Characterising activation functions by their backward dynamics around forward fixed points

Pieter-Jan Hoedt **Sepp Hochreiter** **Günter Klambauer**
LIT AI Lab & Institute for Machine Learning
Johannes Kepler University Linz, Austria
{hoedt, hochreit, klambauer}@ml.jku.at

Abstract

The forward dynamics in neural networks for various activation functions has been studied extensively in the context of initialisation and normalisation strategies, by mean field theory, edge of chaos theory, and fixed point analysis. However, the study of the backward dynamics appears to be largely disconnected to the insights obtained from the forward analysis. We argue that many of the ideas from the forward analysis could and should be applied to backward dynamics. We show that the ideas of mean field theory and fixed point analysis apply to the backward pass and allow to characterise activation functions.

Introduction. The importance of how the variance of data is propagated through the forward pass of a deep neural network (DNN) has been acknowledged already well before the hype of deep learning (DL) took off (LeCun et al., 1998). Glorot et al. (2010) derive an initialisation strategy which aims at preserving variance propagated through a $\tanh$ network. He et al. (2015) build upon this idea to find a proper initialisation scheme for networks with ReLUs, neglecting the backward analysis. Normalisation techniques (Ioffe et al., 2015; Salimans et al., 2016; Ba et al., 2016) aim at controlling the mean and variance of the propagated data. As an alternative, Klambauer et al. (2017) introduced SELUs with corresponding initialisation scheme to create self-normalising neural networks (SNNs). This combination of initialisation and activation function gives rise to a stable fixed point in mean and variance propagation. Poole et al. (2016), Schoenholz et al. (2017), and Yang et al. (2017) analyse the fixed points in covariance and correlation propagation using mean field theory. Their works build upon the edge-of-chaos (EOC) (Langton, 1990), which was introduced to DNNs by Natschläger et al. (2005). However, most of the recent works on initialisation schemes, normalisation, and random networks focus mainly on the forward pass. In this work, we analyse both the fixed point spectrum that different activation functions induce and connect it with an analysis of the backward dynamics (Hoedt, 2017). Concretely, we analyse networks with logistic sigmoid, $\tanh$, ReLU, SiLU (Elfwing et al., 2018; Ramachandran et al., 2018), ELU (Clevert et al., 2016) and SELU (Klambauer et al., 2017) activation functions.

Fixed points in the forward pass. We found that different activation functions give rise to *norm propagation functions* with particular properties. A norm propagation function F_ϕ is a function that maps the second moment of activations of the lower layer to the second moment of the activations of the next layer:

$$F_\phi : \mathbb{R}^+ \to \mathbb{R}^+ : q_x \mapsto F_\phi(q_x\,; g, \sigma_b) = \mathbb{E}_Z \left[\phi \left(Z \sqrt{\sigma_b^2 + g^2 q_x} \right)^2 \right], \tag{1}$$

where ϕ is the activation function, q_x is the second moment of the input layer, g is the variance of the weights scaled by the reciprocal of the number of units in the lower layer, σ_b is the variance of the bias weights, and Z is a standard Gaussian random variable. For details on the norm propagation

Critiquing and Correcting Trends in Machine Learning at NeurIPS 2018 .

function, see Appendix A. We use the norm propagation functions to characterise activation functions because these functions determine how information is propagated through the forward pass. A stable fixed point at zero indicates that the second moment of activations vanishes through layers and thus information is lost. A stable fixed point above zero means that through layers, a certain amount of information is kept. More specifically, from Figure 1, it seems that norm propagation is either linear, convex or concave for most commonly used activation functions. These convexity properties of F_ϕ and H_ϕ obviously have an effect on the fixed point spectrum.

Backward dynamics around forward fixed points. Using back-propagation, the weights and biases in each layer can be updated to minimise some loss function L. Each layer receives an error signal $\boldsymbol{\delta} = \frac{\partial L}{\partial \boldsymbol{s}} \in \mathbb{R}^M$ from its upper layer and passes it through to its lower layer. More specifically, the error at neuron j in the preceding layer is computed by means of the errors in the current layer:

$$\delta_j^\triangleleft = \phi'(s_j^\triangleleft) \sum_{i=1}^{M} \delta_i w_{ij},$$

where we used $\cdot^\triangleleft$ to denote entities from the preceding layer. Just as in the forward propagation, we can consider the entities in the backward propagation as random variables (Schoenholz et al., 2017). The assumptions on the pre-activations and weights can be taken from the forward propagation. For the errors, on the other hand, it is not possible to give much details on the distribution. For the mean of error, $\mu_d = 0$ is a reasonable assumption, since $\mu_w = 0$ and thus the mean of the error cannot propagate through the network. A more troublesome assumption that needs to be made, is that the errors are independent of the pre-activations and weights. It should be noted, however, that a product with some random Gaussian matrix tends to decorrelate inputs and outputs. With these assumptions, we can introduce norm propagation of the errors in the backward pass:

$$B_\phi(q_d\,; q_s, h) = h^2 q_d\, \mathbb{E}\left[\phi'(Z\sqrt{q_s})^2\right], \tag{2}$$

where $h^2 = M\sigma_w^2 = g^2\frac{M}{N}$ and q_s is the variance of the pre-activations. This formulation immediately reveals that the error norm propagation is linear. The effect of the norm of the pre-activations, however, appears to be non-linear. Assuming that q_s converges to some stable fixed point q_s^*, which are closely related to the fixed points of F_ϕ (see Appendix B), we can observe how the error propagates backwards through the network. Figure 2 illustrates B_ϕ for various common activation functions. The main findings are under the assumptions that we have a DNN with $\sigma_b = 0$ and $q_s = 1$ at the input layer (we refer to Appendix C for lemmata with proofs):

ReLU: has linear norm propagation functions. In the special case of $g = \sqrt{2}$, there is a manifold of fixed points with $q \geq 0$ that are neither stable nor unstable (see Figure 1a). The error norm propagation exhibits the exact same behaviour as the forward propagation and is independent of q_s (see Figure 2a).

