Strongly Self-Normalizing Neural Networks with Applications to Implicit Representation Learning

June 15, 2021

Abstract

Recent studies have shown that wide neural networks with orthogonal linear layers and Gaussian Poincaré normalized activation functions avoid vanishing and exploding gradients for input vectors with the correct magnitude. This paper introduces a strengthening of the condition that the activation function must be Gaussian Poincaré normalized which creates robustness to deviations from standard normal distribution in the pre-activations, thereby reducing the dependence on the requirement that the network is wide and that the input vector has the correct magnitude. In implicit representation learning this allows the training of deep networks of this type where the linear layers are no longer constrained to be orthogonal linear transformations. Networks of this type can be fitted to a reference image in 1/100th the number of iterations required by previous methods and to higher accuracy.

1 Introduction

Sitzmann et al.[2] have proposed the use of periodic activation functions, specifically the use of the sine in the fitting of neural networks to, for example, images and signed distance functions. They prove that if the weights are uniformly distributed in $[-c, c]$ where $c = \sqrt{6/f}$ where f is the fan in, then the pre-activations are always standard normal distributed irrespective of the depth of the network. Importantly, the derivatives of these networks are networks of the same type (SIRENs, or sinusoidal representation networks), ensuring that these guarantees are applicable also to derivatives of networks of this type.

However, this provides no guarantees about the derivatives with respect to any parameter.

Lu et al.[1] have proposed a type of neural network, BSNNs (bidirectionally self-normalizing neural networks), for which they prove a similar result: provided that the layers are orthogonal linear transformations that are uniformly distributed on the orthogonal group in the Haar sense followed by activation functions that are Gaussian-Poincaré normalized, meaning that the activation function f satisfies $\mathbb{E}[f(Z)^2] = 1$ and $\mathbb{E}[f'(Z)^2] = 1$ where Z is the standard normal

distribution, f and its derivative are Lipschitz continuous, and the input vector is thin-shell concentrated in the sense that for all $\epsilon > 0$ $\mathbb{P}\left\{|\frac{1}{n}\|x\|_2^2 - 1| > \epsilon\right\} \to 0$ as $n \to \infty$ then the squared magnitude of the input to each layer is n and the magnitude of the derivative of any loss function E with respect to the input to any layer is the same provided that the layers are wide.

The guarantee that a thin-shell concentrated vector has its norm preserved under forward propagation in a BSNN is comparable to the guarantee that in a SIREN the pre-activations are standard normal distributed, but that the derivative of a loss function with respect to the input to any layer always has the same magnitude is an additional guarantee, which when taken together with the first becomes a much stronger assurance with regard to the trainability of deep BSNNs.

In SIRENs the width of the network typically changes in the first and final layers. Consequently a straightforward analogue of a SIREN where all linear layers are orthogonal linear transformations is impossible. We consider a condition under which a BSNN with orthogonal initialization but no orthogonality constraint performs well in implicit representation learning. Among the BSNNs satisfying this condition is one all of whose derivatives are networks of the same type.

2 Motivation

In a BSNN of some fixed finite width the pre-activations are uniformly distributed on the sphere with radius equal to the square root of the dimension, which in high dimension in a certain sense closely approximates the normal distribution so that the trainability guarantees of Lu et al.[1] can be obtained about the distribution of the activations even though the condition they impose is on the expectation of the activation function and its derivatives applied to a standard normal distributed random variable.

It is possible to impose a stronger condition, that $\mathbb{E}[f(W)^2] = \mathbb{E}[f'(W)^2] = 1$ for all distributions W such that $W \sim -W$. This is the case precisely if $f(x)^2 - 1$ and $f'(x)^2 - 1$ are odd functions.

We consider the ideal case where vectors with magnitude exactly equal to the square root of dimension are forward propagated through neural networks of different width with different activation functions, all satisfying that $\mathbb{E}[f(Z)^2] = 1$ where Z is the standard normal distribution and some that $f(x)^2 - 1$ is an odd function.

