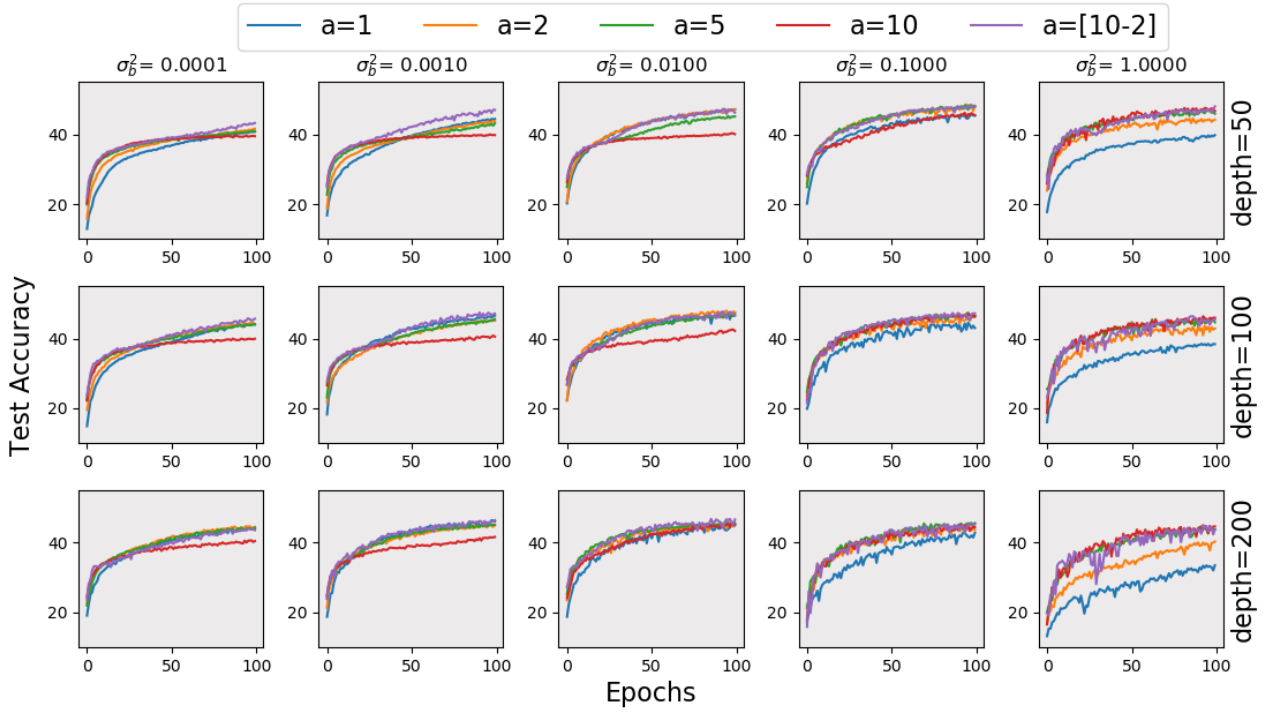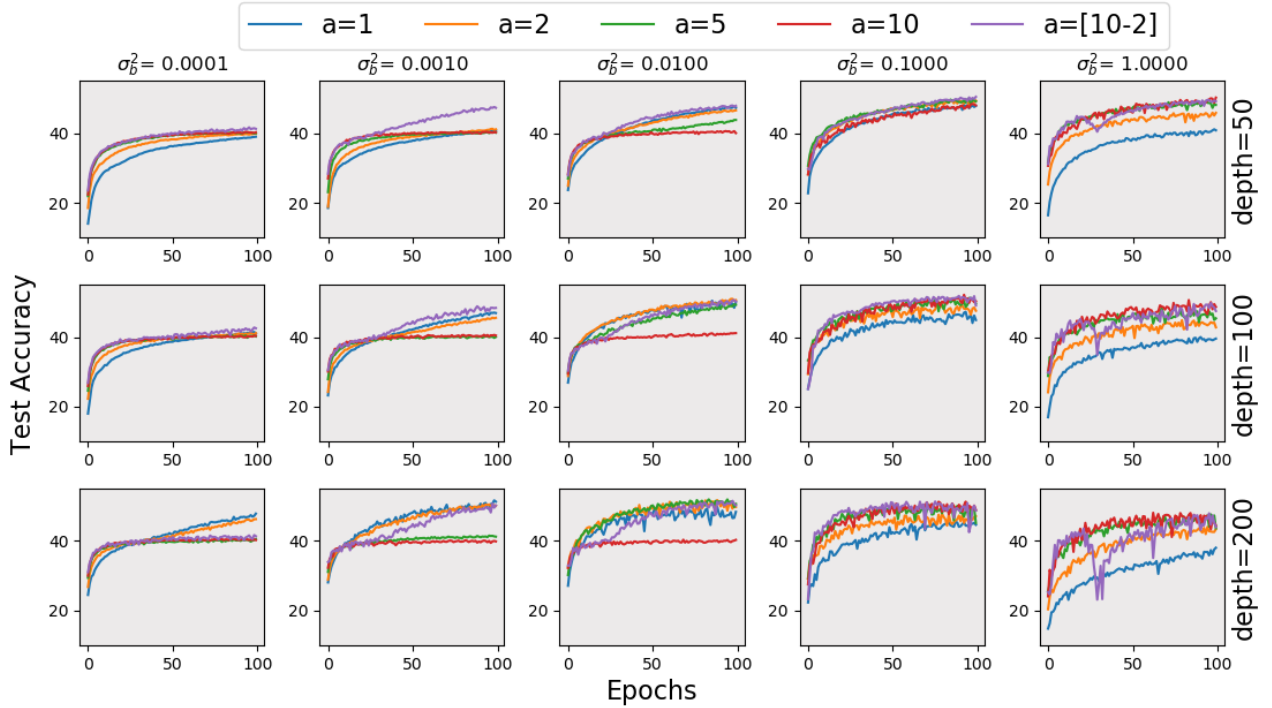
Acknowledgements

(b) Shtanh with orthogonal initialisation

Figure 7: test accuracy on CIFAR-10 at different stages of training. Note that all sub-heatmap plots share the same colour scale.

(a) Shtanh with Gaussian initialisation

(b) Shtanh with orthogonal initialisation

Figure 8: test accuracy during training on CIFAR-10. Networks are trained either with a fixed value of a or with a linearly decreasing from $10 \to 2$ over the first 30 epochs.

References

[1] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 249–256, May 2010.

[2] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *6th International Conference on Learning Representations, ICLR 2018, Workshop Track Proceedings*, 2018.

[3] Ludovic Trottier, Philippe Giguère, and Brahim Chaib-draa. Parametric exponential linear unit for deep convolutional neural networks. In *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, pages 207–214, 2017.

[4] Xinyu Liu and Xiaoguang Di. Tanhexp: A smooth activation function with high convergence speed for lightweight neural networks. *IET Computer Vision*, 15(2):136–150, 2021.

[5] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems 30*, pages 971–980, December 2017.

[6] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *International Conference on Learning Representations (ICLR)*, May 2015.

[7] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 3368–3376, 2016.

[8] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations (ICLR)*, April 2017.

[9] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1924–1932, April 2018.

[10] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning (ICML)*, pages 2672–2680, June 2019.

[11] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(2):107–116, 1998.

[12] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations(ICLR)*, April 2014.

[13] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *International Conference on Neural Information Processing Systems (NeurIPS)*, page 4788–4798, December 2017.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computing*, 9(8):1735–1780, November 1997.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[16] Rupesh . Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2377–2385, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

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

[20] Hamed Aghdam and Elnaz Heravi. *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification*. Springer, January 2017.

[21] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996.

[22] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[23] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations (ICLR)*, pages 1–11, April 2018.

[24] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 571–581, December 2018.

[25] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, Dec 2017.

[26] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[27] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning (ICML)*, pages 2798–2806, 2017.

[28] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[29] Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*. Cambridge Tracts in Mathematics. Cambridge University Press, 2019.

[30] G. B. Folland. *Real analysis : modern techniques and their applications*. Wiley, New York, 1999.

[31] G. K. Karagiannidis and A. S. Lioumpas. An improved approximation for the gaussian q-function. *IEEE Communications Letters*, 11(8):644–646, 2007.

[32] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambertw function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.

A Supporting Lemmas

For the sake of clarity and completeness we recall here two well known Lemmas concerning Lebesgue integrals. Both can be proved using a combination of Lebesgue's dominated convergence theorem and the mean value theorem (See e.g., Chapter 2 of [30]).

Lemma A.1. *Let $(X, \mathcal{F}, \mu)$ be a measure space and $Y \subset \mathbb{R}$ an open interval. Consider a function $f : X \times Y \to \mathbb{R}$ such that the following are true.*

1. *For each $y \in Y$ then the function $f_y : X \to \mathbb{R}$ with $f_y(x) := f(x, y)$ satisfies $f_y(x) \in L^1(X, \mathcal{F}, \mu)$.*

2. *For each $y \in Y$ then $\lim_{y' \to y} f(x, y') = f(x, y)$ almost everywhere in X.*

3. *For each $y \in Y$ there exists an open interval K, with $y \in K$, and a $g_K(x) \in L^1(X, \mathcal{F}, \mu)$ such that*
$$|f(x, y')| \leq g_K(x)$$
   for all $y' \in K$.