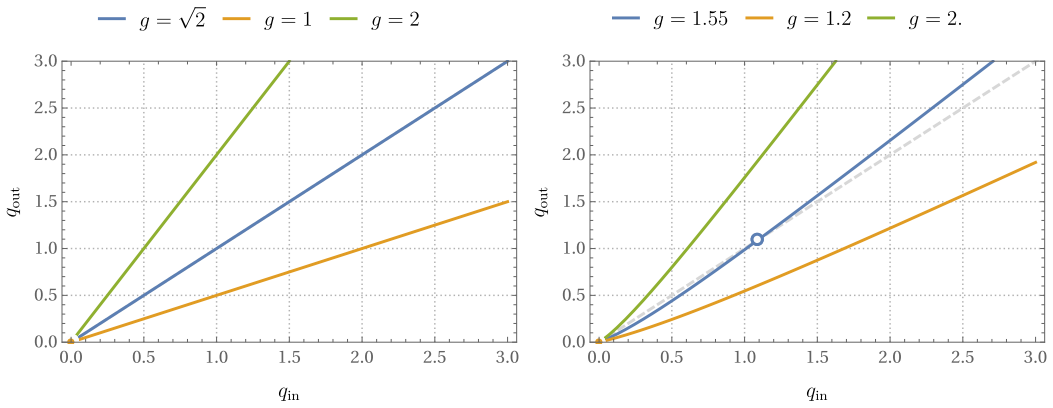
SiLU: has convex norm propagation functions which can only have an unstable fixed point for $q > 0$ (see Figure 1b). The error norm vanishes more in deeper layers than at the input layer if the unstable fixed point is greater than one (see Figure 2b).

$\boldsymbol{\sigma}$: has concave norm propagation functions with a stable fixed point $q > 0$. There is no fixed point at $q = 0$ (see Figure 1c). The error norm is vanishing during back-propagation (see Figure 2c).

tanh: has concave norm propagation functions with a point $q > 0$. There is a stable fixed point at $q = 0$ (see Figure 1d). The error norm propagation is the identity function in layers where the fixed point $q_s^* = 0$ is attained and $h = 1$. Higher h leads to a slightly exploding error norm when $q_s^* = 0$ is reached, but much better propagation in early layers (see Figure 2d).

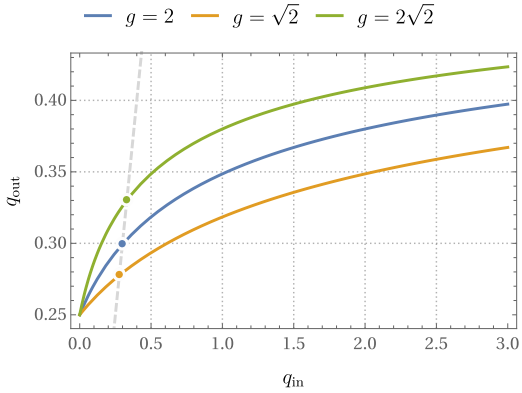
ELU: has concave norm propagation functions that can have a stable fixed point $q > 0$. There is a stable fixed point at $q = 0$ (see Figure 1e). Similar to $\tanh$, error norm propagation can be the identity function. For higher h, however, the fixed point is more realistic to attain and only slightly exploding (see Figure 2e).

SELU: has concave norm propagation functions that can have a stable fixed points at $q = 1$. There is a also stable fixed point at $q = 0$ (see Figure 1f). The error norm propagates similarly as for ELUs. The main difference appears to be that the error norm propagation is close to identity for any choice of h and has an attainable fixed point $q_s^* > 0$ by default (see Figure 2f).