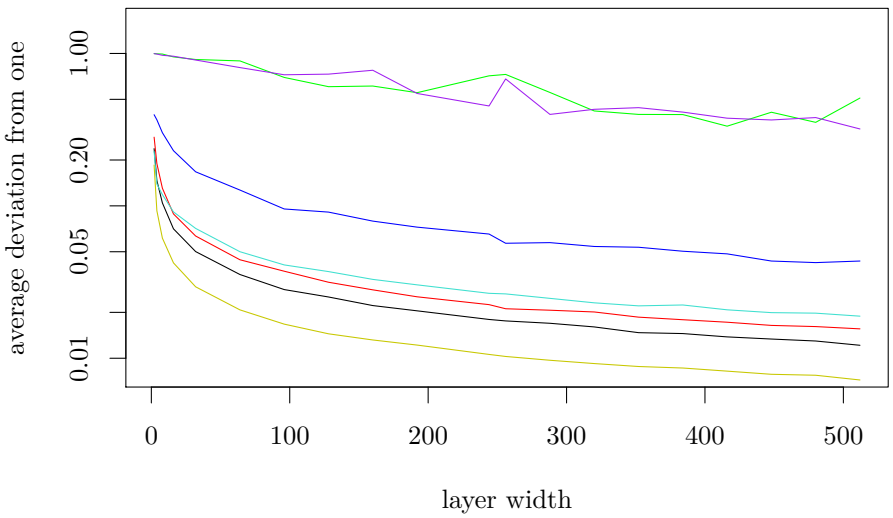

Figure 1: Log plot of $\sum_k |\frac{\|x_k\|}{\sqrt{n}} - 1|$ where x_k is the pre activations over 400 layers, averaged over ten runs, as a function of the network width. Gold: $\sqrt{2/(1+e^{-x})}$, Black: $\sqrt{2}\sin(x+\pi/4)$, Red: GP normalized tanh, Turquoise: GP normalized GELU, Blue: GP normalized ELU, Purple: GP normalized leaky ReLU, Green: GP normalized ReLU.

The functions giving the two lowest average deviation curves (fig. 1) both satisfy the strong condition. In the case of the function achieving the lowest average, $\sqrt{2/(1+e^{-x})}$, without satisfying the part of the Gaussian Poincaré normalization condition that concerns the derivative, so that the guarantees of Lu et al.[1] with regard to backward propagation do not hold.

What the stronger condition assures is not that the magnitudes are preserved more exactly, but that the magnitude of a vector whose magnitude deviates from the square root of the dimension is brought closer to that magnitude.

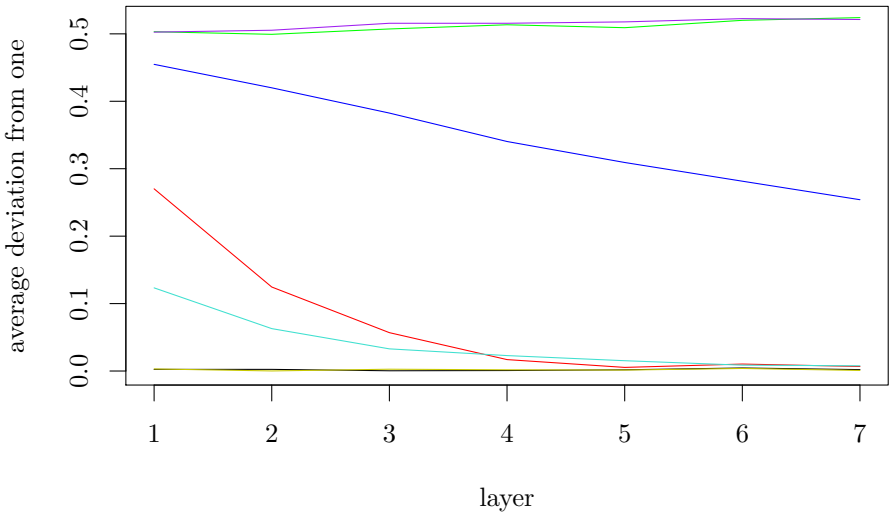

Figure 2: $|\frac{\frac{1}{100}\sum_{l=1}^{100}\|x_k^{(l)}\|}{\sqrt{n}} - 1|$ for k=1,..,7 where $x_k^{(l)}$ are the pre-activations for layer k during forward propagation of $x_0^{(l)}$ and $x_0^{(l)}$ is a random normal vector transformed by normalization to have norm $\frac{\sqrt{n}}{2}$ where n is the dimension. Colour corresponds to activation function. Gold: $\sqrt{2/(1+e^{-x})}$, Black: $\sqrt{2}\sin(x + \pi/4)$, Red: GP normalized tanh, Turquoise: GP normalized GELU, Blue: GP normalized ELU, Purple: GP normalized leaky ReLU, Green: GP normalized ReLU.

