On the Impact of the Activation Function on Deep Neural Networks Training : Supplementary material

Soufiane Hayou, Arnaud Doucet, Judith Rousseau [*]

Department of Statistics
University of Oxford

1 Proofs

We provide in this supplementary material the proofs of theoretical results presented in the main document, and we give additive theoretical and experimental results. For the sake of clarity we recall the results before giving their proofs.

1.1 Convergence to the fixed point: Proposition 1

Lemma 1. *Let $M_\phi := \sup_{x\geq 0}\mathbb{E}[|\phi'^2(xZ) + \phi''(xZ)\phi(xZ)|]$. Suppose $M_\phi < \infty$, then for $\sigma_w^2 < \frac{1}{M_\phi}$ and any σ_b, we have $(\sigma_b, \sigma_w) \in D_{\phi,var}$ and $K_{\phi,var}(\sigma_b,\sigma_w) = \infty$*

Moreover, let $C_{\phi,\delta} := \sup_{x,y\geq 0,|x-y|\leq\delta,c\in[0,1]}\mathbb{E}[|\phi'(xZ_1)\phi'(y(cZ_1+\sqrt{1-c^2}Z_2)|]$. Suppose $C_{\phi,\delta} < \infty$ for some positive δ, then for $\sigma_w^2 < \min(\frac{1}{M_\phi}, \frac{1}{C_\phi})$ and any σ_b, we have $(\sigma_b,\sigma_w) \in D_{\phi,var} \cap D_{\phi,corr}$ and $K_{\phi,var}(\sigma_b,\sigma_w) = K_{\phi,corr}(\sigma_b,\sigma_w) = \infty$.

Proof. To abbreviate the notation, we use $q^l := q_a^l$ for some fixed input a.

Convergence of the variances: We first consider the asymptotic behaviour of $q^l = q_a^l$. Recall that $q^l = F(q^{l-1})$ where

$$F(x) = \sigma_b^2 + \sigma_w^2\mathbb{E}[\phi(\sqrt{x}Z)^2].$$

[*] Email adress:{soufiane.hayou,arnaud.doucet,judith.rousseau}@stats.ox.ac.uk

The first derivative of this function is given by

$$F'(x) = \sigma_w^2 \mathbb{E}[\frac{Z}{\sqrt{x}}\phi'(\sqrt{x}Z)\phi(\sqrt{x}Z)] = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{x}Z)^2 + \phi''(\sqrt{x}Z)\phi(\sqrt{x}Z)], \qquad (1)$$

where we use Gaussian integration by parts, $\mathbb{E}[ZG(Z)] = \mathbb{E}[G'(Z)]$, an identity satisfied by any function G such that $\mathbb{E}[|G'(Z)|] < \infty$.

Using the condition on ϕ, we see that the function F is a contraction mapping for $\sigma_w^2 < \frac{1}{M_\phi}$ and the Banach fixed-point theorem guarantees the existence of a unique fixed point q of F, with $\lim_{l\to+\infty} q^l = q$. Note that this fixed point depends only on F, therefore this is true for any input a and $K_{\phi,var}(\sigma_b, \sigma_w) = \infty$.

Convergence of the covariances: Since $M_\phi < \infty$, then for all $a, b \in \mathbb{R}^d$ there exists l_0 such that $|\sqrt{q_a^l} - \sqrt{q_b^l}| < \delta$ for all $l > l_0$. Let $l > l_0$, using Gaussian integration by parts, we have

$$\frac{dc_{ab}^{l+1}}{dc_{ab}^l} = \sigma_w^2 \mathbb{E}[|\phi'(\sqrt{q_a^l}Z_1)\phi'(\sqrt{q_b^l}(c_{ab}^l Z_1 + \sqrt{1-(c_{ab}^l)^2}Z_2)|].$$

We cannot use the Banach fixed point theorem directly because the integrated function here depends on l through q^l. For ease of notation, we write $c^l := c_{ab}^l$. We have

$$|c^{l+1} - c^l| = |\int_{c^{l-1}}^{c^l} \frac{dc^{l+1}}{dc^l}(x)dx| \le \sigma_w^2 C_\phi |c^l - c^{l-1}|.$$

Therefore, for $\sigma_w^2 < \min(\frac{1}{M_\phi}, \frac{1}{C_\phi})$, c^l is a Cauchy sequence and it converges to a limit $c \in [0, 1]$. At the limit

$$c = f(c) = \frac{\sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(\sqrt{q}z_1)\phi(\sqrt{q}(cz_1 + \sqrt{1-c^2}z_2)))]}{q}.$$

The derivative of this function is given by

$$f'(x) = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z_1)\phi'(\sqrt{q}(xZ_1 + \sqrt{1-x}Z_2)].$$

By assumption on ϕ and the choice of σ_w, we have $\sup_x |f'(x)| < 1$ so f is a contraction and has a unique fixed point. Since $f(1) = 1$ then $c = 1$. The above result is true for any a, b, therefore $K_{\phi,var}(\sigma_b, \sigma_w) = K_{\phi,corr}(\sigma_b, \sigma_w) = \infty$. □

Lemma 2. *Let* $(\sigma_b, \sigma_w) \in D_{\phi,var} \cap D_{\phi,corr}$ *such that* $q > 0$*,* $a, b \in \mathbb{R}^d$ *and* ϕ *an activation function such that* $\sup_{x\in K} \mathbb{E}[\phi(xZ)^2] < \infty$ *for all compact sets* K*. Define* f_l *by* $c_{a,b}^{l+1} = f_l(c_{a,b}^l)$ *and* f *by* $f(x) = \frac{\sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(\sqrt{q}Z_1)\phi(\sqrt{q}(xZ_1 + \sqrt{1-x^2}Z_2))}{q}$*. Then* $\lim_{l\to\infty} \sup_{x\in[0,1]} |f_l(x) - f(x)| = 0$*.*

Proof. For $x \in [0,1]$, we have

$$
\begin{aligned}
f_l(x) - f(x) = &(\frac{1}{\sqrt{q_a^l q_b^l}} - \frac{1}{q})(\sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(\sqrt{q_a^l} Z_1)\phi(\sqrt{q_b^l} u_2(x))]) \\
&+ \frac{\sigma_w^2}{q}(\mathbb{E}[\phi(\sqrt{q_a^l} Z_1)\phi(\sqrt{q_b^l} u_2(x))] - \mathbb{E}[\phi(\sqrt{q} Z_1)\phi(\sqrt{q} u_2(x))]),
\end{aligned}
$$

where $u_2(x) := xZ_1 + \sqrt{1-x^2}Z_2$. The first term goes to zero uniformly in x using the condition on ϕ and Cauchy-Schwartz inequality. As for the second term, it can be written again as

$$
\mathbb{E}[(\phi(\sqrt{q_a^l}Z_1) - \phi(\sqrt{q}Z_1))\phi(\sqrt{q_b^l}u_2(x))] + \mathbb{E}[\phi(\sqrt{q}Z_1)(\phi(\sqrt{q_b^l}u_2(x)) - \phi(\sqrt{q}u_2(x)))].
$$