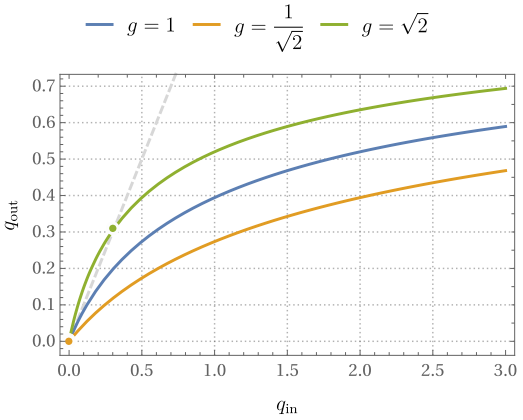


(a) ReLU

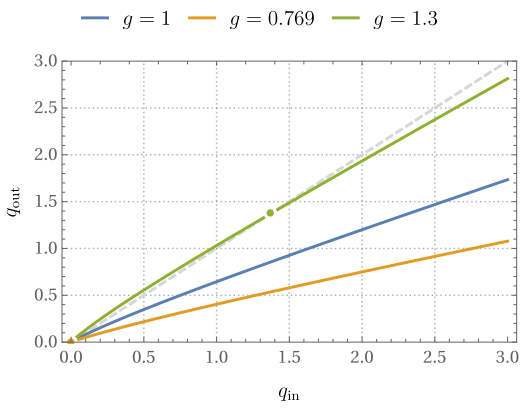
(b) SiLU



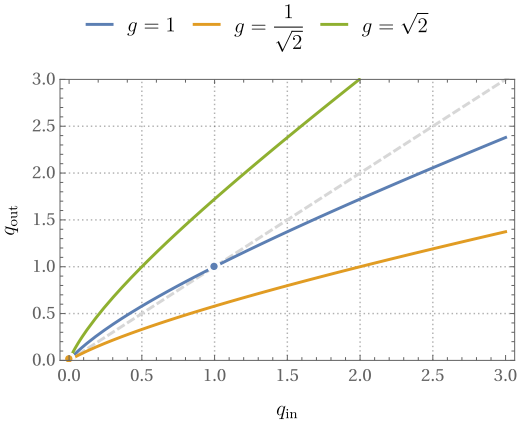
(c) σ



(d) tanh

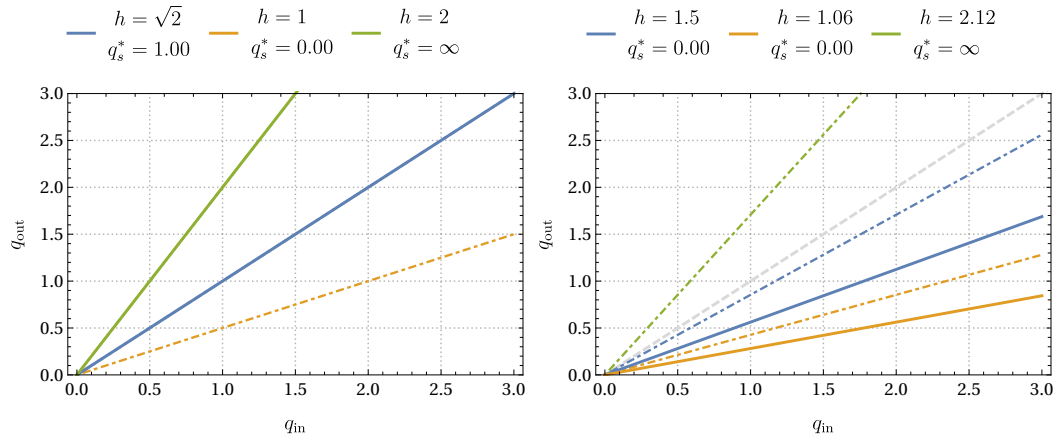


(e) ELU



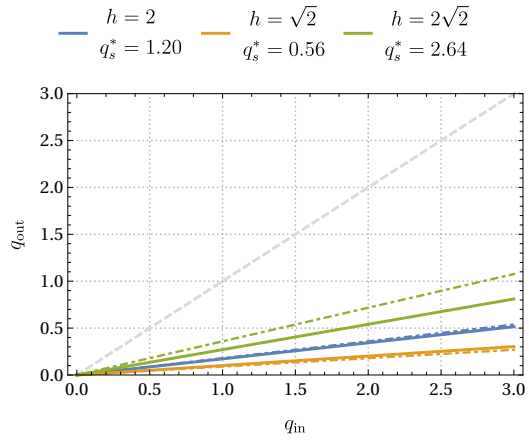
(f) SELU

Figure 1: *Norm propagation functions* F_ϕ for various common activation functions with different gain parameters for weights and $\sigma_b = 0$. The x-axis displays the norm of the neurons in the lower layer, whereas the y-axis displays the norm in the higher layer. **ReLU:** The variance is transformed linearly from one layer to the next, where a gain parameter of $g = \sqrt{2}$ corresponds to the so-called He-initialisation (He et al., 2015), yields the identity function. **SiLU:** The norm propagation function is convex and can only induce a fixed point > 0 that is unstable. **tanh and σ:** The norm propagation function is concave and can yield stable fixed points > 0. The norm propagation function is bounded. Note that $F_\sigma(0) > 0$, since $\sigma(0) = \frac{1}{2}$. **ELU and SELU:** The norm propagation function is concave and can induce a stable fixed point > 0. For SELU, the fixed point is exactly at 1.