We will be concerned with relaxations of networks of the following type:

Definition 1 (Strongly bidirectionally self-normalizing neural networks)**.** *A neural network is said to be a SBSNN if it has orthogonal weight matrices that are uniformly distributed in the Haar sense and the activation is differentiable, Lipschitz continuous with Lipschitz continuous derivative and has the property that $f(x)^2 - 1$ and $f'(x)^2 - 1$ are odd functions.*

3 Characterization of SBSNNs

Lemma 1. *If $f(x)^2 - 1$ and $f'(x)^2 - 1$ are odd functions then there is an even function w taking values in $\{1, -1\}$ such that $f(x)^2 - 1 = \sin(2\int_0^x w(s)ds)$.*

Proof. Let $u(x) = f(x)^2 - 1$ and $v(x) = f'(x)^2 - 1$, then these are odd functions.

Since v is odd $v(x) = -v(-x)$, i.e. $f'(x)^2 - 1 = -(f'(-x)^2 - 1)$ so that $f'(x)^2 + f'(-x)^2 = 2$.

Also, $u'(x) = 2f(x)f'(x)$ so that $u'(x)^2 = 4f(x)^2 f'(x)^2$. Because $u(x) = f(x)^2 - 1$ $f(x)^2 = u(x) + 1$, so that we can write $u'(x)^2 = 4(u(x) + 1)f'(x)^2$.

We now multiply $f'(x)^2 + f(-x)^2 = 2$ by $4(u(x) + 1)f'(x)^2 \cdot 4(u(-x) + 1)f'(-x)^2$ and simplifying using the definition of u' we obtain that $4(u(-x) + 1)u'(x)^2 + 4(u(-x) + 1)^2 = 2 \cdot 4(u(x) + 1) \cdot 4(u(-x) + 1)$.

Further simplifying and using that u is an odd function we obtain $4u'(x)^2(u(x)+1-u(x)+1) = 2\cdot 4^2(1-u(x)^2)$. Further simplifying we obtain $u'(x)^2 = 4(1-u(x)^2)$. Taking the square root we obtain $|u'(x)| = 2\sqrt{1-u(x)^2}$.

Let $\bar{w}(x) = \operatorname{sgn}(u'(x))$. Since u' is the derivative of an odd function it is even and $\bar{w}$ is even.

Now $u'(x) = 2w(x)\sqrt{1-u(x)^2}$, consequently $\frac{du}{\sqrt{1-u(x)^2}} = 2\bar{w}(x)dx$ and $\arcsin(x)+C = 2\int_0^x \bar{w}(s)ds$ for some constant C. Thus $u(x) = \sin(2\int_0^x \bar{w}(s)ds - C)$. Because u is odd it is necessary that $u(0) = 0$ and thus that $C = \pi \cdot n$. Since $\sin(x+\pi) = -\sin(x)$ for all x $u(x) = \sin(2S\int_0^x \bar{w}(s)ds)$ with $S \in \{1, -1\}$ and by setting $w(x) = S\bar{w}(x)$ the conclusion holds. □

Theorem 1. *f is an activation function of an SBSNN, i.e. such that f is differentiable, Lipschitz continuous, has a Lipschitz continuous derivative and satisfies that $f(x)^2 - 1$ and $f'(x)^2 - 1$ are odd functions, precisely if $f(x) = \pm\sqrt{2}\sin(x+\pi/4)$ or $f(x) = \pm\sqrt{2}\cos(x+\pi/4)$.*

Proof. By lemma 1 there is an even function w taking values in $\{1, -1\}$ such that $f(x)^2 - 1 = \sin(2\int_0^x w(s)ds)$. Let $I(x) = \int_0^x w(s)ds$. Then $f(x)^2 = 1 + \sin(2I(x)) = \cos(I(x))^2 + \sin(I(x))^2 + 2\sin(I(x))\cos(I(x)) = (\cos(I(x)) + \sin(I(x)))^2 = (\sin(I(x)+\pi/4))^2$. Thus there exists some function $S(x)$ such that $f(x) = S(x)\sqrt{2}\sin(\int_0^x w(s)ds + \pi/4)$ where $S(x)$ takes values in $\{1, -1\}$.