Then the function $F : Y \to \mathbb{R}$ with $F(y) := \int_X f(x, y)d\mu$ is continuous on Y.

Lemma A.2. *Let $(X, \mathcal{F}, \mu)$ be a measure space and $Y \subset \mathbb{R}$ an open interval. Consider a function $f : X \times Y \to \mathbb{R}$ such that the following are true.*

1. *For each $y \in Y$ then the function $f_y : X \to \mathbb{R}$ with $f_y(x) := f(x, y)$ satisfies $f_y(x) \in L^1(X, \mathcal{F}, \mu)$.*

2. *For each $y \in Y$ then $\frac{\partial f}{\partial y}(x, y)$ exists almost everywhere in X.*

3. *For each $y \in Y$ there exists an open interval K, with $y \in K$, and a $g_K(x) \in L^1(X, \mathcal{F}, \mu)$ such that*
$$|\frac{\partial f}{\partial y}(x, y')| \leq g_K(x)$$
   for all $y' \in K$.

Then the function $F : Y \to \mathbb{R}$ with $F(y) := \int_X f(x, y)dx$ is differentiable on Y with
$$F'(y) = \int_X \frac{\partial f}{\partial y}(x, y)d\mu. \tag{24}$$

We also now present a specific adaptation of integration by parts for piecewise continuously differentiable functions of Gaussian random variables with a finite number of discontinuities and bounded derivative.

Lemma A.3. *Suppose $f : \mathbb{R} \to \mathbb{R}$ is bounded, piecewise continuously differentiable at all but a finite number $T \in \mathbb{N}$ of non-differentiable points, $t_1 < t_2 < ... < t_T$, and has bounded derivative. Then*
$$\int_{\mathbb{R}} z f(z) d\gamma(z) = \sum_i^T \left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{t_i^+}^{t_i^-} + \int_{\mathbb{R}} f'(z) d\gamma(z) \tag{25}$$

Proof. We first note that as f and f' are continuous and continuous almost everywhere but at a finite number of points respectively, then they are clearly measureable with respect to the completion of the Borel sigma algebra on $\mathbb{R}$. Additionally, under the assumptions that f and f' are bounded, it follows that $f(z), z f(z), f'(z) \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \gamma)$ where γ is the standard one dimensional Gaussian measure. Defining, for typographical ease, $g(z) := z f(z)$, $g_-(z) := \max\{-z f(z), 0\}$ and $g_+(z) := \max\{z f(z), 0\}$, then as $|g_\pm(z)| \leq |g(z)|$ it follows that $g_-(z), g_+(z) \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \gamma)$. As both g_- and g_+ are nonnegative functions then $(g_- \mathbf{1}_{[-n,n]})_{n \in \mathbb{N}}$ and $(g_+ \mathbf{1}_{[-n,n]})_{n \in \mathbb{N}}$ are sequences of non-decreasing functions converging to g_- and g_+ respectively. Therefore, by monotone convergence
$$\int_{\mathbb{R}} z f(z) d\gamma(z) = \int_{\mathbb{R}} g_+(z) d\gamma(z) - \int_{\mathbb{R}} g_-(z) d\gamma(z)$$
$$= \lim_{n \to \infty} \int_{-n}^n g_+(z) d\gamma(z) - \lim_{n \to \infty} \int_{-n}^n g_-(z) d\gamma(z)$$
$$= \lim_{n \to \infty} \int_{-n}^n z f(z) d\gamma(z).$$

20

Let $n \in \mathbb{N}$ be any integer such that $n > |t_T|$. For typographical ease let $t_0 := -n$ and $t_{T+1} := n$, we proceed to analyse the integral of interest over $[-n, n]$.

$$\int_{-n}^{n} z f(z) d\gamma(z) = \sum_{i=0}^{T} \int_{t_i}^{t_{i+1}} f(z) z \frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} dz.$$

By construction f is continuously differentiable on each of the above intervals of integration. Standard integration by parts gives

$$\int_{t_i}^{t_{i+1}} f(z) z \frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} dz = \left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{t_i^+}^{t_{i+1}^-} + \int_{t_i}^{t_{i+1}} f'(z) d\gamma(z).$$

Collecting terms it follows that

$$\int_{-n}^{n} z f(z) d\gamma(z) = \sum_{i=0}^{T+1} \left(\left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{t_i^+}^{t_{i+1}^-} + \int_{t_i}^{t_{i+1}} f'(z) d\gamma(z) \right)$$

$$= \left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{-n}^{n} + \sum_{i=1}^{T} \left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{t_i^+}^{t_i^-} + \int_{-n}^{n} f'(z) d\gamma(z).$$

Defining $f'_-(z) := \max\{-f'(z), 0\}$ and $f'_+(z) := \max\{f'(z), 0\}$, then as $|f_\pm(z)| \leq |f(z)|$ it follows that $f_-(z), f_+(z) \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \gamma)$. As both f_- and f_+ are nonnegative functions, $(f_- \mathbf{1}_{[-n,n]})_{n \in \mathbb{N}}$ and $(f_+ \mathbf{1}_{[-n,n]})_{n \in \mathbb{N}}$ are sequences of non-decreasing functions converging to f^+ and f^- respectively. Therefore by monotone convergence

$$\int_{\mathbb{R}} f'(z) d\gamma(z) = \int_{\mathbb{R}} f'_-(z) d\gamma(z) - \int_{\mathbb{R}} f'_+(z) d\gamma(z)$$

$$= \lim_{n \to \infty} \int_{-n}^{n} f'_-(z) d\gamma(z) - \lim_{n \to \infty} \int_{-n}^{n} f'_+(z) d\gamma(z)$$

$$= \lim_{n \to \infty} \int_{-n}^{n} f'(z) d\gamma(z).$$

As a result

$$\int_{\mathbb{R}} z f(z) d\gamma(z) = \lim_{n \to \infty} \int_{-n}^{n} z f(z) d\gamma(z)$$

$$= \lim_{n \to \infty} \left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{-n}^{n} + \sum_{i=1}^{T} \left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{t_i^+}^{t_i^-} + \lim_{n \to \infty} \int_{-n}^{n} f'(z) d\gamma(z)$$

$$= \sum_{i=1}^{T} \left[-\frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}} f(z) \right]_{t_i^+}^{t_i^-} + \int_{\mathbb{R}} f'(z) d\gamma(z)$$

as claimed. □

B   Section 3 proofs

B.1   Lemma 3.1

Lemma 3.1. *Let ϕ be a scaled-bounded activation, see Definition 2.1, and $\sigma_b^2 > 0$. Define*

$$W_\phi(q) := \frac{\mathbb{E}[\phi(\sqrt{q}Z)^2]}{\mathbb{E}[\phi'(\sqrt{q}Z)^2]} + \sigma_b^2$$

for all $q \in \mathbb{R}_{\geq 0}$. Then $W_\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and W has a fixed point $q^* > 0$.

Proof. We first prove that $W_\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. To bound the numerator term then for any scaled-bounded activation function ϕ, see Definition 2.1, it follows that there exists $a, k \in \mathbb{R} > 0$ such that

$$0 \leq \mathbb{E}[\phi(\sqrt{q}Z)^2] < a^2 k^2 < \infty. \tag{26}$$

for all $q \in \mathbb{R}_{\geq 0}$. As $\mathbb{E}[\phi'(\sqrt{q}Z)^2] \geq 0$ it suffices to show that the denominator is nonzero. For a given $q \in \mathbb{R}_{\geq 0}$. as $\phi'(\sqrt{q}z)^2 \geq 0$ for all $z \in \mathbb{R} \backslash \mathcal{D}$, then $\mathbb{E}[\phi'(\sqrt{q}Z)^2] = 0$ implies that $\phi'(\sqrt{q}z) = 0$ almost everywhere for $z \in \mathbb{R}$. This is a contradiction however as there exists by construction an $a > 0$ such that $\phi(\sqrt{q}z) = kz$ for $z \in [-a/\sqrt{q}, a/\sqrt{q}]$. As a result $\mathbb{E}[\phi'(\sqrt{q}Z)^2] > 0$. We therefore conclude that $W_\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

To prove that W_ϕ has a fixed point $q^* > 0$ we need to lower and upper bound $\mathbb{E}[\phi'(\sqrt{q}Z)^2]$. A lower bound can be derived as follows,

$$
\begin{aligned}
\mathbb{E}[\phi'(\sqrt{q}Z)^2] &= \int_{\mathbb{R}} \phi'(\sqrt{q}Z)^2 d\gamma \\
&= 2 \int_0^\infty \phi'(\sqrt{q}Z)^2 d\gamma \\
&= 2 \int_0^{a/\sqrt{q}} k^2 d\gamma + 2 \int_{a/\sqrt{q}}^\infty \phi'(\sqrt{q}Z)^2 d\gamma \\
&> 2k \int_0^{a/\sqrt{q}} \phi'(\sqrt{q}Z)^2 d\gamma \\
&= k^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right).
\end{aligned}
$$

Here the second equality follows from the fact that integrand is an even function and the equality on the third line by the construction of ϕ. The inequality on the fourth line follows from zeroing the second integral which is positive. Furthermore $|\phi'(z)| \leq k$ almost everywhere by construction, therefore we conclude that

$$
k^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right) \leq \mathbb{E}[\phi'(\sqrt{q}Z)^2] < k^2 < \infty \tag{27}
$$

for all $q \in \mathbb{R}_0$.