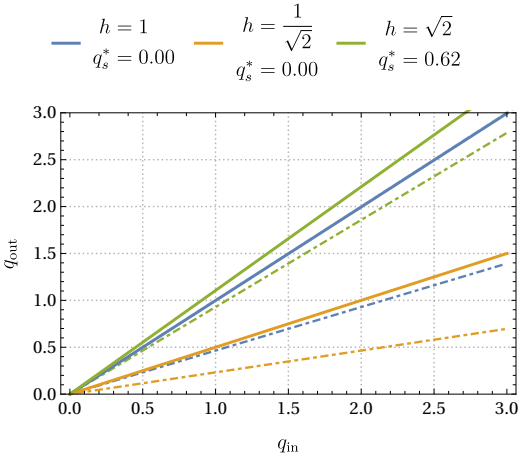


(a) ReLU

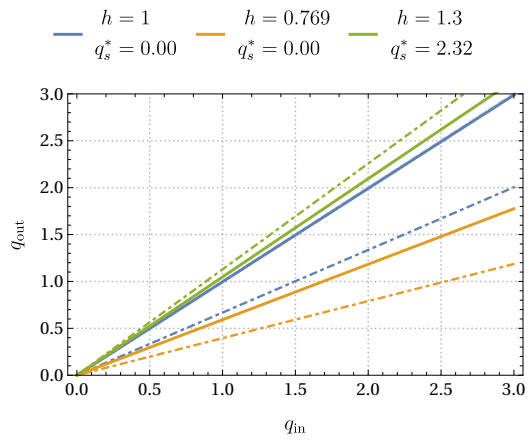
(b) SiLU



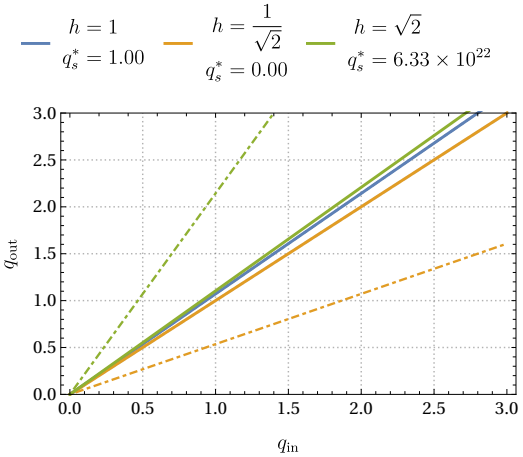
(c) σ



(d) tanh



(e) ELU



(f) SELU

Figure 2: *Norm propagation functions* for the error propagation of various common activation functions with different choices for $h = g\sqrt{\frac{M}{N}}$. The x-axis displays the norm of the error signal in the upper layer, whereas the y-axis displays the norm at the lower layer. The solid and dot-dashed lines correspond to the norm mapping for the error signal at the stable fixed point in pre-activation norms, $q_s = q_s^*$, resp. the pre-activation variance in the first layer, $q_s = 1$.

References

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). *Layer normalization*. NIPS 2016 Deep Learning Symposium.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2016). "Fast and Accurate Ceep Network Learning by Exponential Linear Units (ELUs)". In: *ICLR 2016*. Conference Track. 4th International Conference on Learning Representations. (San Juan, Puerto Rico, May 2–4, 2016). arXiv.org.

Elfwing, Stefan, Eiji Uchibe, and Kenji Doya (2018). "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning". In: *Neural Networks* 107. Special issue on deep reinforcement learning, pp. 3–11. ISSN: 0893-6080.

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of Machine Learning Research*. Thirteenth International Conference on Artificial Intelligence and Statistics. (Sardinia, Italy, May 13–15, 2010). Vol. 9. PMLR, pp. 249–256.

He, Kaiming et al. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. International Conference on Computer Vision 2015. (Santiago, Chile, Dec. 11–18, 2015). IEEE, pp. 1026–1034.

Hoedt, Pieter-Jan (2017). "Moment Dynamics in Self-Normalising Neural Networks". MA thesis. Johannes Kepler University Linz.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of Machine Learning Research*. 32nd International Conference on Machine Learning. (Lille, France, July 6–11, 2015). Vol. 37. PMLR, pp. 448–456.

Klambauer, Günter et al. (2017). "Self-Normalizing Neural Networks". In: *Advances in Neural Information Processing Systems 30*. Neural Information Processing Systems 2017. (Long Beach, CA, USA, Dec. 4–9, 2017). Vol. 31. Curran Associates, Inc., pp. 971–980.

Langton, Chris G. (1990). "Computation at the edge of chaos: Phase transitions and emergent computation". In: *Physica D: Nonlinear Phenomena* 42.1, pp. 12–37. ISSN: 0167-2789.

LeCun, Yann et al. (1998). "Efficient BackProp". In: *Neural Networks: Tricks of the Trade*. Ed. by Genevieve B. Orr and Klaus-Robert Müller. Vol. 1524. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 9–50. ISBN: 978-3-540-49430-0.

Natschläger, Thomas, Nils Bertschinger, and Robert A. Legenstein (2005). "At the Edge of Chaos: Real-time Computations and Self-Organized Criticality in Recurrent Neural Networks". In: *Advances in Neural Information Processing Systems 17*. Neural Information Processing Systems 2004. (Vancouver, BC, Canada, Dec. 13–18, 2004). Vol. 18. MIT Press, pp. 145–152.

Poole, Ben et al. (2016). "Exponential expressivity in deep neural networks through transient chaos". In: *Advances in Neural Information Processing Systems 29*. Neural Information Processing Systems 2016. (Barcelona, Spain, Dec. 5–10, 2016). Vol. 30. Curran Associates, Inc., pp. 3360–3368.

Ramachandran, Prajit, Barret Zoph, and Quoc V Le (2018). "Searching for Activation Functions". In: *ICLR 2018*. Workshop Track. 6th International Conference on Learning Representations. (Vancouver, BC, Canada, Apr. 30–May 3, 2018). openreview.net.

Salimans, Tim and Diederik P Kingma (2016). "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks". In: *Advances in Neural Information Processing Systems 29*. Neural Information Processing Systems 2016. (Barcelona, Spain, Dec. 5–10, 2016). Vol. 30. Curran Associates, Inc., pp. 901–909.