Using Cauchy-Schwartz and the condition on ϕ, both terms can be controlled uniformly in x by an integrable upper bound. We conclude using dominated convergence. □

Lemma 3 (Weak EOC). *Let ϕ be a ReLU-like function with λ, β defined as above. Then f_l' does not depend on l, and having $f_l'(1) = 1$ and q^l bounded is only achieved for the singleton $(\sigma_b, \sigma_w) = (0, \sqrt{\frac{2}{\lambda^2+\beta^2}})$. The Weak EOC is defined as this singleton.*

Proof. We write $q^l = q_a^l$ throughout the proof. Note first that the variance satisfies the recursion:

$$
q^{l+1} = \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(Z)^2] q^l = \sigma_b^2 + \sigma_w^2 \frac{\lambda^2 + \beta^2}{2} q^l. \tag{2}
$$

For all $\sigma_w < \sqrt{\frac{2}{\lambda^2+\beta^2}}$, $q = \sigma_b^2 \left(1 - \sigma_w^2(\lambda^2 + \beta^2)/2\right)^{-1}$ is a fixed point. This is true for any input, therefore $K_{\phi,var}(\sigma_b, \sigma_w) = \infty$ and (i) is proved.

Now, the EOC equation is given by $\chi_1 = \sigma_w^2 \mathbb{E}[\phi'(Z)^2] = \sigma_w^2 \frac{\lambda^2+\beta^2}{2}$. Therefore, $\sigma_w^2 = \frac{2}{\lambda^2+\beta^2}$. Replacing σ_w^2 by its critical value in (2) yields

$$
q^{l+1} = \sigma_b^2 + q^l.
$$

Thus $q = \sigma_b^2 + q$ if and only if $\sigma_b = 0$, otherwise q^l diverges to infinity. So the frontier is reduced to a single point $(\sigma_b^2, \sigma_w^2) = (0, \mathbb{E}[\phi'(Z)^2]^{-1})$, and the variance does not depend on l.

□

Proposition 1 (EOC acts as Residual connections)**.** *Consider a ReLU network with parameters* $(\sigma_b^2, \sigma_w^2) = (0, 2) \in EOC$ *and let* c_{ab}^l *be the corresponding correlation. Consider also a ReLU network with simple residual connections given by*

$$\overline{y}_i^l(a) = \overline{y}_i^{l-1}(a) + \sum_{j=1}^{N_{l-1}} \overline{W}_{ij}^l \phi(\overline{y}_j^{l-1}(a)) + \overline{B}_i^l,$$

where $\overline{W}_{ij}^l \sim \mathcal{N}(0, \frac{\overline{\sigma}_w^2}{N_{l-1}})$ *and* $\overline{B}_i^l \sim \mathcal{N}(0, \overline{\sigma}_b^2)$*. Let* $\overline{c}_{ab}^l$ *be the corresponding correlation. Then, by taking* $\overline{\sigma}_w > 0$ *and* $\overline{\sigma}_b = 0$*, there exists a constant* $\gamma > 0$ *such that*

$$1 - c_{ab}^l \sim \gamma(1 - \overline{c}_{ab}^l) \sim \frac{9\pi^2}{2l^2}$$

as $l \to \infty$*.*

Proof. Let us first give a closed-form formula of the correlation function f of a ReLU network. In this case, we have $f(x) = 2\mathbb{E}[(Z_1)_+(xZ_1 + \sqrt{1-x^2}Z_2)_+]$ where $(x)_+ := x1_{x>0}$. Let $x \in [0,1]$, f is differentiable and satisfies

$$f'(x) = 2\mathbb{E}[1_{Z_1>0}1_{xZ_1+\sqrt{1-x^2}Z_2>0}],$$

which is also differentiable. Simple algebra leads to

$$f"(x) = \frac{1}{\pi\sqrt{1-x^2}}.$$

Since $\arcsin'(x) = \frac{1}{\sqrt{1-x^2}}$ and $f'(0) = 1/2$,

$$f'(x) = \frac{1}{\pi}\arcsin(x) + \frac{1}{2}.$$

Using the fact that $\int \arcsin = x\arcsin + \sqrt{1-x^2}$ and $f(1) = 1$, we conclude that for $x \in [0,1]$, $f(x) = \frac{1}{\pi}x\arcsin(x) + \frac{1}{\pi}\sqrt{1-x^2} + \frac{1}{2}x$.
For the residual network, we have $\overline{q}_a^l = \overline{q}_a^{l-1} + \overline{\sigma}_w^2\mathbb{E}[\phi(\sqrt{\overline{q}_a^{l-1}}Z)^2] = (1 + \frac{\overline{\sigma}_w^2}{2})\overline{q}_a^{l-1}$.
Let $\delta = \frac{1}{1+\frac{\overline{\sigma}_w^2}{2}}$. We have

$$\begin{aligned}\overline{c}_{ab}^l &= \delta\overline{c}_{ab}^{l-1} + \delta\overline{\sigma}_w^2\mathbb{E}[\phi(Z_1)\phi(U_2(\overline{c}_{ab}^{l-1}))] \\ &= \overline{c}_{ab}^{l-1} + \delta\frac{\overline{\sigma}_w^2}{2}(f(\overline{c}_{ab}^{l-1}) - \overline{c}_{ab}^{l-1})\end{aligned}$$

Now, we use Taylor expansion near to conclude. However, since f is not differentiable in 1 for all orders, we use a change of variable $x = 1 - t^2$ with t close to 0, then

$$\arcsin(1-t^2) = \frac{\pi}{2} - \sqrt{2}t - \frac{\sqrt{2}}{12}t^3 + O(t^5),$$

so that

$$\arcsin(x) = \frac{\pi}{2} - \sqrt{2}(1-x)^{1/2} - \frac{\sqrt{2}}{12}(1-x)^{3/2} + O((1-x)^{5/2}),$$

and

$$x\arcsin(x) = \frac{\pi}{2}x - \sqrt{2}(1-x)^{1/2} + \frac{11\sqrt{2}}{12}(1-x)^{3/2} + O((1-x)^{5/2}).$$

Since

$$\sqrt{1-x^2} = \sqrt{2}(1-x)^{1/2} - \frac{\sqrt{2}}{4}(1-x)^{3/2} + O((1-x)^{5/2}),$$

we obtain that

$$f(x) \underset{x\to 1-}{=} x + \frac{2\sqrt{2}}{3\pi}(1-x)^{3/2} + O((1-x)^{5/2}). \tag{3}$$

Since $(f(x)-x)' = \frac{1}{\pi}(\arcsin(x) - \frac{\pi}{2}) < 0$ and $f(1) = 1$, for all $x \in [0,1)$, $f(x) > x$. If $c^l < c^{l+1}$ then by taking the image by f (which is increasing because $f' \geq 0$) we have that $c^{l+1} < c^{l+2}$, and we know that $c^1 = f(c^0) \geq c^0$, so by induction the sequence c^l is increasing, and therefore it converges to the fixed point of f which is 1.