We now prove that W_ϕ is continuous. Due to fact that $\mathbb{E}[\phi'(\sqrt{q}Z)^2] > 0$ then W_ϕ has no singularities. It therefore suffices to prove that both $q \mapsto \mathbb{E}[\phi(\sqrt{q}Z)^2]$ and $q \mapsto \mathbb{E}[\phi'(\sqrt{q}Z)^2]$ are continuous functions on $\mathbb{R}_{\geq 0}$. In both cases we achieve this by applying Lemma A.1 in the context of the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \gamma)$, where γ denotes the standard one dimensional Gaussian measure and $\mathcal{B}(\mathbb{R})$ the completion of the Borel σ algebra on the real numbers. For $q \mapsto \mathbb{E}[\phi(\sqrt{q}Z)^2]$ with $q \in \mathbb{R}_{\geq 0}$ then condition 1 is satisfied due to (26), condition 2 follows from the continuity of ϕ and condition 3 is satisfied by $g(z) := a^2 k^2$. For $q \mapsto \mathbb{E}[\phi'(\sqrt{q}Z)^2]$ with $q \in \mathbb{R}_{\geq 0}$ then condition 1 is satisfied due to (27), condition 2 follows from the fact that ϕ' is continuous almost everywhere in $\mathbb{R}$ and condition 3 is satisfied for all $q \in \mathbb{R}_{\geq 0}$ by $g(z) := k^2$. We conclude then that W_ϕ is continuous on $\mathbb{R}_{\geq 0}$.

To prove that 0 is not a fixed point, observe that $\frac{\mathbb{E}[\phi(\sqrt{q}Z)^2]}{\mathbb{E}[\phi'(\sqrt{q}Z)^2]} \geq 0$ and therefore $W_\phi(q) \geq \sigma_b^2 > 0$ for all $q \in \mathbb{R}_{\geq 0}$. Using (26) and (27) we may in addition derive the following upper bound on W_ϕ,

$$
\sigma_b^2 \leq W_\phi(q) < \frac{a^2}{\mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right)} + \sigma_b^2 =: U(q) \tag{28}
$$

for all $q \in \mathbb{R}_{\geq 0}$. Observe that as W_ϕ is continuous and $W_\phi(0) > 0$, then if W_ϕ has no fixed points it must hold that $W_\phi(q) > q$ for all $q \in \mathbb{R}_{\geq 0}$. Otherwise, by the intermediate value theorem, the function $W_\phi(p) - p$ must have a root and hence W must have a fixed point $q^* > 0$. Considering the upper bound U on W_ϕ from equation 28, assume that $U(q) > q$ for all $q \in \mathbb{R}_{\geq 0}$. Then

$$
\frac{a^2}{\mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right)} + \sigma_b^2 > q,
$$

$$
a^2 + \sigma_b^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right) > q\,\mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right)
$$

However, $\lim_{q \to \infty} a^2 + \sigma_b^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right) = a^2 < \infty$ while $\lim_{q \to \infty} q\,\mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right) = \infty$. As $q\,\mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right)$ is continuous then there must exist a $q \in \mathbb{R}_{\geq 0}$ such that $U(q) < q$, which is a contradiction. We therefore conclude that $W_\phi(q) < U(q) < q$ for some $q \in \mathbb{R}_{\geq 0}$ and that therefore W_ϕ must have a fixed point $q^* > 0$. □

B.2   Corollary 3.1.1

Corollary 3.1.1. *Let ϕ be a scaled-bounded activation, see Definition 2.1, $\sigma_b^2 > 0$ and suppose*

$$\chi_1 := \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q^*}Z)^2] = 1,$$

where $q^* > 0$ is a fixed point of W_ϕ, defined in (18). Then q^* is a fixed point of the associated variance function V_ϕ, defined in (3).

Proof. From Lemma 3.1 it holds that there exists $q^* > 0$ such that $W_\phi(q^*) = q^*$. Inspecting (3) and (18) it follows that $V_\phi(q^*) = W_\phi(q^*) = q^*$, therefore q^* is a fixed point of V_ϕ. □

B.3   Lemma 3.2

Lemma 3.2. *Let ϕ be a scaled-bounded activation, see Definition 2.1, $\sigma_b^2 > 0$ and suppose*

$$\chi_1 := \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q^*}Z)^2] = 1,$$

where $q^* > 0$ is a fixed point of W_ϕ. In addition, assume that all inputs $\mathbf{x}$ are normalised so that $||\mathbf{x}||_2^2 = q^*$. Then the associated correlation map $R_{\phi,q*}$, defined in (5), is fixed at each layer $l \in [L]$, satisfies $R_{\phi,q*} : [-1, 1] \to [-1, 1]$ and is differentiable with

$$R'_{\phi,q^*}(\rho) = \sigma_w^2 \mathbb{E}[\phi'(U_1)\phi'(U_2)]$$

for all input correlations $\rho \in (-1, 1)$.

Proof. Recall that $U_1 := \sqrt{q^*}Z_1$ and $U_2 := \sqrt{q^*}(\rho Z_1 + \sqrt{1 - \rho^2}Z_2)$. From Corollary 3.1.1 we know that q^* is a fixed point of V_ϕ and therefore the variance of all inputs, given the assumed normalisation, will remain fixed at q^* for all layers of the network. Since $q^* > 0$, then $R_{\phi,q*} : [-1, 1] \to [-1, 1]$ by construction as long as the correlation function (5) is finite for any $\rho \in [-1, 1]$. This follows by Cauchy-Schwarz,

$$|\mathbb{E}[\phi(U_1)\phi(U_2)]| \leq \mathbb{E}[\phi(U_1)^2]^{1/2}\mathbb{E}[\phi'(U_2)^2]^{1/2} < a^2 k^2 < \infty. \tag{29}$$

It remains to be proved that R_{ϕ,q^*} is differentiable on $(-1, 1)$ and to derive (21). To this end it suffices to show that $H(\rho) := \mathbb{E}[\phi(U_1)\phi(U_2)]$ is differentiable on $(-1, 1)$ and derive an expression for its derivative. We rewrite H as follows,

$$H(\rho) := \int_{\mathbb{R} \times \mathbb{R}} \phi(u_1)\phi(u_2)d\gamma^{(2)}(z_1, z_2)$$

recalling that $\gamma^{(2)}$ denotes the standard two dimensional Gaussian measure. We proceed by applying Lemma A.2 in the context of the measure space $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \gamma^2)$, where $\mathcal{B}(\mathbb{R}^2)$ denotes the completion of the Borel σ-algebra on $\mathbb{R}^2$, and the interval $(-1, 1)$. First observe that condition 1 of Lemma A.2 is satisfied as

$$\int_{\mathbb{R} \times \mathbb{R}} |\phi(u_1)\phi(u_2)|d\gamma^{(2)}(z_1, z_2) < a^2 k^2 < \infty.$$

For condition 2, by construction $(\phi \circ u_2)(\rho, z_1, z_2)$ is non-differentiable only on the set

$$\bigcup_{i=1}^{|\mathcal{D}|}\{(z_1, z_2) : \rho z_1 + \sqrt{1 - \rho^2}z_2 = \frac{d_i}{\sqrt{q^*}}\}.$$

This is the union of a finite number of one dimensional lines in $\mathbb{R}^2$ and hence has measure 0. Therefore, for each $\rho \in (-1, 1)$, then $\frac{\partial \phi \circ u_2}{(\partial \rho)}(x, z_1, z_2)$ exists almost everywhere and hence condition 2 is also satisfied. Finally, for condition 3 note that this partial derivative can be expressed as

$$\frac{\partial \phi \circ u_2}{\partial \rho}(\rho, z_1, z_2) = \frac{\partial u_2}{\partial \rho}(\rho, z_1, z_2)\phi'(u_2)$$

$$= \sqrt{q^*}\left(z_1 - \frac{\rho}{\sqrt{1 - \rho^2}}z_2\right)\phi'(u_2),$$

from which it follows

$$\left|\frac{\partial \phi \circ u_2}{\partial \rho}(\rho, z_1, z_2)\right| \leq \sqrt{q^*}\left(|z_1| + \frac{|\rho|}{\sqrt{1-\rho^2}}|z_2|\right)|\phi'(u_2)|$$

$$< \sqrt{q^*}k\left(|z_1| + \frac{|\rho|}{\sqrt{1-\rho^2}}|z_2|\right).$$

For any $\rho \in (-1, 1)$ consider the open interval $(-1 + \delta(\rho), 1 - \delta(\rho))$, where $\delta(\rho) := \frac{|1-\rho|}{2}$. It follows that

$$\sup_{\rho \in (-1+\delta(\rho), 1-\delta(\rho))} \left|\frac{\partial \phi \circ u_2}{\partial \rho}(\rho, z_1, z_2)\right| \leq \sqrt{q^*}k\left(|z_1| + \frac{1 - \delta(\rho)}{\sqrt{2\delta(\rho) - \delta(\rho)^2}}|z_2|\right) =: g_K(z_1, z_2).$$