Saxe, Andrew M, James L McClelland, and Surya Ganguli (2014). "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks". In: *ICLR 2014*. Conference Track. 2nd International Conference on Learning Representations. (Banff, AB, Canada, Apr. 14–16, 2014). openreview.net.

Schoenholz, Samuel S et al. (2017). "Deep information propagation". In: *ICLR 2017*. Conference Track. 5th International Conference on Learning Representations. (Toulon, France, Apr. 24–26, 2018). openreview.net.

Yang, Ge and Samuel Schoenholz (2017). "Mean Field Residual Networks: On the Edge of Chaos". In: *Advances in Neural Information Processing Systems 30*. Neural Information Processing Systems 2017. (Long Beach, CA, USA, Dec. 4–9, 2017). Vol. 31. Curran Associates, Inc., pp. 7103–7114.

A Fixed points in the forward pass

Consider a DNN consisting of L fully connected layers. Each layer maps inputs $\boldsymbol{x} \in \mathbb{R}^N$ to activations $\boldsymbol{a} \in \mathbb{R}^M$, using some activation function ϕ. The mapping of each layer is specified by its weight matrix $\mathbf{W} \in \mathbb{R}^{M \times N}$ and bias vector $\boldsymbol{b} \in \mathbb{R}^M$. Concretely, the activation of neuron i can be computed using:

$$s_i = b_i + \sum_{j=1}^{N} w_{ij} x_j \qquad\qquad a_i = \phi(s_i),$$

where we use $\boldsymbol{s} \in \mathbb{R}^M$ to denote the pre-activations.

In analogy with (Poole et al., 2016) and (Schoenholz et al., 2017), we will consider biases and weights to be random variables with mean $\mu_b = \mu_w = 0$ and variances σ_b^2 resp. $\sigma_w^2 = g^2 \frac{1}{N}$, where g is the *gain factor* for the weights (cfr. Saxe et al., 2014). Furthermore, a single input unit x is assumed to have mean μ_x and variance σ_x^2. Assuming that N is large, the central limit theorem can be applied to conclude that the pre-activations $S \sim \mathcal{N}(0, \sigma_s)$. With the necessary independence assumptions, it can easily be verified that $\sigma_s^2 = \sigma_b^2 + g^2(\sigma_x^2 + \mu_x^2)$, since $g^2 = N\sigma_w^2$ and $\mu_w = \mu_b = 0$. We will reserve the random variable $Z \sim \mathcal{N}(0, 1)$ to denote standard Gaussian Variables.

Propagation of the first moment of activations We first introduce the concept of the *mean propagation function*, which maps the mean of the activations in one layer to the average activation in the next layer. Using the Gaussianity of the pre-activations and their variance, the mapping can be defined as:

$$M_\phi : \mathbb{R} \to \mathbb{R} : m \mapsto M_\phi(m\,; \sigma_x, g, \sigma_b) = \mathbb{E}_Z \left[\phi\left(Z \sqrt{\sigma_b^2 + g^2(\sigma_x^2 + m^2)} \right) \right], \tag{3}$$

where μ_x and σ_x are the mean resp. variance of the inputs — or the activations from the previous layer — and ϕ is the activation function.

Propagation of the second moment of activations Similarly, we can calculate the second moment of the activations in the higher layer, to which Poole et al., 2016 also refer as the normalised squared length of a vector. The effect of a fully connected layer on the second moment of its inputs, $q_x = \sigma_x^2 + \mu_x^2$, is given by the *norm propagation function*:

$$F_\phi : \mathbb{R}^+ \to \mathbb{R}^+ : q_x \mapsto F_\phi(q_x\,; g, \sigma_b) = \mathbb{E}_Z \left[\phi\left(Z \sqrt{\sigma_b^2 + g^2 q_x} \right)^2 \right]. \tag{4}$$

This mapping computes the norm of the activations from the second moment of the inputs to the layer. SNNs are the result of analysing the fixed points of mean and norm propagation functions simultaneously (Klambauer et al., 2017).

Propagation of the second moment of pre-activations Instead of studying the moments of inputs and activations, Poole et al. (2016) focused on how the correlation of pre-activations in one layer affect those in the next. Translating this approach to the moments of the pre-activations, we find a second norm propagation function:

$$H_\phi : \mathbb{R}^+ \to \mathbb{R}^+ : q_s \mapsto H_\phi(q_s\,; g, \sigma_b) = \sigma_b^2 + g^2\, \mathbb{E}_Z \left[\phi(Z\sqrt{q_s})^2 \right]. \tag{5}$$

This mapping keeps the effect of the g and σ_b outside of the non-linearity. The advantage of this second norm propagation function compared to its activation-level counterpart, is that it allows to derive results even when the expectation can not be written down analytically, as is the case for e.g. $\phi = \tanh$. We show that the fixed points of F_ϕ and H_ϕ are connected and that the stability of the fixed points is the same, see Lemma 1 in appendix B.