Using a Taylor expansion of f near 1, we have

$$\overline{c}^l_{ab} = \overline{c}^{l-1}_{ab} + \delta\frac{2\sqrt{2}}{3\pi}(1-\overline{c}^{l-1}_{ab})^{3/2} + O((1-\overline{c}^{l-1}_{ab})^{5/2})$$

and

$$c^l_{ab} = c^{l-1}_{ab} + \frac{2\sqrt{2}}{3\pi}(1-c^{l-1}_{ab})^{3/2} + O((1-c^{l-1}_{ab})^{5/2}).$$

Now let $\gamma_l := 1 - c^l_{ab}$ for a, b fixed. We note $s = \frac{2\sqrt{2}}{3\pi}$, from the series expansion we have that $\gamma_{l+1} = \gamma_l - s\gamma_l^{3/2} + O(\gamma_l^{5/2})$ so that

$$\begin{aligned}\gamma_{l+1}^{-1/2} &= \gamma_l^{-1/2}(1 - s\gamma_l^{1/2} + O(\gamma_l^{3/2}))^{-1/2} = \gamma_l^{-1/2}(1 + \frac{s}{2}\gamma_l^{1/2} + O(\gamma_l^{3/2})) \\ &= \gamma_l^{-1/2} + \frac{s}{2} + O(\gamma_l).\end{aligned}$$

Thus, as l goes to infinity

$$\gamma_{l+1}^{-1/2} - \gamma_l^{-1/2} \sim \frac{s}{2}$$

and by summing and equivalence of positive divergent series

$$\gamma_l^{-1/2} \sim \frac{s}{2} l.$$

Therefore, we have $1 - c_{ab}^l \sim \frac{9\pi^2}{2l^2}$. Using the same argument for $\overline{c}_{al}^l$, we conclude. □

Proposition 2. *Let $\phi \in \mathcal{D}_g^1$ be non ReLU-like function. Assume $V[\phi]$ is non-decreasing and $V[\phi']$ is non-increasing. Let $\sigma_{max} := \sqrt{\sup_{x \geq 0} |x - \frac{V[\phi](x)}{V[\phi'](x)}|}$ and for $\sigma_b < \sigma_{max}$ let q_{σ_b} be the smallest fixed point of the function $\sigma_b^2 + \frac{V[\phi]}{V[\phi']}$. Then we have $EOC = \{(\sigma_b, \frac{1}{\sqrt{\mathbb{E}[\phi'(\sqrt{q}Z)^2]}}) : \sigma_b < \sigma_{max}\}$.*

To prove Proposition 2, we need to introduce some lemmas. The next lemma gives a characterization of ReLU-like activation functions.

Lemma 1.1 (A Characterization of ReLU-like activations)**.** *Let $\phi \in \mathcal{D}^1(\mathbb{R}, \mathbb{R})$ such that $\phi(0) = 0$ and ϕ' non-identically zero. We define the function e for non-negative real numbers by*

$$e(x) = \frac{V[\phi](x)}{V[\phi'](x)} = \frac{\mathbb{E}[\phi(\sqrt{x}Z)^2]}{\mathbb{E}[\phi'(\sqrt{x}Z)^2]}$$

Then, for all $x \geq 0$, $e(x) \leq x$.
Moreover, the following statements are equivalent

- *There exists $x_0 > 0$ such that $e(x_0) = x_0$.*
- *ϕ is ReLU-like, i.e. there exists $\lambda, \beta \in \mathbb{R}$ such that $\phi(x) = \lambda x$ if $x > 0$ and $\phi(x) = \beta x$ if $x \leq 0$.*

Proof. Let $x > 0$. We have for all $z \in \mathbb{R}, \phi(\sqrt{x}z) = \sqrt{x} \int_0^z \phi'(\sqrt{x}u)du$. This yields

$$\begin{aligned}
\mathbb{E}[\phi(\sqrt{x}Z)^2] &= x\mathbb{E}[\big(\int_0^Z \phi'(\sqrt{x}u)du\big)^2] \\
&\leq x\mathbb{E}[|Z| \int_0^{|Z|} \phi'(\sqrt{x}u)^2 du] \\
&= x\mathbb{E}[Z \int_0^Z \phi'(\sqrt{x}u)^2 du] \\
&= x\mathbb{E}[\phi'(\sqrt{x}Z)^2 du]
\end{aligned}$$

where we have used Cauchy-Schwartz inequality and Gaussian integration by parts. Therefore $e(x) \leq x$.
Now assume there exists $x_0 > 0$ such that $e(x_0) = x_0$. We have

$$\begin{aligned}\mathbb{E}[\phi(\sqrt{x_0}Z)^2] &= x_0\mathbb{E}[\big(\int_0^Z \phi'(\sqrt{x_0}u)du\big)^2] \\ &= x_0\mathbb{E}[1_{Z>0}\big(\int_0^Z \phi'(\sqrt{x_0}u)du\big)^2] + x_0\mathbb{E}[1_{Z\leq 0}\big(\int_Z^0 \phi'(\sqrt{x_0}u)du\big)^2] \\ &\leq x_0\mathbb{E}[1_{Z>0}\int_0^Z 1du \int_0^Z \phi'(\sqrt{x_0}u)^2du] + x_0\mathbb{E}[1_{Z\leq 0}\int_Z^0 1du \int_Z^0 \phi'(\sqrt{x_0}u)^2du].\end{aligned}$$

The equality in Cauchy-Schwartz inequality implies that
- For almost every $z > 0$, there exists λ_z such that $\phi'(\sqrt{x_0}u) = \lambda_z$ for all $u \in [0, z]$.
- For almost every $z < 0$, there exists β_z such that $\phi'(\sqrt{x_0}u) = \beta_z$ for all $u \in [z, 0]$.
Therefore, λ_z, β_z are independent of z, and ϕ is ReLU-like.

It is easy to see that for ReLU-like activations, $e(x) = x$ for all $x \geq 0$. □

The next trivial lemma provides a sufficient condition for the existence of a fixed point of a shifted function.