Letting $\kappa_\delta := \frac{1-\delta(\rho)}{\sqrt{2\delta(\rho)-\delta(\rho)^2}}$, then by applying the Fubini-Tonelli theorem it follows that

$$\int_{\mathbb{R}^2} g_K(z_1, z_2)d\gamma^{(2)}(z_1, z_2) = \sqrt{q^*}k\left(\int_{\mathbb{R}^2}|z_1|d\gamma^{(2)}(z_1, z_2) + \kappa_\delta\int_{\mathbb{R}^2}|z_2|d\gamma^{(2)}(z_1, z_2)\right)$$

$$= \sqrt{q^*}k\sqrt{\frac{2}{\pi}}(1 + \kappa_\delta)$$

$$< \infty.$$

Hence for any $\rho \in (-1, 1)$ there exists an open interval $K = (-1 + \delta(\rho), 1 - \delta(\rho))$ with $\rho \in K$ and an integrable function $g_K(z_1, z_2) : \mathbb{R}^2 \to \mathbb{R}$, such that

$$\left|\frac{\partial(\phi \circ u_2)}{\partial \rho}(\rho, z_1, z_2)\right| \leq g_K(z_1, z_2).$$

We conclude then that condition 3 is also satisfied, and therefore, by Lemma A.2, H is differentiable on $(-1, 1)$. The derivative of H can be expressed as follows,

$$H'(\rho) = \int_{\mathbb{R}^2} \phi(u_1)\frac{\partial u_2}{\partial \rho}\phi'(u_2)d\gamma^{(2)}(z_1, z_2)$$

$$= \sqrt{q^*}\int_{\mathbb{R}}\int_{\mathbb{R}} \phi(u_1)\left(z_1 - \frac{\rho}{\sqrt{1-\rho^2}}z_2\right)\phi'(u_2)d\gamma(z_1)d\gamma(z_2) \tag{30}$$

$$= \sqrt{q^*}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} z_1\phi(u_1)\phi'(u_2)d\gamma(z_1)\right)d\gamma(z_2) - \frac{\rho\sqrt{q^*}}{\sqrt{1-\rho^2}}\int_{\mathbb{R}} \phi(u_1)\left(\int_{\mathbb{R}} z_2\phi'(u_2)d\gamma(z_2)\right)d\gamma(z_1),$$

where the third equality in the above follows by applying the Fubini-Tonelli theorem. We proceed first to derive an expression for $H'(0)$ and then to analyse $H'(\rho)$ for $\rho \in (-1, 0) \cup (0, 1)$. In all that follows, and as ϕ is odd, we let $-d_T ... < -d_2 < -d_1 < d_1 < d_2 .. < d_T$ be the elements of $\mathcal{D}$ where $T := |\mathcal{D}|/2$. From (30) it follows that

$$H'(0) = \sqrt{q^*}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} z_1\phi(\sqrt{q^*}z_1)\phi'(\sqrt{q^*}z_2)d\gamma(z_1)\right)d\gamma(z_2)$$

$$= \sqrt{q^*}\int_{\mathbb{R}} \phi'(\sqrt{q^*}z_2)\left(\int_{\mathbb{R}} z_1\phi(\sqrt{q^*}z_1)d\gamma(z_1)\right)d\gamma(z_2)$$

By construction $\phi(\sqrt{q^*}z_1)$ is differentiable at all points other than

$$s_i := \frac{d_i}{\sqrt{q^*}},$$

$$l_i := -\frac{d_i}{\sqrt{q^*}}$$

for $i \in [T]$. Applying Lemma A.3

$$\int_{\mathbb{R}} z_1\phi(\sqrt{q^*}z_1)d\gamma(z_1) = -\sum_i^T\left(\left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\right]_{l_i^+}^{l_i^-} + \left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\right]_{s_i^+}^{s_i^-}\right)$$

$$+ \sqrt{q^*}\int_{\mathbb{R}} \phi'(\sqrt{q^*}z_1)d\gamma(z_1)$$

$$= \sqrt{q^*}\int_{\mathbb{R}} \phi'(\sqrt{q^*}z_1)d\gamma(z_1).$$

Here the second equality follows from the continuity of ϕ. Therefore

$$
\begin{aligned}
H'(0) &= \sqrt{q^*} \int_{\mathbb{R}} \phi'(\sqrt{q^*} z_2) \left(\sqrt{q^*} \int_{\mathbb{R}} \phi'(\sqrt{q^*} z_1) d\gamma(z_1) \right) d\gamma(z_2) \\
&= q^* \int_{\mathbb{R}^2} \phi'(\sqrt{q^*} z_1) \phi'(\sqrt{q^*} z_2) d\gamma^{(2)}(z_1, z_2)
\end{aligned}
\tag{31}
$$

To derive an expression for $H'(\rho)$ for any $\rho \in (-1, 0) \cup (0, 1)$, we apply integration by parts to each of the inner integrals in (30) using Lemma A.3. Starting with the second inner integral, with $\rho \in (-1, 0) \cup (0, 1)$ and $z_1 \in \mathbb{R}$ fixed, we define $\psi : \mathbb{R} \to \mathbb{R}$ as the function $z_2 \mapsto (\phi' \circ u_2)(z_1, z_2, \rho)$. By construction ϕ' is continuously differentiable at all points other than the following,

$$
\begin{aligned}
p_i(z_1) &:= \frac{d_i}{\sqrt{q^*(1 - \rho^2)}} - z_1 \frac{\rho}{\sqrt{1 - \rho^2}}, \\
n_i(z_1) &:= -\frac{d_i}{\sqrt{q^*(1 - \rho^2)}} - z_1 \frac{\rho}{\sqrt{1 - \rho^2}}
\end{aligned}
$$

for all $i \in [T]$. Applying Lemma A.3 it follows that

$$
\begin{aligned}
\int_{\mathbb{R}} z_2 \psi(z_2) d\gamma(z_2) &= -\sum_i^T \left(\left[\frac{\exp(-\frac{1}{2} z_2^2)}{\sqrt{2\pi}} \psi(z_2) \right]_{n_i^+}^{n_i^-} + \left[\frac{\exp(-\frac{1}{2} z_2^2)}{\sqrt{2\pi}} \psi(z_2) \right]_{p_i^+}^{p_i^-} \right) + \int_{\mathbb{R}} \psi'(z_2) d\gamma(z_2) \\
&=: -\kappa_1(z_1) + \int_{\mathbb{R}} \psi'(z_2) d\gamma(z_2).
\end{aligned}
$$

Therefore the second integral on the final line of (30) can be expressed as

$$
\begin{aligned}
& \int_{\mathbb{R}} \phi(u_1) \left(\int_{\mathbb{R}} z_2 \phi'(u_2) d\gamma(z_2) \right) d\gamma(z_1) \\
&= -\int_{\mathbb{R}} \phi(u_1) \kappa_1(z_1) d\gamma(z_1) + \int_{\mathbb{R}} \phi(u_1) \int_{\mathbb{R}} \psi'(z_2) d\gamma(z_2) d\gamma(z_1) \\
&= -\int_{\mathbb{R}} \phi(u_1) \kappa_1(z_1) d\gamma(z_1) + \sqrt{q^*(1 - \rho^2)} \int_{\mathbb{R}^2} \phi(u_1) \phi''(u_2) d\gamma^{(2)}(z_1, z_2).
\end{aligned}
$$