Since S takes values in $\{1, -1\}$ and $\sqrt{2}\sin(\int_0^x w(s)ds + \pi/4)$ is continuous, if S jumps at any input where $\sin(\int_0^x w(s)ds + \pi/4)$ is not zero, then $f(x)$ is not continuous. Consequently $S(x)$ changes sign only at points where $\sin(\int_0^x w(s)ds + \pi/4) = 0$.

Thus at points where $\sin(\int_0^x w(s)ds) \neq 0$ $f'(x) = S(x)\sqrt{2}\cos(\int_0^x w(s)ds + \pi/4)w(x)$. Continuity of $f'(x)$ then requires that $S(x)$ jumps precisely where $w(x)$ jumps.

Thus $w(x)$ and $S(x)$ may only jump at point where $\sin(\int_0^x w(s)ds + \pi/4) = 0$ and if they jump at such a point they must jump together. Consequently we can write $S(x) = Cw(x)$ where $C \in \{1, -1\}$. Thus $f'(x) =$
$= \sqrt{2}Cw(x)\cos(\int_0^x w(s)ds + \pi/4)$. If w jumps at x, then since $\int_0^x w(s)ds = 0$ $\cos(\int_0^x w(s)ds) = 1$, so f' jumps at x, and f' is not continuous.

Thus w is constant and $f(x) = C\sin(Wx + \pi/4)$ where $C, W \in \{1, -1\}$.

The reverse direction is trivial. □

Lipschitz continuity of derivatives and the activation function and its derivative is a requirement of the theory of Lu et al.[1], but Gaussian Poincaré normalized ReLU activation functions, which do not have a derivative everywhere and which have a derivative which is not Lipschitz continuous still behave largely according to the theory, with wide Gaussian Poincaré normalized ReLU networks preserving norms in practice (fig. 1). Consequently there is reason to relax the continuity conditions so as to admit functions for which these guarantees do not hold but which are still in the spirit of the theory. For this reason we consider the following theorem:

Theorem 2. *Let w be an even function taking values in $\{1,-1\}$ such that $\partial \int_0^x w(s)ds$ exists where ∂ is a left- or right derivative and S any function taking values in $\{1,-1\}$ with jumps only when $\sin(\int_0^x w(s)ds + \pi/4) = 0$, then $f(x) = \sqrt{2}S(x)\sin(\int_0^x w(s)ds + \pi/4)$ is such that $f(x)^2 - 1$ and $(\partial f)(x)^2 - 1$ are odd functions.*

Proof. Let as before $I(x) = \int_0^x w(s)ds$. Then $I(x)$ is odd.

$f(x)^2 - 1 = 2\sin^2(I(x) + \pi/4) - 1 = 2\left(\frac{1}{\sqrt{2}}(\sin(I(x)) + \cos(I(x)))\right)^2 - 1 =$
$= 1 + 2\sin(I(x))\cos(I(x)) - 1 = 2\sin(I(x))\cos(I(x))$ is an odd function since $I(x)$ and $\sin(x)\cos(x)$ are odd functions.

Let $g(x) = S(x)\sqrt{2}\sin(x + \pi/4)$. Because $S(x)$ jumps only when $\sin(I(x) + \pi/4)$ is zero ∂g exists and is $\lim_{a\uparrow x} S(a)\sqrt{2}\cos(x + \pi/4)$ in the case of the left derivative and $\lim_{a\downarrow x} S(a)\sqrt{2}\cos(x + \pi/4)$ in the case of the right derivative.

Similarly, since $(\partial I)(x)$ exists it is $\lim_{a\uparrow x} w(a)$ in the case when ∂ is the left derivative and $\lim_{a\downarrow x} w(a)$ in the case of the right derivative.

Consequently $\partial f = \partial(g \circ I)$ exists and is $(\partial g)(I(x))(\partial I)(x)$. $(\partial I)(x)$ is in $\{1,-1\}$ so $((\partial f)(x))^2 = (\partial g)(I(x))^2 = (\lim_{a\uparrow x} S(a)\sqrt{2}\cos(I(x) + \pi/4))^2 = 2\cos^2(I(x) + \pi/4)$.