Lemma 1.2. *Let $g \in \mathcal{C}^0(\mathbb{R}^+, \mathbb{R})$ such that $g(0) = 0$ and $g(x) \leq x$ for all $x \in \mathbb{R}^+$. Let $t_{max} := \sup_{x\geq 0}|x - g(x)|$ (t_{max} may be infinite). Then, for all $t \in [0, t_{max})$, the shifted function $t + g(.)$ has a fixed point.*

Proof. Let $t \in [0, t_{max})$. There exists $x_0 > 0$ such that $t + g(.) < x_0 - g(x_0) + g(.)$. So we have $t + g(0) = t$ and $t + g(x_0) < x_0$, which means that $t + g(.)$ crosses the identity line, therefore the fixed point exists. □

Corollary 1.1. *Let $\phi \in \mathcal{D}^1(\mathbb{R}, \mathbb{R})$ such that ϕ is non ReLU-like. Let $t_{max} = \sup_{x\geq 0}|x - \frac{V[\phi](x)}{V[\phi'](x)}|$. Then, For any $\sigma_b^2 \in [0, t_{max})$, the shifted function $\sigma_b^2 + \frac{V[\phi]}{V[\phi']}$ has a fixed point q. Moreover, by taking q to be the greatest fixed point, we have $\lim_{\sigma_b\to 0} q = 0$.*

The limit of q is zero because it is a fixed point of the function $\frac{V[\phi](x)}{V[\phi'](x)}$ which has only 0 as a fixed point for non ReLU-like functions.

Corollary 1.1 proves the existence of a fixed point for the shifted function $\sigma_b^2 + \frac{V[\phi]}{V[\phi']}$, which is a necessary condition for $(\sigma_b, 1/\sqrt{V[\phi'](q)})$ to be in the EOC where q is the smallest fixed point. It is not a sufficient condition because q may not be the smallest fixed point of $\sigma_b^2 + \frac{1}{V[\phi'](q)}V[\phi]$. We further analyse this problem hereafter.

Definition 1 (Permissible couples). *Let $g, h \in \mathcal{C}(\mathbb{R}^+, \mathbb{R}^+)$ and $c > 0$. Define the function $k(x) = c + \frac{g(x)}{h(x)}$ for $x \geq 0$ and let $q = \inf\{x : k(x) = x\}$. We say that (g, h) is permissible if for any $c \geq 0$ such that $q < \infty$, q is the smallest fixed point of the function $c + \frac{g(.)}{h(q)}$.*

Lemma 1.3. *Let $g, h \in \mathcal{C}(\mathbb{R}^+, \mathbb{R}^+)$. Then the following statements are equivalent*

1. *(g, h) is permissible.*
2. *For any $c > 0$ such that q is finite, we have $g(q) - g(x) < (q - x)h(q)$ for $x \in [0, q)$.*

Proof. If q is a fixed point of $c+\frac{g(.)}{h(.)}$, then q is clearly a fixed point of $I(x) = c+\frac{1}{h(q)}g(x)$. Having q is the smallest fixed point of $c + \frac{g(.)}{h(q)}$ is equivalent to $c + \frac{g(x)}{h(q)} > x$ for all $x \in [0, q)$. Since $c = q - \frac{g(q)}{h(q)}$, we conclude. □

Corollary 1.2. *Let $g, h \in \mathcal{C}(\mathbb{R}^+, \mathbb{R}^+)$. Assume h is non-increasing, then (g, h) is permissible.*

Proof. Since h is non-increasing, we have for $x \in [0, q)$, $g(q) - g(x) \leq h(q)(q - c) - \frac{h(q)}{h(x)}g(x) = h(q)(q - (c + \frac{g(x)}{h(x)}))$. We conclude using the fact that $c + \frac{g(x)}{h(x)} > x$ for $x \in [0, q)$. □

Corollary 1.3. *Let ϕ be a non ReLU-like function. Assume $V[\phi]$ is non-decreasing and $(V[\phi], V[\phi'])$ is permissible. Then, for any $\sigma_b^2 < t_{max} := \sup_{x \geq 0} |x - e(x)|$, by taking $\sigma_w^2 = \frac{1}{\mathbb{E}[\phi'(\sqrt{q}Z)^2]}$, we have $(\sigma_b, \sigma_w) \in EOC$. Moreover, we have $\lim_{\sigma_b \to 0} q = 0$.*

We can omit the condition '$V[\phi]$ is non-decreasing' by choosing a small t_{max}. Indeed, by taking a small σ_b, the limiting variance q is small, and we know that $V[\phi]$ is increasing near 0 because $V[\phi]'(0) = \phi'(0)^2 > 0$.

The proof of Proposition 2 is straightforward from corollary A.3.

Lemma 4. *Let ϕ be a Tanh-like activation function, then ϕ satisfies all conditions of Proposition 2 and $EOC = \{(\sigma_b, \frac{1}{\sqrt{\mathbb{E}[\phi'(\sqrt{q}Z)^2]}}) : \sigma_b \in \mathbb{R}^+\}$.*

Proof. For $x \geq 0$, we have $V[\phi]'(x) = \frac{1}{x}\mathbb{E}[\sqrt{x}Z\phi'(\sqrt{x}Z)\phi(\sqrt{x}Z)] \geq 0$, so $V[\phi]$ is non-decreasing. Moreover, $V[\phi']'(x) = \frac{1}{x}\mathbb{E}[\sqrt{x}Z\phi''(\sqrt{x}Z)\phi'(\sqrt{x}Z)] \leq 0$, therefore $V[\phi']$ is non-increasing. To conclude, we still have to show that $t_{max} = \infty$.

Using the second condition on ϕ, there exists $M > 0$ such that $|\phi'(y)|^2 \geq Me^{-2\alpha|y|}$. Let $x > 0$. we have

$$\begin{aligned}\mathbb{E}[\phi'(\sqrt{x}Z)^2] &\geq M\mathbb{E}[e^{-2\alpha|\sqrt{x}Z|}] \\ &= 2M\int_0^\infty e^{-2\alpha\sqrt{x}Z}\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz \\ &= 2Me^{2\alpha^2 x}\Psi(2\alpha\sqrt{x}) \\ &\sim \frac{2M}{2\alpha\sqrt{x}}\end{aligned}$$

where Ψ is the Gaussian cumulative function and where we used the asymptotic approximation $\Psi(x) \sim \frac{e^{-x^2/2}}{x}$ for large x.
Using this lower bound and the upper bound on ϕ, there exists $x_0, k > 0$ such that for $x > x_0$, we have $x - \frac{V[\phi](x)}{V[\phi'](x)} \geq x - k\sqrt{x} \to \infty$ which concludes the proof. □

Proposition 3 (Convergence rate for smooth activations)**.** *Let $\phi \in \mathcal{A}$ such that ϕ non-linear (i.e. $\phi^{(2)}$ is non-identically zero). Then, on the EOC, we have $1 - c^l \sim \frac{\beta_q}{l}$ where $\beta_q = \frac{2\mathbb{E}[\phi'(\sqrt{q}Z)^2]}{q\mathbb{E}[\phi''(\sqrt{q}Z)^2]}$.*

Proof. We first prove that $\lim_{l\to\infty} c^l = 1$ on the EOC. Let $x \in [0, 1)$ and $u_2(x) := xZ_1 + \sqrt{1 - x^2}Z_2$, we have

$$\begin{aligned}f'(x) &= \sigma_w^2\mathbb{E}[\phi'(\sqrt{q}Z_1)\phi'(\sqrt{q}u_2(x))] \\ &\leq \sigma_w^2(\mathbb{E}[\phi'(\sqrt{q}Z_1)^2])^{1/2}(\mathbb{E}[\phi'(\sqrt{q}u_2(x))^2])^{1/2} \\ &= 1\end{aligned}$$

where we have used Cauchy Schwartz inequality and the fact the $\sigma_w^2 = \frac{1}{\mathbb{E}[\phi'(\sqrt{q}Z)^2]}$. Moreover, the equality holds if and only if there exists a constant s such that $\phi'(\sqrt{q}(xz_1 + \sqrt{1 - x^2}z_2)) = s\phi'(\sqrt{q}z_1)$ for almost any $z_1, z_2 \in \mathbb{R}$, which is equivalent to having ϕ' equal to a constant almost everywhere on $\mathbb{R}$, hence ϕ is linear and q does not exists. This proves that for all $x \in [0, 1)$, $f'(x) < 1$. Integrating both sides between x and 1 yields $f(x) > x$ for all $x \in [0, 1)$. Therefore c^l is non-decreasing and converges to the fixed point of f which is 1.