Analysing $\kappa_1(z_1)$, then by construction $\psi(n_i) = \phi'(-d_i)$ and $\psi(p_i) = \phi'(d_i)$. Furthermore, as ϕ is odd and continuous then ϕ' is even and as a result $\phi'(-d_i^-) = \phi'(d_i^+)$ and $\phi'(-d_i^+) = \phi'(d_i^-)$. Therefore

$$
\begin{aligned}
\kappa_1(z_1) &= \sum_i^T \left(\left[\frac{\exp(-\frac{1}{2} z^2)}{\sqrt{2\pi}} \psi(z) \right]_{n_i^+}^{n_i^-} + \left[\frac{\exp(-\frac{1}{2} z^2)}{\sqrt{2\pi}} \psi(z) \right]_{p_i^+}^{p_i^-} \right) \\
&= \sum_i^T \left(\frac{\exp(-\frac{1}{2} n_i^2)}{\sqrt{2\pi}} (\psi(n_i^-) - \psi(n_i^+)) + \frac{\exp(-\frac{1}{2} p_i^2)}{\sqrt{2\pi}} (\psi(p_i^-) - \psi(p_i^+)) \right) \\
&= \sum_i^T \left(\frac{\exp(-\frac{1}{2} n_i^2)}{\sqrt{2\pi}} (\phi'(-d_i^-) - \phi'(-d_i^+)) + \frac{\exp(-\frac{1}{2} p_i^2)}{\sqrt{2\pi}} (\phi'(d_i^-) - \phi'(d_i^+)) \right) \\
&= \sum_i^T \left(\frac{\exp(-\frac{1}{2} n_i^2)}{\sqrt{2\pi}} (\phi'(d_i^+) - \phi'(d_i^-)) + \frac{\exp(-\frac{1}{2} p_i^2)}{\sqrt{2\pi}} (\phi'(d_i^-) - \phi'(d_i^+)) \right) \\
&= \frac{1}{\sqrt{2\pi}} \sum_i^T (\phi'(d_i^+) - \phi'(d_i^-)) \left(\exp(-\frac{1}{2} n_i^2) - \exp(-\frac{1}{2} p_i^2) \right).
\end{aligned}
$$

Expanding the integral involving κ_1 then

$$
\int_{\mathbb{R}} \phi(u_1) \kappa_1(z_1) d\gamma(z_1) = \sum_i^T (\phi'(d_i^+) - \phi'(d_i^-)) \int_{\mathbb{R}} \phi(u_1) \frac{\exp(-\frac{1}{2} n_i^2) - \exp(-\frac{1}{2} p_i^2)}{\sqrt{2\pi}} d\gamma(z_1).
$$

25

Observe that as

$$p_i(z_1)^2 = \frac{d_i^2}{q^*(1-\rho^2)} - z_1 \frac{2d_i\rho}{\sqrt{q^*}(1-\rho^2)} + z_1^2 \frac{\rho^2}{1-\rho^2}$$

$$n_i(z_1)^2 = \frac{d_i^2}{q^*(1-\rho^2)} + z_1 \frac{2d_i\rho}{\sqrt{q^*}(1-\rho^2)} + z_1^2 \frac{\rho^2}{1-\rho^2}$$

then clearly $n_i(z_1)^2 = p_i(-z_1)^2$. Let $\beta_i(z_1) := e^{-\frac{1}{2}n_i(z_1)^2} - e^{-\frac{1}{2}p_i(z_1)^2}$ for all $i \in [T]$, then these β_i are odd functions as

$$\beta_i(-z_1) = e^{-\frac{1}{2}n_i(-z_1)^2} - e^{-\frac{1}{2}p_i(-z_1)^2}$$

$$= e^{-\frac{1}{2}p_i(z_1)^2} - e^{-\frac{1}{2}n_i(z_1)^2}$$

$$= -\beta_i(z_1).$$

As the product of two odd functions is odd then

$$\int_{\mathbb{R}} \phi(u_1)\kappa_1(z_1)d\gamma(z_1) = \frac{1}{\sqrt{2\pi}} \sum_i^T \left(\phi'(d_i^+) - \phi'(d_i^-) \right) \int_{\mathbb{R}} \phi(u_1)\beta_i(z_1)d\gamma(z_1) = 0$$

and so

$$\int_{\mathbb{R}} \phi(u_1) \left(\int_{\mathbb{R}} z_2 \phi'(u_2)d\gamma(z_2) \right) d\gamma(z_1) = \sqrt{q^*(1-\rho^2)} \int_{\mathbb{R}^2} \phi(u_1)\phi''(u_2)d\gamma^{(2)}(z_1, z_2). \tag{32}$$

Now we turn our attention to the first inner integral on the final line of (30). With $\rho \in (-1, 0) \cup (0, 1)$ and $z_1 \in \mathbb{R}$ fixed, and defining for typographical ease $\varphi : \mathbb{R} \to \mathbb{R}$ as the function $z_1 \mapsto (\phi' \circ u_2)(z_1, z_2, \rho)$, then $\phi(u_1)\phi'(u_2)$ is continuously differentiable at all points other than the following,

$$o_i(z_2) := \frac{d_i}{\rho\sqrt{q^*}} - z_2 \frac{\sqrt{1-\rho^2}}{\rho}$$

$$e_i(z_2) := -\frac{d_i}{\rho\sqrt{q^*}} - z_2 \frac{\sqrt{1-\rho^2}}{\rho}$$

$$s_i := \frac{d_i}{\sqrt{q^*}}$$

$$l_i := -\frac{d_i}{\sqrt{q^*}}$$

for all $i \in [T]$. Applying Lemma A.3,

$$\int_{\mathbb{R}} z_1\phi(\sqrt{q^*}z_1)\varphi(z_1)d\gamma(z_1) = -\sum_i^T \left(\left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{e_i^+}^{e_i^-} + \left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{o_i^+}^{o_i^-} \right)$$

$$- \sum_i^T \left(\left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{l_i^+}^{l_i^-} + \left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{s_i^+}^{s_i^-} \right)$$

$$+ \int_{\mathbb{R}} \phi(\sqrt{q^*}z_1)\varphi'(z_1) + \sqrt{q^*}\phi'(\sqrt{q^*}z_1)\varphi(z_1)d\gamma(z_1).$$

As both ϕ and φ are continuous at l_i and s_i for all $i \in [T]$, then the left and right limits of $\phi(\sqrt{q^*}z_1)\varphi(z_1)$ at these points are equal. Therefore

$$\int_{\mathbb{R}} z_1\phi(\sqrt{q^*}z_1)\varphi(z_1)d\gamma(z_1) = -\sum_i^T \left(\left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{e_i^+}^{e_i^-} + \left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}}\phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{o_i^+}^{o_i^-} \right)$$

$$+ \int_{\mathbb{R}} \phi(\sqrt{q^*}z_1)\varphi'(z_1) + \sqrt{q^*}\phi'(\sqrt{q^*}z_1)\varphi(z_1)d\gamma(z_1)$$

$$=: -\kappa_2(z_2) + \int_{\mathbb{R}} \phi(\sqrt{q^*}z_1)\varphi'(z_1) + \sqrt{q^*}\phi'(\sqrt{q^*}z_1)\varphi(z_1)d\gamma(z_1).$$

Analysing κ_2, then by construction $\varphi(e_i) = \phi'(-d_i)$ and $\varphi(o_i) = \phi'(d_i)$. Similar to before it follows that

$$\kappa_2(z_2) = \sum_{i=1}^{T} \left(\left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}} \phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{e_i^+}^{e_i^-} + \left[\frac{\exp(-\frac{1}{2}z_1^2)}{\sqrt{2\pi}} \phi(\sqrt{q^*}z_1)\varphi(z_1) \right]_{o_i^+}^{o_i^-} \right)$$

$$= \sum_{i=1}^{T} \left(\frac{\exp(-\frac{1}{2}e_i^2)}{\sqrt{2\pi}} \phi(\sqrt{q^*}e_i) \left(\varphi(-d_i^-) - \varphi(-d_i^+) \right) + \frac{\exp(-\frac{1}{2}o_i^2)}{\sqrt{2\pi}} \phi(\sqrt{q^*}o_i) \left(\varphi(d_i^-) - \varphi(d_i^+) \right) \right)$$

$$= \sum_{i=1}^{T} \left(\frac{\exp(-\frac{1}{2}e_i^2)}{\sqrt{2\pi}} \phi(\sqrt{q^*}e_i) \left(\varphi(d_i^+) - \varphi(d_i^-) \right) + \frac{\exp(-\frac{1}{2}o_i^2)}{\sqrt{2\pi}} \phi(\sqrt{q^*}o_i) \left(\varphi(d_i^-) - \varphi(d_i^+) \right) \right)$$

$$= \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{T} \left(\left(\varphi(d_i^+) - \varphi(d_i^-) \right) \left(\exp(-\frac{1}{2}e_i^2)\phi(\sqrt{q^*}e_i) - \exp(-\frac{1}{2}o_i^2)\phi(\sqrt{q^*}o_i) \right) \right)$$