B Fixed Point Equivalence on Activation and Pre-activation Levels

A few simple insights follow directly from the definitions of the propagation functions:

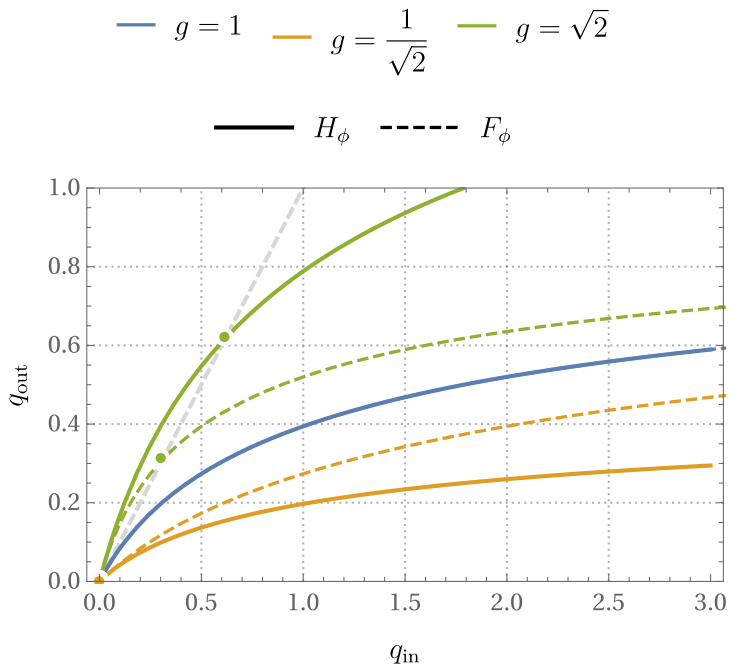


Figure 3: Norm propagation functions for both F_ϕ, for activation norms, and H_ϕ, for pre-activation norms. This illustrates the results obtained in lemma 1

Proposition 1. *For $\sigma_b = 0$ and $g = 1$, both norm propagation functions (4) and (5) are equivalent, i.e. $F_\phi(q\,; 1, 0) = H_\phi(q\,; 1, 0)$.*

Proposition 2. *Let ϕ be an activation function satisfying $\phi(0) = 0$ and $\sigma_b = 0$, then the norm propagation functions (4) and (5) have a trivial fixed point, $F_\phi(q^*\,; g, 0) = q^* = H_\phi(q^*\,; g, 0)$, at $q^* = 0$.*

A more interesting result is illustrated in Figure 3 and formulated in the following lemma:

Lemma 1. *q_x^* is a fixed point of F_ϕ if and only if $q_s^* = \sigma_b^2 + g^2 q_x^*$ is a fixed point of H_ϕ. Furthermore, the stability of both fixed points is the same.*

Proof. Let q_x^* be a fixed point of F_ϕ, then $q_x^* = \mathbb{E}\left[\phi(Z\sqrt{\sigma_b^2 + g^2 q_x^*})^2\right]$ and thus

$$H_\phi(\sigma_b^2 + g^2 q_x^*\,; g, \sigma_b) = \sigma_b^2 + g^2\,\mathbb{E}\left[\phi\left(Z\sqrt{\sigma_b^2 + g^2 q_x^*}\right)^2\right] = \sigma_b^2 + g^2 q_x^*,$$

i.e. $q_s^* = \sigma_b^2 + g^2 q_x^*$ is a fixed point of H_ϕ. In the opposite direction, the proof is equally straightforward, once it is has been observed that a fixed point of H_ϕ can not be smaller than σ_b^2.

For the stability of the fixed point, consider the derivatives

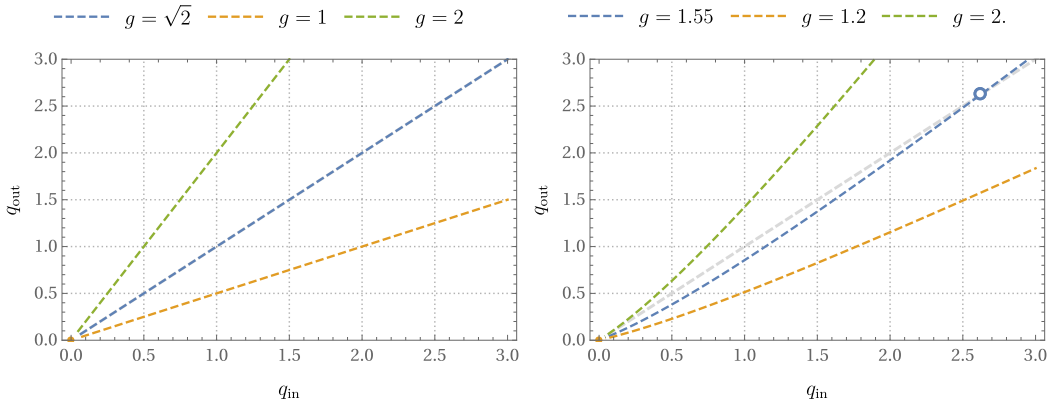
$$H'_\phi(q_s\,; g, \sigma_b) = g^2\,\mathbb{E}\left[\phi'(Z\sqrt{q_s})^2 + \phi(Z\sqrt{q_s})\phi''(Z\sqrt{q_s})\right]$$
$$F'_\phi(q_x\,; g, \sigma_b) = g^2\,\mathbb{E}\left[\phi'(Z\sqrt{\sigma_b^2 + g^2 q_x})^2 + \phi(Z\sqrt{\sigma_b^2 + g^2 q_x})\phi''(Z\sqrt{\sigma_b^2 + g^2 q_x})\right]$$

and observe that,

$$H'_\phi(q_s^*\,; g, \sigma_b) = H'_\phi(\sigma_b^2 + g^2 q_x^*\,; g, \sigma_b) = F'_\phi(q_x^*\,; g, \sigma_b),$$