Now we want to prove that f admits a Taylor expansion near 1. It is easy to do that if $\phi \in \mathcal{D}_g^3$. Indeed, using the conditions on ϕ, we can easily see that f has a third derivative at 1 and we have

$$\begin{aligned}f'(1) &= \sigma_w^2\mathbb{E}[\phi'(\sqrt{q}Z)^2] \\ f''(1) &= \sigma_w^2 q\mathbb{E}[\phi''(\sqrt{q}Z)^2].\end{aligned}$$

A Taylor expansion near 1 yields

$$\begin{aligned} f(x) &= 1 + f'(1)(x-1) + \frac{(x-1)^2}{2} f''(1) + O((x-1)^3) \\ &= x + \frac{(x-1)^2}{\beta_q} + O((x-1)^3). \end{aligned}$$

The proof is a bit more complicated for general $\phi \in \mathcal{A}$. We prove the result when $\phi^{(2)}(x) = 1_{x<0} g_1(x) + 1_{x \geq 0} g_2(x)$. The generalization to the whole class is straightforward. Let us first show that there exists $g \in \mathcal{C}^1$ such that $f^{(3)}(x) = \frac{1}{\sqrt{1-x^2}} g(x)$.
We have

$$\begin{aligned} f''(x) &= \sigma_w^2 q \mathbb{E}[\phi''(\sqrt{q} Z_1) \phi''(\sqrt{q} U_2(x))] \\ &= \sigma_w^2 q \mathbb{E}[\phi''(\sqrt{q} Z_1) 1_{U_2(x)<0} g_1(\sqrt{q} U_2(x))] + \sigma_w^2 q \mathbb{E}[\phi''(\sqrt{q} Z_1) 1_{U_2(x)>0} g_2(\sqrt{q} U_2(x))]. \end{aligned}$$

Let $G(x) = \mathbb{E}[\phi''(\sqrt{q} Z_1) 1_{U_2(x)<0} g_1(\sqrt{q} U_2(x))]$ then

$$\begin{aligned} G'(x) = \mathbb{E}[\phi''(\sqrt{q} Z_1)(Z_1 - \frac{x}{\sqrt{1-x^2}} Z_2) \delta_{U_2(x)=0} \frac{1}{\sqrt{1-x^2}} g_1(\sqrt{q} U_2(x))] \\ + \mathbb{E}[\phi''(\sqrt{q} Z_1) 1_{U_2(x)<0} \sqrt{q} (Z_1 - \frac{x}{\sqrt{1-x^2}} Z_2) g_1'(\sqrt{q} U_2(x))]. \end{aligned}$$

After simplification, it is easy to see that $G'(x) = \frac{1}{\sqrt{1-x^2}} G_1(x)$ where $G_1 \in \mathcal{C}^1$. By extending the same analysis to the second term of f'', we conclude that there exists $g \in \mathcal{C}^1$ such that $f^{(3)}(x) = \frac{1}{\sqrt{1-x^2}} g(x)$.

Let us now derive a Taylor expansion of f near 1. Since $f^{(3)}$ is potentially non defined at 1, we use the change of variable $x = 1 - t^2$ to compensate this effect. Simple algebra shows that the function $t \to f(1-t^2)$ has a Taylor expansion near 0

$$f(1-t^2) = 1 - t^2 f'(1) + \frac{t^4}{2} f''(1) + O(t^5).$$

Therefore,

$$f(x) = 1 + (x-1) f'(1) + \frac{(x-1)^2}{2} f''(1) + O((x-1)^{5/2}).$$

Note that this expansion is weaker than the expansion when $\phi \in \mathcal{D}_g^3$.

Denote $\lambda_l := 1 - c^l$, we have

$$\lambda_{l+1} = \lambda_l - \frac{\lambda_l^2}{\beta_q} + O(\lambda_l^{5/2})$$

therefore,

$$\begin{aligned}
\lambda_{l+1}^{-1} &= \lambda_l^{-1}(1 - \frac{\lambda_l}{\beta_q} + O(\lambda_l^{3/2}))^{-1} \\
&= \lambda_l^{-1}(1 + \frac{\lambda_l}{\beta_q} + O(\lambda_l^{3/2})) \\
&= \lambda_l^{-1} + \frac{1}{\beta_q} + O(\lambda_l^{1/2}).
\end{aligned}$$

By summing (divergent series), we conclude that $\lambda_l^{-1} \sim \frac{l}{\beta_q}$. □

Proposition 4. *Let $\phi \in \mathcal{A}$ be a non-linear activation function such that $\phi(0) = 0$, $\phi'(0) \neq 0$. Assume that $V[\phi]$ is non-decreasing and $V[\phi'])$ is non-increasing, and let $\sigma_{max} > 0$ be defined as in Proposition 2. Define the gradient with respect to the l^{th} layer by $\frac{\partial E}{\partial y^l} = (\frac{\partial E}{\partial y_i^l})_{1 \leq i \leq N_l}$ and let $\tilde{Q}_{ab}^l = \mathbb{E}[\frac{\partial E}{\partial y_a^l}^T \frac{\partial E}{\partial y_b^l}]$ denote the covariance matrix of the gradients during backpropagation. Recall that $\beta_q = \frac{2\mathbb{E}[\phi'(\sqrt{q}Z)^2]}{q\mathbb{E}[\phi''(\sqrt{q}Z)^2]}$.*

Then, for any $\sigma_b < \sigma_{max}$, by taking $(\sigma_b, \sigma_w) \in EOC$ we have

- $\sup_{x \in [0,1]} |f(x) - x| \leq \frac{1}{\beta_q}$
- *For $l \geq 1$, $|\frac{\mathrm{Tr}(\tilde{Q}_{ab}^l)}{\mathrm{Tr}(\tilde{Q}_{ab}^{l+1})} - 1| \leq \frac{2}{\beta_q}$*

Moreover, we have

$$\lim_{\substack{\sigma_b \to 0 \\ (\sigma_b, \sigma_w) \in EOC}} \beta_q = \infty.$$

To prove this result, let us first prove a more general result.