Note that $o_i(-z_2) = -e_i(z_2)$, therefore $\phi(\sqrt{q^*}o_i(-z_2)) = \phi(-\sqrt{q^*}e_i(z_2)) = -\phi(\sqrt{q^*}e_i(z_2))$ as ϕ is odd. Likewise $e_i(-z_2) = -o_i(z_2)$ and so $\phi(\sqrt{q^*}e_i(-z_2)) = \phi(-\sqrt{q^*}o_i(z_2)) = -\phi(\sqrt{q^*}o_i(z_2))$. Furthermore as

$$o_i(z_2)^2 = \frac{d_i^2}{\rho^2 q^*} - z_2 \frac{\sqrt{1-\rho^2}}{\rho} + z_2^2 \frac{1-\rho^2}{\rho^2},$$

$$e_i(z_2)^2 = \frac{d_i^2}{\rho^2 q^*} + z_2 \frac{\sqrt{1-\rho^2}}{\rho} + z_2^2 \frac{1-\rho^2}{\rho^2}$$

then $o_i(-z_2)^2 = e_i(z_2)^2$. Analogous to β_i, we now define $\Gamma_i(z_2) := \exp(-\frac{1}{2}e_i^2)\phi(\sqrt{q^*}e_i) - \exp(-\frac{1}{2}o_i^2)\phi(\sqrt{q^*}o_i)$. The fact that $\Gamma_i(z_2)$ is odd follows from

$$\Gamma_i(-z_2) = \exp(-\frac{1}{2}e_i(-z_2)^2)\phi(\sqrt{q^*}e_i(-z_2))) - \exp(-\frac{1}{2}o_i(-z_2)^2)\phi(\sqrt{q^*}o_i(-z_2))$$

$$= -\exp(-\frac{1}{2}o_i(z_2)^2)\phi(\sqrt{q^*}o_i(z_2))) + \exp(-\frac{1}{2}e_i(z_2)^2)\phi(\sqrt{q^*}e_i(z_2))$$

$$= -\Gamma_i(z_2).$$

Analysing the first integral on the last line of 30), then as Γ_i is odd for each $i \in [T]$

$$\int_{\mathbb{R}} \left(\int_{\mathbb{R}} z_1 \phi(u_1)\phi'(u_2) d\gamma(z_1) \right) d\gamma(z_2)$$

$$= -\int_{\mathbb{R}} \kappa_2(z_2) d\gamma(z_2) + \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(\sqrt{q^*}z_1)\varphi'(z_1) + \sqrt{q^*}\phi'(\sqrt{q^*}z_1)\varphi(z_1) d\gamma(z_1) d\gamma(z_2)$$

$$= -\frac{1}{\sqrt{2\pi}} \sum_{i=1}^{T} \left(\varphi(d_i^+) - \varphi(d_i^-) \right) \int_{\mathbb{R}} \Gamma_i(z_2) d\gamma(z_2) + \int_{\mathbb{R}^2} \phi(\sqrt{q^*}z_1)\varphi'(z_1) d\gamma^{(2)}(z_1, z_2) \qquad (33)$$

$$+ \sqrt{q^*} \int_{\mathbb{R}^2} \phi'(\sqrt{q^*}z_1)\varphi(z_1) d\gamma^{(2)}(z_1, z_2)$$

$$= \sqrt{q^*}\rho \int_{\mathbb{R}^2} \phi(u_1)\phi''(u_2) d\gamma^{(2)}(z_1, z_2) + \sqrt{q^*} \int_{\mathbb{R}^2} \phi'(u_1)\phi'(u_2) d\gamma^{(2)}(z_1, z_2).$$

Substituting (33) and (32) into (30) it follows that

$$H'(\rho) = \sqrt{q^*} \left(\int_{\mathbb{R}} \left(\int_{\mathbb{R}} z_1 \phi(u_1)\phi'(u_2) d\gamma(z_1) \right) d\gamma(z_2) - \frac{\rho}{\sqrt{1-\rho^2}} \int_{\mathbb{R}} \phi(u_1) \left(\int_{\mathbb{R}} z_2 \phi(u_2) d\gamma(z_2) \right) d\gamma(z_1) \right)$$

$$= q^* \left(\rho \int_{\mathbb{R}^2} \phi(u_1)\phi''(u_2) d\gamma^{(2)}(z_1, z_2) + \int_{\mathbb{R}^2} \phi'(u_1)\phi'(u_2) d\gamma(z_1, z_2) - \rho \int_{\mathbb{R}^2} \phi(u_1)\phi''(u_2) d\gamma^{(2)}(z_1, z_2) \right)$$

$$= q^* \int_{\mathbb{R}^2} \phi'(u_1)\phi'(u_2) d\gamma^{(2)}(z_1, z_2).$$

It therefore follows for all $\rho \in (-1, 1)$ that

$$R'_{\phi,q^*}(\rho) = \sigma_w^2 \int_{\mathbb{R}^2} \phi'(u_1)\phi'(u_2) d\gamma^{(2)}(z_1, z_2) = \sigma_w^2 \mathbb{E}[\phi'(U_1)\phi'(U_2)]$$

as claimed.   □

27

B.4   Lemma 3.3

Lemma 3.3. *Under the same conditions and assumptions as in Lemma 3.2, it holds that*

$$\max_{\rho \in [0,1]} |R_{\phi,q^*}(\rho) - \rho| = \frac{\sigma_b^2}{q^*}.$$

Proof.   Observe that

$$\begin{aligned}
R_{\phi,q^*}(0) &= \frac{\mathbb{E}[\phi(\sqrt{q^*}Z_1)\phi(\sqrt{q^*}Z_2)]}{q^*\mathbb{E}[\phi'(\sqrt{q^*}Z)^2]} + \frac{\sigma_b^2}{q^*} \\
&= \frac{\mathbb{E}[\phi(\sqrt{q^*}Z_1)]^2}{q^*\mathbb{E}[\phi'(\sqrt{q^*}Z)^2]} + \frac{\sigma_b^2}{q^*}.
\end{aligned}$$

As ϕ is odd then $\mathbb{E}[\phi(\sqrt{q^*}Z_1)] = 0$ and therefore

$$R_{\phi,q^*}(0) = \frac{\sigma_b^2}{q^*} > 0.$$

Lemma 3.3 follows then as long as $|R_{\phi,q^*}(\rho) - \rho| < R_{\phi,q^*}(0)$ for all $\rho \in [0,1]$. As

$$R_{\phi,q^*}(1) = \frac{\sigma_w^2}{q^*}\mathbb{E}[\phi(\sqrt{q^*}Z_1)^2] + \frac{\sigma_b^2}{q^*} = \frac{V(q^*)}{q^*} = 1$$

then $|R_{\phi,q^*}(1) - 1| = 0 < R_{\phi,q^*}(0)$. All that remains to show is that the inequality holds for $\rho \in (0,1)$. We proceed using an approach similar to that used to prove Proposition 3 in [10]. Using Lemma 3.2 then for any $\rho \in (0,1)$ we have

$$\begin{aligned}
R'_{\phi,q^*}(\rho) &= \frac{\mathbb{E}[\phi'(U_1)\phi'(U_2)]}{\mathbb{E}[\phi'(\sqrt{q^*}Z)^2]} \\
&\leq \frac{\mathbb{E}[\phi'(\sqrt{q^*}Z_1)^2]^{\frac{1}{2}}\mathbb{E}[\phi'(U_2)^2]^{\frac{1}{2}}}{\mathbb{E}[\phi'(\sqrt{q^*}Z)^2]} \\
&= \frac{\mathbb{E}[\phi'(U_2)^2]^{\frac{1}{2}}}{\mathbb{E}[\phi'(\sqrt{q^*}Z)^2]^{\frac{1}{2}}} \\
&= 1.
\end{aligned}$$

The inequality on the second line of the above follows from Cauchy-Schwarz and the equalities on the third and fourth lines are due to the fact that $Z_1, Z, U_2 \sim \mathcal{N}(0,1)$ are all identically distributed. Note that equality holds iff either $\rho = 0$ or there exists an $\alpha \in \mathbb{R}$ such that $\phi'(U_1) = \alpha\phi'(U_2)$. Since Z_1 and Z_2 are i.i.d. this can only occur if ϕ' is a constant, which in turn would imply that ϕ must be linear. However, by construction linear functions are clearly not scaled-bounded activations and therefore for any $\rho \in (0,1)$ it holds that

$$R'_{\phi,q^*}(\rho) < 1.$$

For $\rho \in (0,1)$ then integrating both sides of the above inequality and applying the fundamental theorem of calculus we have

$$\int_0^\rho R'_{\phi,q^*}(t)dt < \int_0^\rho 1dt \implies R_{\phi,q^*}(\rho) - \rho < R_{\phi,q^*}(0)$$

and

$$\int_\rho^1 R'_{\phi,q^*}(t)dt < \int_\rho^1 1dt \implies \rho - R_{\phi,q^*}(\rho) < 0.$$