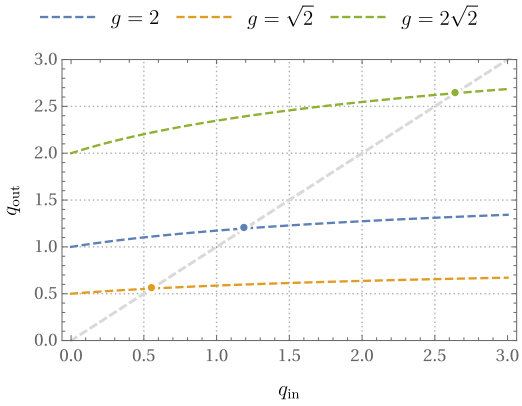
i.e. the stability of both fixed points is the same. □

For the sake of comparison and completeness, we re-compiled Figure 1 using H_ϕ instead of F_ϕ in Figure 4.

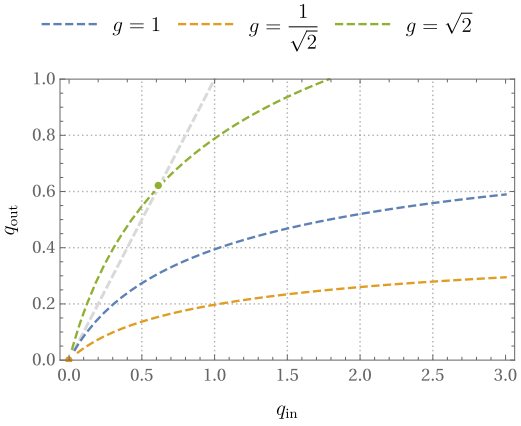


(a) ReLU

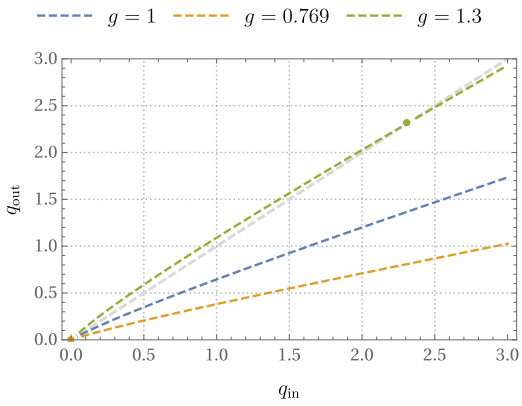
(b) SiLU



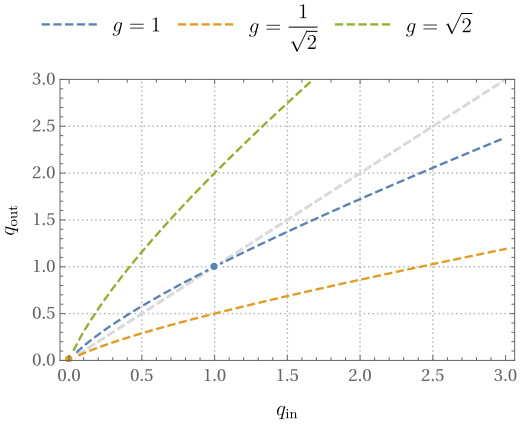
(c) σ



(d) tanh



(e) ELU



(f) SELU

Figure 4: *Norm propagation functions* H_ϕ for various common activation functions with different gain parameters for weights and $\sigma_b = 0$, cf. Figure 1.

C Convexity

Lemma 2. *The linear and* ReLU *activation functions give rise to linear norm propagation functions.*

Proof. Consider the propagation functions for $g = 1$ and $\sigma_b = 0$:

$$F_{\mathrm{id}}(q\,;1,0) = q \qquad\qquad F_{\mathrm{ReLU}}(q\,;1,0) = \frac{1}{2}q.$$

Since squeezing by a factor g^2 and/or shifting with a term σ_b^2 either horizontally (F_ϕ) or vertically (H_ϕ) preserves this linearity, we can conclude that F_{id} and F_{ReLU} must be linear. □

Lemma 3. *Let $f_\phi(x) := \phi(x)^2$ satisfying $f_\phi^{(4)} \in L^2$, then the norm propagation functions (4) and (5) are convex (concave) if $f_\phi^{(4)}(x)$ has a strictly positive (negative) Fourier transform, $\widehat{f_\phi^{(4)}}(\omega)$. Under mild conditions on the lower order derivatives of f_ϕ, also the positivity (negativity) of $\widehat{f_\phi}(\omega)$ or the negative Fourier transform of $f_\phi''(x)$ can be used.*

Proof. Without loss of generality, let $g = 1$ and $\sigma_b = 0$ so that propagation functions (4) and (5) are equivalent. Since

$$\begin{aligned}
\frac{\partial\,\mathbb{E}\left[f(Z\sqrt{q})\right]}{\partial q} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{q}}f'(z\sqrt{q})ze^{-\frac{z^2}{2}}\,\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi q}}\left(\left[-f'(z\sqrt{q})e^{-\frac{z^2}{2}}\right]_{-\infty}^{\infty} + \sqrt{q}\int_{-\infty}^{\infty}f''(z\sqrt{q})e^{-\frac{z^2}{2}}\,\mathrm{d}z\right) \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}f''(z\sqrt{q})e^{-\frac{z^2}{2}}\,\mathrm{d}z
\end{aligned}$$

if

$$\lim_{x\to\pm\infty} f'(x)e^{-x^2} = 0,$$

the convexity of F_ϕ is entirely specified by:

$$F_\phi''(q\,;1,0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{4}f_\phi^{(4)}(\sqrt{q}z)e^{-\frac{z^2}{2}}\,\mathrm{d}z.$$