Consequently $(\partial f)(x)^2 - 1 = 2\cos^2(I(x) + \pi/4)\bar{w}(x)^2 - 1 =$
$= 2(\frac{1}{\sqrt{2}}(\cos(I(x)) - \sin(I(x)))^2 - 1 = -2\cos(I(x))\sin(I(x))$, which is an odd function. □

We visualize two activation functions in this class along with a sinusoidal-type SBSNN:

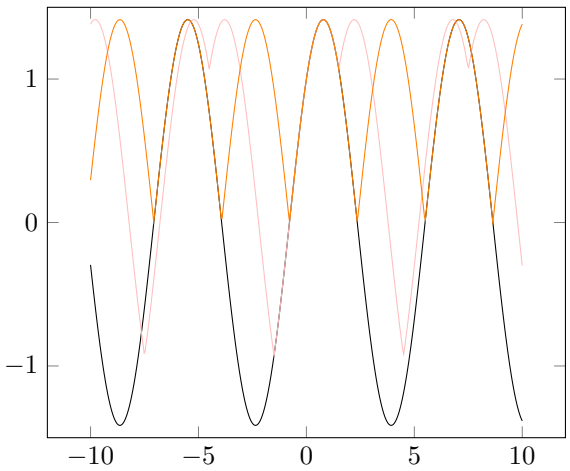

Figure 3: Examples of functions such that $f(x)^2 - 1$ and $(\partial f(x))^2 - 1$ are odd functions, where ∂ is the derivative if it exists, or a one-sided derivative if it does not. Black: $\sqrt{2}\sin(x + \pi/4)$, Pink: $\sqrt{2}\sin(T(x) + \pi/4)$ where T is an odd triangle wave with amplitude 1.5 and period 6, Orange: $\sqrt{2}|\sin(x + \pi/4)|$.

4 All derivatives of SBSNNs are SBSNNs

The derivative of an activation function of an SBSNN is again an SBSNN. As was shown in the previous sections, these functions are $f(x) = S\sqrt{2}\sin(wx + \pi/4)$ where $S, w \in \{1, -1\}$. Knowing that these functions and their derivatives are Lipschitz continuous we it is sufficient to prove the following theorem, which we consider more directly, in terms of distributions, instead of even and odd functions (recall that $\mathbb{E}[f(W)^2 - 1]$ for all W such that $W \sim -W$ iff $f(x)^2 - 1$ is an odd function):

Theorem 3. *Let W and $-W$ have the same distribution and $f(x) = S\sqrt{2}\sin(\frac{\pi}{4} + wx)$, then $\mathbb{E}[f^{(n)}(W)] = 1$ for all n.*

Proof. If g is an odd function, i.e. $g(x) = -g(-x)$ then $\mathbb{E}[g(W)] = \mathbb{E}[-g(-W)] = -\mathbb{E}[g(-W)]$. Since $-W$ has the same distribution as W it follows that $\mathbb{E}[g(W)] = -\mathbb{E}[g(W)] = 0$.

Because $f'' = -f$, it is sufficient to check that $\mathbb{E}[f(W)^2] = \mathbb{E}[f'(W)^2] = 1$ to ensure that $\mathbb{E}[f^{(n)}(W)] = 1$ for all n.

For the first equality $\mathbb{E}[(\sin(Z) + \cos(Z))^2] = \mathbb{E}[\sin^2(Z) + 2\sin(Z)\cos(Z) + \cos(Z)^2] = \mathbb{E}[1 + 2\cos(Z)\sin(Z)] = 1$, since the function $\cos(x)\sin(x)$ is odd. For second equality $\mathbb{E}[(\cos(Z) - \sin(Z))^2] = \mathbb{E}[\cos(Z)^2 - 2\sin(Z)\cos(Z) + \sin^2(Z)] = \mathbb{E}[1 - 2\sin(Z)\cos(Z)] = 1$, again since the the function $\sin(x)\cos(x)$ is odd. □

5 Ablation: fitting SBSNNs to images without using positional encoding

Sitzmann et al.[2] have demonstrated that a SIREN can be fitted to the 256×256 cameraman reference image to a MSE of 10^{-5} using 10000 iterations, demonstrating its benefits relative to tanh, ReLU MLPs and over ReLU network having positional encoders and ReLU with initial radial basis function layers on this problem, none of which can achive MSE higher than 10^{-4}.

This can be exceeded by a SBSNNs without an positional encoder and instead an initial SIREN layer in around 500 iterations using a training schedule involving reduction of an initial learning rate of $2 \cdot 10^{-4}$ by half when the mean square error during training has a plateau lasting 40 epochs (fig. 4).