As $R_{\phi,q^*}(0) > 0$ then we conclude for $\rho \in (0,1)$ that $|R_{\phi,q^*}(\rho) - \rho| < R_{\phi,q^*}(0)$. Therefore

$$|R_{\phi,q^*}(\rho) - \rho| \leq R_{\phi,q^*}(0) = \frac{\sigma_b^2}{q^*}$$

for all $\rho \in [0,1]$ as claimed.   □

B.5 Lemma 3.4

Lemma 3.4. *Under the same conditions and assumptions as in Lemma 3.2, and defining* $y := \frac{\sigma_b^2}{a^2}$, *then*

$$\Lambda(y) < \frac{a}{\sqrt{q^*}} < \left(\frac{8}{\pi}\right)^{1/6} y^{-1/3},$$

where $\Lambda(y)$ *is defined as*

$$\left(W_0\left(\frac{2}{\pi}\left(\sqrt{\frac{8}{\pi}}\exp\left(-\frac{\left(\sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}\right)^2}{2}\right)\left(\frac{1}{\left(\sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}\right)} + \left(\sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}\right)\right)\right)^{-2}\right)\right)^{1/2}$$

and W_0 *denotes the principal branch of the Lambert W function.*

Proof. To derive the upper bound on $a/\sqrt{q^*}$ we study lower bounds for $V_\phi(q)$. To this end we first lower bound $\mathbb{E}[\phi(\sqrt{q}Z)^2]$. The fact that ϕ is odd implies ϕ^2 is even, therefore

$$\mathbb{E}[\phi(\sqrt{q}Z)^2] = 2\left(\int_0^{a/\sqrt{q}} k^2 q z^2 d\gamma(z) + \int_{a/\sqrt{q}}^\infty \phi(\sqrt{q}z)^2 d\gamma(z)\right)$$

$$\geq 2\left(k^2 q \int_0^{a/\sqrt{q}} z^2 d\gamma(z)\right)$$

$$= qk^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right) - \sqrt{\frac{2}{\pi}} ak^2 \sqrt{q} \exp\left(-\frac{a^2}{2q}\right).$$

To now upper bound $\mathbb{E}[\phi'(\sqrt{q^*}Z)^2]$ we use the fact that $|\phi'(z)| \leq k$, which implies

$$\mathbb{E}[\phi'(\sqrt{q^*}Z)^2] \leq 2k^2 \int_0^\infty d\gamma(z)$$

$$= k^2.$$

As $q^* = V_\phi(q^*)$, it therefore follows that

$$q^* \geq q^* \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) - \sqrt{\frac{2}{\pi}} a\sqrt{q^*} \exp\left(-\frac{a^2}{2q^*}\right) + \sigma_b^2.$$

Rearranging and dividing by a^2 gives

$$\frac{q^*}{a^2} \mathrm{erfc}\left(\frac{a}{\sqrt{2q^*}}\right) \geq \frac{\sigma_b^2}{a^2} - \frac{\sqrt{q^*}}{a}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{a^2}{2q^*}\right).$$

For typographical ease we now substitute $x = \frac{a}{\sqrt{q^*}}$ and multiply by x^2,

$$\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) \geq x^2 \frac{\sigma_b^2}{a^2} - x\sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right).$$

It is known that $\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) \leq \sqrt{\frac{2}{\pi}}\frac{1}{x}\exp\left(-\frac{x^2}{2}\right) \leq \sqrt{\frac{2}{\pi}}\frac{1}{x}$ (see e.g., [31]). Additionally it holds that $\exp\left(-\frac{x^2}{2}\right) < \frac{1}{x^2}$, this follows from the fact that $x > 2\ln(x)$, which can be proved using elementary calculus. Using these results we formulate the following inequality,

$$x^2 \frac{\sigma_b^2}{a^2} - x\sqrt{\frac{2}{\pi}}\frac{1}{x^2} < \sqrt{\frac{2}{\pi}}\frac{1}{x},$$

which simplifies to

$$\frac{a}{\sqrt{q^*}} =: x < \left(\frac{8}{\pi}\right)^{1/6} \frac{a^{2/3}}{\sigma_b^{2/3}} = \left(\frac{8}{\pi}\right)^{1/6} y^{-1/3}.$$

as claimed.

The derivation of the lower bound is more involved, we proceed by deriving an upper bound on $V_\phi(q)$ which will allow us to upper bound q^*. First, as ϕ is upper bounded by $k^2 a^2$ then

$$\mathbb{E}[\phi(\sqrt{q}Z)^2] \leq 2\left(\int_0^{a/\sqrt{q}} k^2 q z^2 d\gamma(z) + k^2 a^2 \int_{a/\sqrt{q}}^\infty d\gamma(z)\right)$$

$$= k^2 q \mathrm{erf}\left(\frac{a}{\sqrt{2q}}\right) - k^2 \sqrt{\frac{2}{\pi}} a \sqrt{q} \exp\left(-\frac{a^2}{2q}\right) + k^2 a^2 \mathrm{erfc}\left(\frac{a}{\sqrt{2q}}\right).$$

A simple lower bound for $\mathbb{E}[\phi'(\sqrt{q^*}Z)^2]$ is as follows,

$$\mathbb{E}[\phi'(\sqrt{q^*}Z)^2] \geq 2k^2 \int_0^{a/\sqrt{q^*}} \gamma(z)$$

$$= k^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right).$$

As a result of these inequalities we may formulate the following inequality,

$$q^* \leq \frac{k^2 q^* \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) - k^2 \sqrt{\frac{2}{\pi}} a \sqrt{q^*} \exp\left(-\frac{a^2}{q^*}\right) + k^2 a^2 \mathrm{erfc}\left(\frac{a}{\sqrt{2q^*}}\right)}{k^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right)} + \sigma_b^2.$$

Rearranging this expression gives

$$q^* \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) \leq q^* \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right) - \sqrt{\frac{2}{\pi}} a \sqrt{q^*} \exp\left(-\frac{a^2}{2q^*}\right) + a^2 \mathrm{erfc}\left(\frac{a}{\sqrt{2q^*}}\right) + \sigma_b^2 \mathrm{erf}\left(\frac{a}{\sqrt{2q^*}}\right).$$

Further simplification leads to

$$\sqrt{\frac{2}{\pi}} a \sqrt{q^*} \exp\left(-\frac{a^2}{2q^*}\right) \leq (a^2 - \sigma_b^2)\mathrm{erfc}\left(\frac{a}{\sqrt{2q^*}}\right) + \sigma_b^2$$

$$< a^2 \mathrm{erfc}\left(\frac{a}{\sqrt{2q^*}}\right) + \sigma_b^2.$$

Dividing by a^2 and making the substitution $x = \frac{a}{\sqrt{q^*}}$ we arrive at the key inequality

$$\sqrt{\frac{2}{\pi}} \frac{1}{x} \exp\left(-\frac{x^2}{2}\right) < \mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) + \frac{\sigma_b^2}{a^2}. \tag{34}$$

Recalling that our objective is to lower bound the quantity $\frac{a}{\sqrt{q^*}}$, then by construction any value of x which satisfies the above inequality is a viable candidate. The challenging aspect now is to find a non-trivial candidate, in particular one that scales appropriately with a and σ_b^2. For typographical ease we now define $g(x) := \sqrt{\frac{2}{\pi}} \frac{1}{x} \exp\left(-\frac{x^2}{2}\right)$ and $h(x) := \mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) + \frac{\sigma_b^2}{a^2}$. The derivatives of each of these function are as follows,

$$g'(x) = -\sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right)\left(\frac{1}{x^2} + 1\right)$$

$$h'(x) = -\sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right),$$

note that $g'(x) > h'(x)$ for all $x \in \mathbb{R}$. We now identify a candidate lower bound by considering the linear equation $u(x)$ defined by being tangent to $g(x)$ at $x = \delta := \sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}$. Here W_0 refers to the principle branch of the Lambert W function (see e.g., [32]). The line $u(x)$ therefore has the form

$$u(x) = -\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\delta^2}{2}\right)\left(\frac{1}{\delta^2} + 1\right)x + \beta.$$

Observe that as $u(\delta) = y = g(\delta)$ by construction, then

$$\beta = y + \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\delta^2}{2}\right)\left(\frac{1}{\delta} + \delta\right)$$

$$= \sqrt{\frac{2}{\pi}}\frac{1}{\delta}\exp\left(-\frac{\delta^2}{2}\right) + \sqrt{\frac{2}{\pi}}\exp\left(-\frac{\delta^2}{2}\right)\left(\frac{1}{\delta} + \delta\right)$$

$$= \sqrt{\frac{2}{\pi}}\exp\left(-\frac{\delta^2}{2}\right)\left(\frac{2}{\delta} + \delta\right)$$