Proposition 5 (How close is f to the identity function?)**.** *Let $\phi \in \mathcal{D}^2(\mathbb{R}, \mathbb{R}) - \{0\}$ and $(\sigma_b, \sigma_w) \in D_{\phi,var}$ with q the corresponding limiting variance. Then,*

$$\sup_{x \in [0,1]} |f(x) - x| \leq |\sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z)^2] - 1| + \frac{\sigma_w^2}{2} q \mathbb{E}[\phi''(\sqrt{q}Z)^2]$$

Proof. Using a second order Taylor expansion, we have for all $s \in [0, 1]$

$$|f(x) - f(1) - f'(1)(x - 1)| \leq \frac{(1 - x)^2}{2} \sup_{\theta \in [0,1]} |f''(\theta)|.$$

We have $f(1) = 1$. Therefore $|f(x) - x| \leq (1-x)|f'(1) - 1| + \frac{(1-x)^2}{2} \sup_{\theta\in[0,1]} |f''(\theta)|$. For $\theta \in [0,1]$, we have

$$\begin{aligned} f''(\theta) &= \sigma_w^2 q \mathbb{E}[\phi''(\sqrt{q}Z_1)\phi''(\sqrt{q}U_2(\theta))] \\ &\leq \sigma_w^2 q \mathbb{E}[\phi''(\sqrt{q}Z)^2] \\ &= \frac{\sigma_w^2}{2} q \mathbb{E}[\phi''(\sqrt{q}Z)^2] \end{aligned}$$

using Cauchy-Schwartz inequality. □

As a result, for $\phi \in \mathcal{D}^2(\mathbb{R},\mathbb{R}) - \{0\}$ and $(\sigma_b, \sigma_w) \in EOC$ with q the corresponding limiting variance, we have

$$\sup_{x\in[0,1]} |f(x) - x| \leq \frac{q\mathbb{E}[\phi''(\sqrt{q}Z)^2]}{2\mathbb{E}[\phi'(\sqrt{q}Z)^2]} = \frac{1}{\beta_q}$$

which is the first result of Proposition 4.

Now let us prove the second result for gradient backpropagation, we show that under some assumptions, our results of forward information propagation generalize to the back-propagation of the gradients. Let us first recall the results in Schoenholz et al. [2017] (we use similar notations hereafter).

Let E be the loss we want to optimize. The backpropagation process is given by the equations

$$\frac{\partial E}{\partial W_{ij}^l} = \delta_i^l \phi(y_j^{l-1})$$

$$\delta_i^l = \frac{\partial E}{\partial y_i^l} = \phi'(y_i^l) \sum_{j=1}^{N_{l+1}} \delta_j^{l+1} W_{ji}^{l+1}.$$

Although δ_i^l is non Gaussian (unlike y_i^l), knowing how $\tilde{q}_a^l = \mathbb{E}[(\delta_i^l)^2]$ changes back through the network will give us an idea about how the norm of the gradient changes. Indeed, following this approach, and using the approximation that the weights used during forward propagation are independent from those used for backpropagation, Schoenholz et al. [2017] showed that

$$\tilde{q}_a^l = \tilde{q}_a^{l+1} \frac{N_{l+1}}{N_l} \chi_1$$

where $\chi_1 = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z)^2]$.

Considering a constant width network, authors concluded that χ_1 controls also the depth scales of the gradient norm, i.e. $\tilde{q}_a^l = \tilde{q}_a^L e^{-(L-l/\xi_\Delta)}$ where $\xi_\Delta^{-1} = -\log(\chi_1)$. So in the ordered phase, gradients can propagate to a depth of ξ_Δ without being exponentially small, while in the chaotic phase, gradient explode exponentially. On the EOC ($\chi_1 = 1$), the depth scale is infinite so the gradient information can also propagate deeper without being exponentially small.
The following result shows that our previous analysis on the EOC extends to the backpropagation of gradients, and that we can make this propagation better by choosing a suitable activation function and an initialization on the EOC. We use the following approximation to ease the calculations: the weights used in forward propagation are independent from those used in backward propagation.

Proposition 6 (Better propagation for the gradient)**.** *Let a and b be two inputs and $(\sigma_b, \sigma_w) \in D_{\phi,var}$ with q the limiting variance. We define the covariance between the gradients with respect to layer l by $\tilde{q}_{ab}^l = \mathbb{E}[\delta_i^l(a)\delta_i^l(b)]$. Then, we have*

$$|\frac{\tilde{q}_{ab}^l}{\tilde{q}_{ab}^{l+1}} \times \frac{N_l}{N_{l+1}} - 1| \leq |\sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z)^2] - 1| + (1 - c_{ab}^l)\sigma_w^2 q \mathbb{E}[\phi''(\sqrt{q}Z)^2] \rightarrow_{\sigma_b \rightarrow 0} 0.$$

Proof. We have

$$\begin{aligned}
\tilde{q}_{ab}^l &= \mathbb{E}[\delta_i^l(a)\delta_i^l(b)] \\
&= \mathbb{E}[\phi'(y_i^l(a))\phi'(y_i^l(b)) \sum_{j=1}^{N_{l+1}} \delta_j^{l+1}(a) W_{ji}^{l+1} \sum_{j=1}^{N_{l+1}} \delta_j^{l+1}(b) W_{ji}^{l+1}] \\
&= \mathbb{E}[\phi'(y_i^l(a))\phi'(y_i^l(b))] \times \mathbb{E}[\delta_j^{l+1}(a)\delta_j^{l+1}(b)] \times \mathbb{E}[\sum_{j=1}^{N_{l+1}} (W_{ji}^{l+1})^2] \\
&\approx \tilde{q}_{ab}^{l+1} \frac{N_{l+1}}{N_l} \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z_1)\phi'(\sqrt{q}U_2(c_{ab}^l))] \\
&= \tilde{q}_{ab}^{l+1} \frac{N_{l+1}}{N_l} f'(c_{ab}^l).
\end{aligned}$$

We conclude using the fact that $|f'(x) - 1| \leq |f'(1) - 1| + (1 - x) f''(1)$ □

The dependence in the width of the layer is natural since it acts as a scale for the covariance. We define the gradient with respect to the l^{th} layer by $\frac{\partial E}{\partial y^l} = (\frac{\partial E}{\partial y_i^l})_{1 \leq i \leq N_l}$ and let $\tilde{Q}_{ab}^l = \mathbb{E}[\frac{\partial E}{\partial y_a^l}^T \frac{\partial E}{\partial y_b^l}]$ denote the covariance matrix of the gradients during backpropagation. Then, on the EOC, we have

$$|\frac{\text{Tr}(\tilde{Q}_{ab}^l)}{\text{Tr}(\tilde{Q}_{ab}^{l+1})} - 1| \leq (1 - c_{ab}^l) \frac{q \mathbb{E}[\phi''(\sqrt{q}Z)^2]}{\mathbb{E}[\phi'(\sqrt{q}Z)^2]} \leq \frac{2}{\beta_q}.$$

So again, the quantity $|\phi|_{EOC}$ controls the vanishing of the covariance of the gradients during backpropagation. This was expected because linear activation functions do not change the covariance of the gradients.