Using Plancherel's identity, i.e.

$$\int_{-\infty}^{\infty} f(x)\overline{g(x)}\,\mathrm{d}x = \int_{-\infty}^{\infty}\hat{f}(\omega)\overline{\hat{g}(\omega)}\,\mathrm{d}\omega,$$

where $\hat{f}(\omega)$ is the Fourier transform of some square-integrable function $f(x) \in L^2(\mathbb{R})$, the curvature can also be written in terms of the Fourier transform of $f_\phi^{(4)}$:

$$F_\phi''(q\,;1,0) = C_0\int_{-\infty}^{\infty}\widehat{f_\phi^{(4)}}(\omega)e^{-q\frac{\omega^2}{2}}\,\mathrm{d}\omega,$$

with C_0 some (irrelevant) positive constant.

Additionally, if the functions $f_\phi^{(k)}$ for $k \in \{0,1,2,3\}$ are integrable and differentiable in the sense of distributions, the curvature of F_ϕ can be expressed by means of the Fourier transform of f_ϕ'' or f_ϕ:

$$F_\phi''(q\,;1,0) = -C_1\int_{-\infty}^{\infty}\omega^2\widehat{f_\phi''}(\omega)e^{-q\frac{\omega^2}{2}}\,\mathrm{d}\omega = C_2\int_{-\infty}^{\infty}\omega^4\hat{f}_\phi(\omega)e^{-q\frac{\omega^2}{2}}\,\mathrm{d}\omega,$$

with C_1 and C_2 positive constants.

Therefore, if $\widehat{f_\phi^{(4)}}(\omega)$ or $\widehat{f_\phi}(\omega)$ are positive (negative), F_ϕ'' will be positive (negative), making F_ϕ convex (concave). Alternatively, convexity (concavity) is guaranteed by negativity (positivity) of $\widehat{f_\phi''}(\omega)$, which is equivalent to positivity (negativity) of the Fourier transform of $-f_\phi''(x)$. □

Lemma 4. *The* $\tanh$ *and logistic sigmoid activation functions give rise to concave moment propagation functions.*

Proof. Consider the Fourier transform, $\widehat{f''_{\tanh}}(\omega)$, of the second derivative of $f_{\tanh}(x)$:

$$\widehat{f''_{\tanh}}(\omega) = \frac{8\pi^4\omega^3}{\sinh(\omega\pi^2)}.$$

Since $-\widehat{f''_{\tanh}}(\omega) = \frac{-8\pi^4\omega^3}{\sinh(\omega)}$ is strictly negative, we can conclude that $F_{\tanh}$ must be concave (cfr. lemma 3).

The logistic sigmoid can be written as $\sigma(x) = \frac{\tanh(x/2)+1}{2}$, and since

$$\mathbb{E}\left[\left(\frac{\tanh(X/2)+1}{2}\right)^2\right] = \frac{1}{4}\,\mathbb{E}\left[\tanh(X/2)^2\right] + \frac{1}{2}\,\mathbb{E}\left[\tanh(X/2)\right] + \frac{1}{4} = \frac{1}{4}\,\mathbb{E}\left[\tanh(X/2)^2\right] + \frac{1}{4},$$

F_σ must have the same convexity properties as $F_{\tanh}$. Note that $\mathbb{E}\left[\tanh(X)\right] = 0$ since $\tanh$ is odd. □

Lemma 5. *The* ELU *and* SELU *activation functions give rise to concave norm propagation functions.*

Proof. Consider the Fourier transform, $\widehat{f^{(4)}_{\mathrm{ELU}}}(\omega)$, of the generalised function, $f^{(4)}_{\mathrm{ELU}}(x)$:

$$\frac{1}{(4\pi^2\omega^2+1)\,(\pi^2\omega^2+1)}\left(-24\pi^4\omega^4 + 4\pi\omega\left[1+7\pi^2\omega^2\right]i\right).$$

First of all, note that the imaginary part of the transform,

$$\frac{4\pi\omega\left[1+7\pi^2\omega^2\right]}{(4\pi^2\omega^2+1)\,(\pi^2\omega^2+1)},$$

is an odd function and therefore disappears when integrated over from $-\infty$ to ∞ — we refer to the use of Plancherel's identity in the proof of lemma 3. Since the denominator is non-negative and $-24\pi^4\omega^4 \leq 0$, the real part of the Fourier transform must be negative. Using lemma 3, we can conclude that the norm propagation function for the ELU activation function is concave.

Because SELU is nothing more than a scaled version (with positive scale factor), F_{SELU} must also be concave. □

Proposition 3. *If the norm propagation functions are convex and* $\sigma_b = 0$*, there are no stable fixed points* $q^* > 0$*.*

Proposition 4. *If the norm propagation functions are concave and* $F'_\phi(0\,;g,0) > 1$*, there is exactly one stable fixed point* $q^* > 0$*.*