We now identify a candidate lower bound γ as the solution of $g(\gamma) = \beta$, the solution of which can be expressed using the Lambert W function, $\gamma = \sqrt{W_0\left(\frac{2}{\pi}\beta^{-2}\right)}$. Consider the following compositions of functions, $\Phi_g(y) := g(\gamma(\beta(\delta(y))))$ and $\Psi_h(y) := h(\gamma(\beta(\delta(y))))$. To ensure that $\gamma \leq \frac{a}{\sqrt{q^*}}$, i.e., that γ is indeed a valid lower bound, then it must hold that $\Phi_g(y) \geq \Phi_h(y)$. We refer the reader to Figure B.5 for a graphical description of this approach to identifying a candidate lower bound. Our task then is to inspect the range of $y \in \mathbb{R}_{\geq 0}$ values for which this inequality holds true. First we note that $\Phi_f(0) = \Phi_h(0) = 0$. Second we observe h, g, γ, β and δ are differentiable functions on $\mathbb{R}_{>0}$. As

$$\frac{d\delta}{dy} = -\frac{\sqrt{W_0\left(\frac{2}{\pi}y^{-2}\right)}}{y\left(W_0\left(\frac{2}{\pi}y^{-2}\right)+1\right)} < 0,$$

$$\frac{d\beta}{d\delta} = -\frac{\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\delta^2}{2}\right)\left(\delta^4 + \delta^2 + 2\right)}{\delta^2} < 0,$$

$$\frac{d\gamma}{d\beta} = -\frac{\sqrt{W_0\left(\frac{2}{\pi}\beta^{-2}\right)}}{\beta\left(W_0\left(\frac{2}{\pi}\beta^{-2}\right)+1\right)} < 0$$

then $\frac{d\gamma}{dy} = \frac{d\gamma}{d\beta}\frac{d\beta}{d\delta}\frac{d\delta}{dy} < 0$. Applying the chain rule

$$\Phi'_g(y) = \frac{dg}{d\gamma}\frac{d\gamma}{dy}$$

$$= -\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2}\right)\left(\frac{1}{\gamma^2}+1\right)\frac{d\gamma}{dy}$$

$$= \sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2}\right)\left(\frac{1}{\gamma^2}+1\right)\left|\frac{d\gamma}{dy}\right|$$

and

$$\Phi'_h(y) = \frac{dh}{d\gamma}\frac{d\gamma}{dy}$$

$$= \sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2}\right)\left|\frac{d\gamma}{dy}\right|.$$

The fact that

$$\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2}\right)\left(\frac{1}{\gamma^2}+2\right)\left|\frac{d\gamma}{dy}\right| > \sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2}\right)\left|\frac{d\gamma}{dy}\right|,$$

which can be further simplified to

$$\left(\frac{1}{\gamma^2}+1\right) > 1,$$

holds for all $\gamma \in \mathbb{R}$ implies that $\Phi'_g(y) > \Phi'_h(y)$ for all $y \in \mathbb{R}_{>0}$. Therefore

$$\int_0^y \Phi'_g(t) - \Phi'_h(t)dt > 0$$

for all $y \in \mathbb{R}_{>0}$. Now applying the fundamental theorem of calculus

$$\int_0^y \Phi'_g(t) - \Phi'_h(t)dt = \Phi_g(y) - \Phi_h(y) + \Phi_h(0) - \Phi_f(0)$$

$$= \Phi_g(y) - \Phi_h(y)$$

we conclude that $\Phi_g(y) > \Phi_h(y)$ for all $y \in \mathbb{R}_{>0}$. As a result, γ is a valid lower bound for $\frac{a^2}{\sqrt{q^*}}$ as long as $\frac{\sigma_b^2}{a^2} > 0$. Finally, to recover the statement of the theorem, we define the composite function $\Lambda : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ as $\Lambda(y) := \gamma(\beta(\delta(y)))$. $\square$
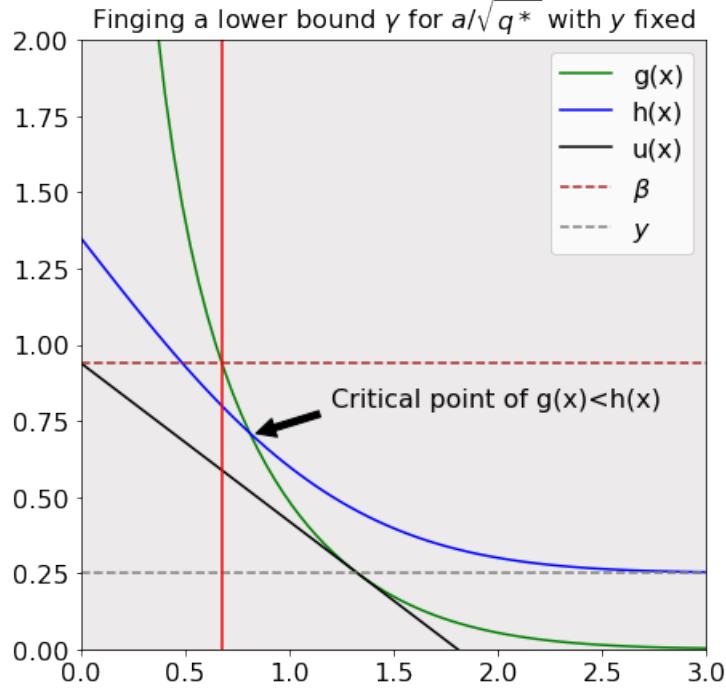

Figure 9: illustration of proof approach to finding a lower bound γ for $\frac{a}{\sqrt{q^*}}$ for a given value of $y = \sigma_b^2/a^2$ (in this example $y = 0.25$). The red line indicates the value on the x-axis defined to be γ. This is computed by first identifying where the line $u(x)$, which lies tangent to the point where g hits the asymptotic limit of $h(x)$ which in turn corresponds to the value of y, intercepts the y-axis, denoted as β. We define γ then to be the point at which the horizontal line at β (the upper dashed brown line) intercepts $g(x)$. To ensure that β intercepts $g(x)$ above the critical point of the inequality $g(x) < h(x)$, it is necessary to check that $h(\gamma) < g(\gamma)$ is true.
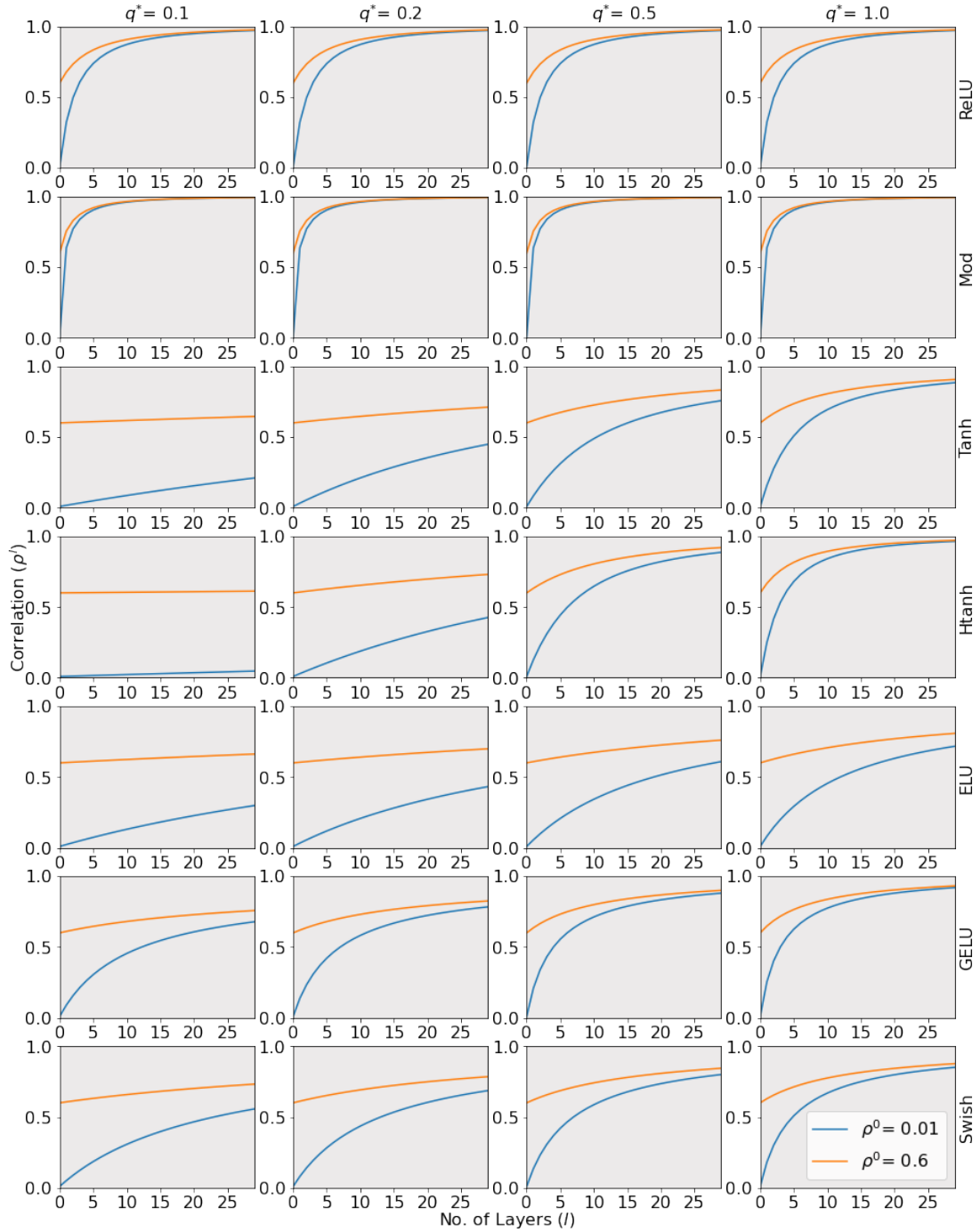
B.6   Convergence of preactivation correlations with depth for a range of activation functions

Figure 10: evolution of numerically computed preactivation correlations from two initial correlations, 0.01 and 0.6, across a range of activation functions and fixed point values.