2 Further theoretical results

2.1 Results on the Edge of Chaos

The next lemma shows that under some conditions, the EOC does not include couples (σ_b, σ_w) with small $\sigma_b > 0$.

Lemma 5 (Trivial EOC)**.** *Assume there exists $M > 0$ such that $\mathbb{E}[\phi''(xZ)\phi(xZ)] > 0$ for all $x \in]0, M[$. Then, there exists $\sigma > 0$ such that $EOC \cap ([0, \sigma) \times \mathbb{R}^+) = \{(0, \frac{1}{|\phi'(0)|})\}$. Moreover, if $M = \infty$ then $EOC = \{(0, \frac{1}{|\phi'(0)|})\}$.*

Activation functions that satisfy the conditions of Lemma 5 cannot be used with small $\sigma_b > 0$ (note that using $\sigma_b = 0$ would lead to $q = 0$ which is not practical for the training), therefore, the result of Proposition 4 do not apply in this case. However, as we will see hereafter, SiLU (a.k.a Swish) has a partial EOC, and still allows better information propagation (Proposition 3) compared to ReLU even if σ_b not very small.

Proof. It is clear that $(0, \frac{1}{|\phi'(0)|}) \in EOC$. For $\sigma_b > 0$ we denote by q the smallest fixed point of the function $\sigma_b^2 + \frac{V[\phi]}{V[\phi']}$ (which is supposed to be the limiting variance on the EOC). Using the condition on ϕ and the fact that $\lim_{\sigma_b \to 0} q = 0$, there exists $\sigma > 0$ such that for $\sigma_b < \sigma$ we have $\mathbb{E}[\phi''(\sqrt{q}Z)\phi(\sqrt{q}Z)] > 0$. Now let us prove that for $\sigma_b \in]0, \sigma[$, the limiting variance does not satisfy the EOC equation.
Let $t_{max} = \sqrt{\sup_{x>0} |x - \frac{V[\phi]}{V[\phi']}|}$ and $\sigma_b \in]0, \min(t_{max}, \sigma)[$. Recall that for all $x \geq 0$ we have that

$$F'(x) = \sigma_w^2(\mathbb{E}[\phi'(\sqrt{x}Z)^2] + \mathbb{E}[\phi''(\sqrt{x}Z)\phi(\sqrt{x}Z)])$$

Using $\sigma_w^2 = 1/V[\phi'](q)$ (EOC equation) we have that $F'(q) = 1 + \sigma_w^2 \mathbb{E}[\phi''(\sqrt{q}Z)\phi(\sqrt{q}Z)]) > 1$. Therefore, the function $\sigma_b^2 + \frac{1}{V[\phi'](q)} V[\phi]$ crosses the identity in a point $\hat{q} < q$, hence $(\sigma_b, \sigma_w) \notin D_{\phi,var}$. Therefore, for any $\sigma_b \in]0, \sigma[$, there is no σ_w such that $(\sigma_b, \sigma_w) \in EOC$.

If $M = \infty$, the previous analysis is true for any $\sigma > 0$, by taking the limit $\sigma \to \infty$, we conclude. □

This is true for activations such as Shifted Softplus (a shifted version of Softplus in order to have $\phi(0) = 0$) and SiLU (a.k.a Swish).

Corollary 1. *$EOC_{SSoftplus} = \{(0,2)\}$ and there exists $\sigma > 0$ such that $EOC_{SiLU} \cap ([0, \sigma[\times \mathbb{R}^+) = \{(0,2)\}$*

Proof. let $s(x) = \frac{1}{1+e^{-x}}$ for all $x \in \mathbb{R}$ (sigmoid function).

1. Let $sp(x) = \log(1+e^x) - \log(2)$ for $x \in \mathbb{R}$ (Shifted Softplus). We have $sp'(x) = s(x)$ and $sp''(x) = s(x)(1-s(x))$. For $x > 0$ we have

$$\begin{aligned}\mathbb{E}[sp''(xZ)sp(xZ)] &= \mathbb{E}[s(xZ)(1-s(xZ))sp(xZ)] \\ &= \mathbb{E}[1_{Z>0}(s(xZ)(1-s(xZ))sp(xZ))] + \mathbb{E}[1_{Z<0}(s(xZ)(1-s(xZ))sp(xZ))] \\ &= \mathbb{E}[1_{Z>0}(s(xZ)(1-s(xZ))sp(xZ))] + \mathbb{E}[1_{Z<0}(s(xZ)(1-s(xZ))sp(-xZ))] \\ &= \mathbb{E}[1_{Z>0}(s(xZ)(1-s(xZ))(sp(xZ)+sp(-xZ)))] > 0,\end{aligned}$$

where we have used the fact that $sp(y) + sp(-y) = \log(\frac{2+e^y+e^{-y}}{4}) > 0$ for all $y > 0$. We conclude using Lemma 5.

2. Let $si(x) = xs(x)$ (SiLU activation function, known also as Swish). We have $si'(x) = s(x) + xs(x)(1-s(x))$ and $si''(x) = s(x)(1-s(x))(2 + x(1-2s(x)))$. Using the same technique as for SSoftplus, we have for $x > 0$

$$\begin{aligned}\mathbb{E}[si''(xZ)si(xZ)] &= \mathbb{E}[xZ \times s(xZ)^2 \times (1-s(xZ))(2 + xZ(1-2))] \\ &= \mathbb{E}[1_{Z>0}G(xZ)],\end{aligned}$$

where $G(y) = ys(y)(1-s(y))(2+y(1-2s(y)))(2s(y)-1)$. The only term that changes sign is $(2+y(1-2s(y)))$. It is positive for small y and negative for large y. We conclude that there $M > 0$ such that $\mathbb{E}[si''(xZ)si(xZ)] > 0$ for $x \in]0, M[$.

□

2.2 Beyond the Edge of Chaos

Can we make the distance between f and the identity function small independently from the choice of σ_b? The answer is yes if we select the right activation function. Let us first define a semi-norm on $\mathcal{D}^2(\mathbb{R}, \mathbb{R})$.

Definition 2 (EOC semi-norm). *The semi-norm $|.|_{EOC}$ is defined on $\mathcal{D}^2(\mathbb{R}, \mathbb{R})$ by $|\phi|_{EOC} = \sup_{y \in \mathbb{R}^+} \frac{y\mathbb{E}[\phi''(\sqrt{y}Z)^2]}{\mathbb{E}[\phi'(\sqrt{y}Z)^2]}$.*
$|.|_{EOC}$ is a norm on the quotient space $\mathcal{D}^2(\mathbb{R}, \mathbb{R})/\mathcal{L}(\mathbb{R})$ where $\mathcal{L}(\mathbb{R})$ is the space of linear functions.