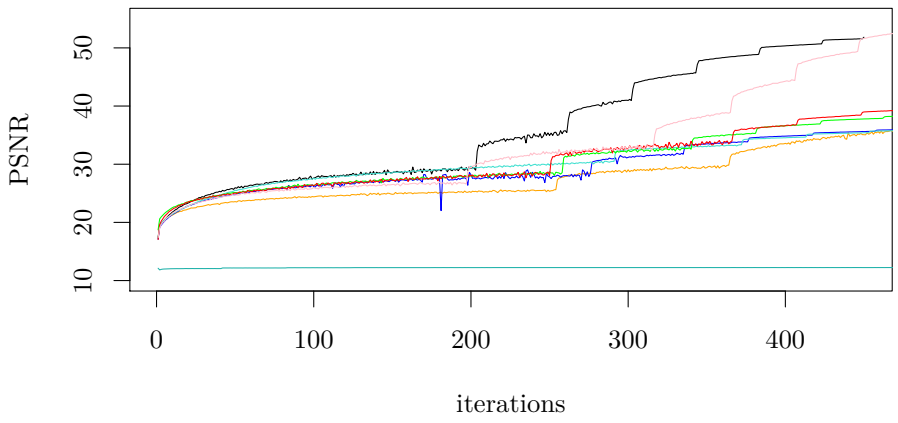

Figure 4: PSNR during training of network with five hidden layers, width 192, orthogonal initialization and an initial SIREN layer. Black: $\sqrt{2}\sin(x+\pi/4)$, Pink: $\sqrt{2}\sin(T(x)+\pi/4)$ (triangle wave), Red: GPN Tanh, Green: GPN ReLU, Blue: GPN ELU, Orange: $\sqrt{2}|\sin(x+\pi/4)|$, Sea green: SIREN (special initialization).

6 Fitting images using positional encoders

Positional encoders can be seen to allow higher accuracy and faster fitting.

Mildenhall et al. [3] describe a positional encoder assigning to each co-ordinate the vector $(\sin(2^0\pi p), \cos(2^0\pi p), ..., \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p))$ where p is the co-ordinate. These vectors have squared magnitude equal to one half their dimension, and by scaling them by $\sqrt{2}$ we obtain a vector of the required magnitude.

The method of Mildenhall et al. was originally applied to five-dimensional input, but when applied to two-dimensional input it introduces obvious patterns in the form of correlations between pixels along lines where one co-ordinate is constant.

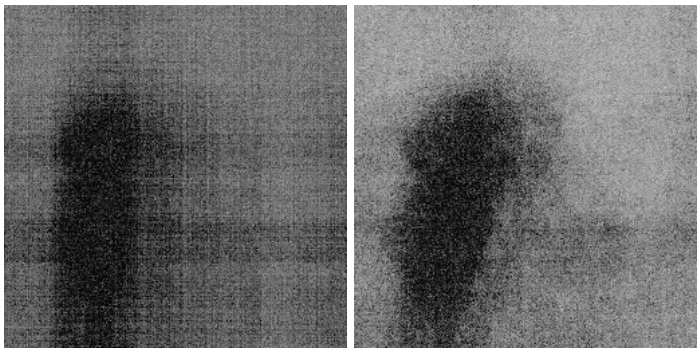

Figure 5: Early frame during fitting. Left: Using a scaled encoder of the Mildenhall et al. type, Right: Using the modified encoder. The image being fitted is the Cameraman test image.

These patterns can be removed by a slight modification of the encoder of Mildenhall et al.: taking the input co-ordinate pair $(x, y)^T$ we construct two additional co-ordinate pairs by rotating the original co-ordinate pair by one third and two-thirds a turn around the origin to obtain two more $D_{2\pi/3}(x, y)^T$, $D_{4\pi/3}(x, y)^T$ where D_θ denotes a rotation in the x-y plane by θ, and applying the encoder of Mildenhall et al. to each and concatenating the three outputs. Use of this positional encoder avoids line artefacts early (fig. 5) in the network fitting and can be seen to improve final mean squared error (fig. 6 and fig. 7).