When $|\phi|_{EOC}$ is small, ϕ is close to a linear function, which implies that the function $\frac{V[\phi]}{V[\phi']}$ defined on $\mathbb{R}^+$ is close to the identity function. Thus, for a fixed σ_b, we expect q to become arbitrarily big when $|\phi|_{EOC}$ goes to zero.

Lemma 2.1. *Let $(\phi_n)_{n\in\mathbb{N}}$ be a sequence of functions such that $\lim_{n\to\infty}|\phi_n|_{EOC} = 0$. Let $\sigma_b > 0$ and assume that for all $n \in \mathbb{N}$ there exists $\sigma_{w,n}$ such that $(\sigma_w, \sigma_{w,n}) \in EOC$. Let q_n be the limiting variance. Then $\lim_{n\to\infty} q_n = \infty$*

Proof. The proof is straightforward knowing that $f(0) \leq \frac{1}{2}|\phi_n|_{EOC}$, which implies that $\frac{\sigma_b^2}{q} \leq \frac{1}{2}|\phi_n|_{EOC}$. □

Corollary 2.1. *Let $\phi \in \mathcal{D}^2(\mathbb{R},\mathbb{R}) - \{0\}$ and $(\sigma_b, \sigma_w) \in EOC$ with q the corresponding limiting variance. Then,*

$$\sup_{x\in[0,1]} |f(x) - x| \leq \frac{1}{2}|\phi|_{EOC}.$$

Corollary 2.1 shows that by taking an activation function ϕ such that $|\phi|_{EOC}$ is small and by initializing the network on the EOC, the correlation function is close to the identity function, i.e., the signal propagates deeper through the network. However, note that there is a trade-off to take in account here: we loose expressiveness by taking $|\phi|_{EOC}$ too small, because this would imply that ϕ is close to a linear function. So there is a trade-off between signal propagation and expressiveness We check this finding with activation functions of the form $\phi_\alpha(x) = x + \alpha\text{Tanh}(x)$. Indeed, we have $|\phi_\alpha|_{EOC} \leq \alpha^2 \sup_{y\in\mathbb{R}^+} \mathbb{E}[\text{Tanh}''(\sqrt{x}Z)^2] \to_{\alpha\to 0} 0$. So by taking small α, we would theoretically provide deeper signal propagation. However, note that we loose expressiveness as α goes to zero because ϕ_α becomes closer to the identity function. So There is also a trade-off here. The difference with Proposition 4 is that here we can compensate the expressiveness issue by adding more layers (see e.g. Montufar et al. [2014] who showed that expressiveness grows exponentially with depth).

3 Experiments

3.1 Training with RMSProp

For RMSProp, the learning rate 10^{-5} is nearly optimal for networks with depth $L \leq 200$ (for deeper networks, 10^{-6} gives better results). This learning rate was found by a grid search with exponential step of size 10.

Figure 1 shows the training curves of ELU, ReLU and Tanh on MNIST for a network with depth 200 and width 300. Here also, ELU and Tanh perform better than ReLU.

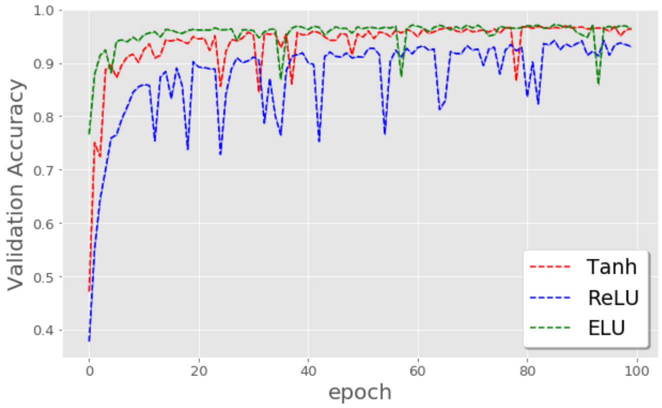

Figure 1: 100 epochs of the training curves of ELU, ReLU and Tanh networks of depth 200 and width 300 on MNIST with RMSProp

This confirms that the result of Proposition 3 is independent of the training algorithm. ELU has faster convergence than Tanh. This could be explained by the saturation problem of Tanh.

3.2 Training with activation $\phi_\alpha(x) = x + \alpha\textbf{Tanh}(x)$

As we have already mentioned, ϕ_α satisfies all conditions of Proposition 3. Therefore, we expect it to perform at least better than ReLU for deep neural networks. Figure 2 shows the training curve for width 300 and depth 200 with different activation functions. $\phi_{0.5}$ has approximately similar performance as ELU and better than Tanh and ReLU. Note that ϕ_α does not suffer form saturation of the gradient, which could explain why it performs better than Tanh.

3.3 Impact of $\phi''(0)$

Since we usually take σ_b small on the EOC, then having $\phi''(0) = 0$ would make the coefficient β_q even bigger. We test this result on SiLU (a.k.a Swish) for depth 70. SiLU is defined by

$$\phi_{SiLU}(x) = x \text{ sigmoid}(x)$$

we have $\phi''(0) = 1/2$. consider a modified SiLU (MSiLU) defined by

$$\phi_{MSiLU}(x) = x \text{ sigmoid}(x) + (e^{-x^2} - 1)/4$$

We have $\phi''_{MSiLU}(0) = 0$.
Figure 3 shows the the training curves (test accuracy) of SiLU and MSiLU on MNIST

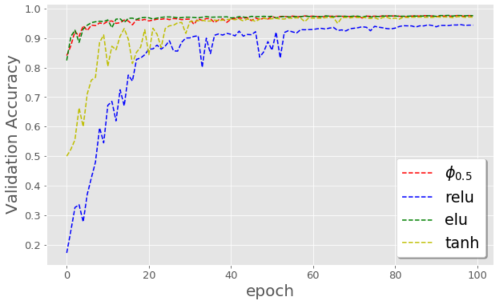

Figure 2: 100 epochs of the training curves of ELU, ReLU, Tanh and $\phi_{0.5}$ networks of depth 200 and width 300 on MNIST with SGD

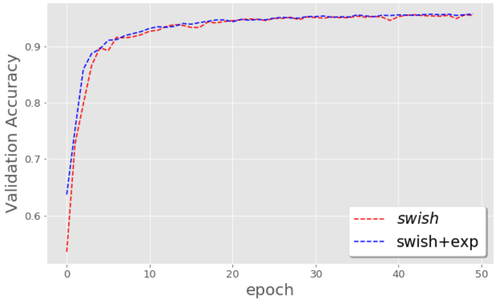

Figure 3: 50 epochs of the training curves of SiLU and MSiLU on MNIST with SGD

with SGD. MSiLU performs better than SiLU, expecially at the beginning of the training.

References

S.S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. *5th International Conference on Learning Representations*, 2017.

G.F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, 27: 2924–2932, 2014.