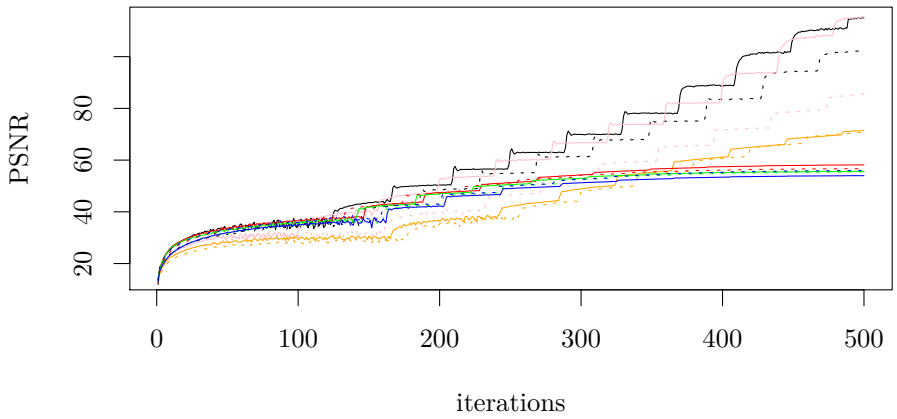

Figure 6: PSNR during training of a network with five hidden layers of width 192 and orthogonal initialization. Dotted lines use Mildenhall et al. type encoders. Solid coloured lines use Mildenhall et al. type encoder applied to three rotated vectors. Activation function is by colour. Black: $\sqrt{2}\sin(x + \pi/4)$, Pink: $\sqrt{2}\sin(T(x) + \pi/4)$ where T is an odd triangle wave with amplitude 1.5 and period 6, Orange: $\sqrt{2}|\sin(x + \pi/4)|$, Red: GP normalized tanh, Turquoise: GP normalized GELU, Blue: GP normalized ELU, Green: GP normalized ReLU.

With more aggressive learning rates this permits very fast fitting to accuracy exceeding that achievable with SIREN in around 100 iterations, giving the claim of the abstract (fig. 7).

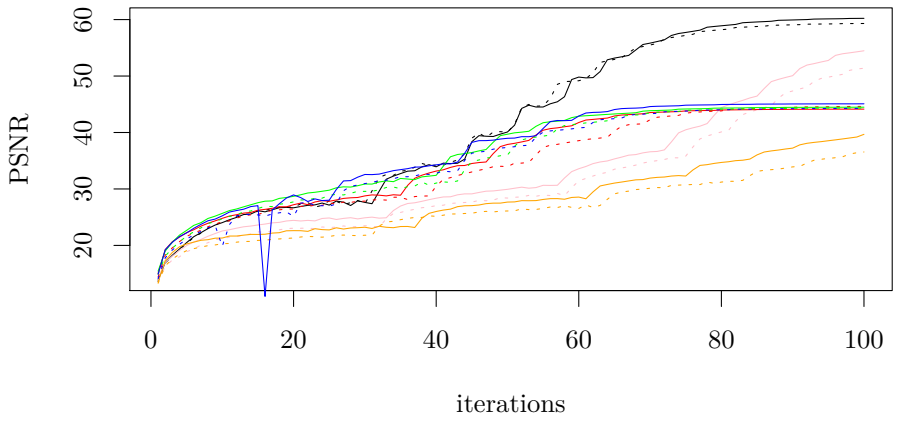

Figure 7: PSNR during training of a network with five hidden layers of width 192 and orthogonal initialization using aggressive learning rate reduction. Dotted lines indicate use of a Mildenhall et al. type encoder, while solid lines correspond to the rotated encoder. Colour indicates activation function. Black: $\sqrt{2}\sin(x + \pi/4)$, Pink: $\sqrt{2}\sin(T(x) + \pi/4)$ where T is an odd triangle wave with amplitude 1.5 and period 6, Orange: $\sqrt{2}|\sin(x+\pi/4)|$, Red: GP normalized tanh, Turquoise: GP normalized GELU, Blue: GP normalized ELU, Green: GP normalized ReLU.

References

[1] Lu, Y. Gould. S. & Ajanthan T. (2020) Bidirectionally Self-Normalizing Neural Networks *arXiv preprint arXiv:2006.12169*.

[2] Sitzmann, V. Martel, J.N.P., Bergman, A.W., Lindell, D.B. & Wetzstein, G. (2020) Implicit Neural Representations with Periodic Activation Functions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan M.F. and Lin, H. (eds.) *Advances in Neural Information Processing Systems 33*, pp. 7462–7473. Curran Associates, Inc.

[3] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. & Ng R. (2020) NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Vedaldi, A., Bischof, H., Brox, Th. and Frahm, J.M. (eds.) *Computer Vision – ECCV 2020, 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pp. 405–421, Springer Verlag