SIREN is characterized by the use of a sine activation function with a weight initialization where synaptic weights are chosen independently from a uniform distribution $\mathcal{U}(-c/\sqrt{n}, c\sqrt{n})$. A SIREN neuron is intended to give Arcsine$(-1,1)$ distributed output for pre-activations which are $\mathcal{U}([-1,1])$ distributed in the case of the first layer or $N(0, c^2/6)$-distributed where $c = \sqrt{6}$ in the case of the later layers.

In both the case where the pre-activations are uniformly distributed and where they are normal distributed there are problems with the justification of the SIREN initialization, although it will ultimately be found that the SIREN initialization is still essentially reasonable.

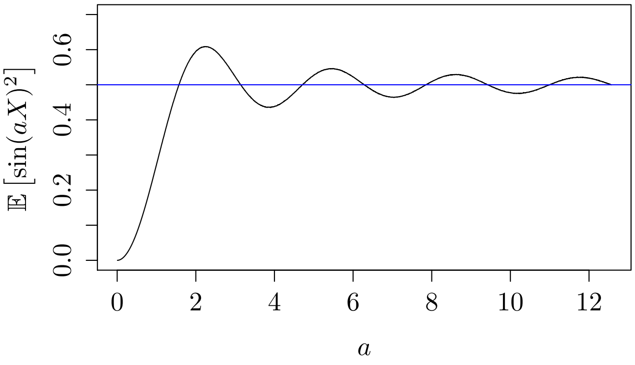
However, this necessitates reconsidering some of the assumptions about what constitutes a good choice of parameters in the weight initialization of SIRENs that lead to an improvement in the kind of PSNR that can be achieved.

Furthermore, independently of this the method described by Sitzmann et al. can be sped up greatly by reducing the learning rate during training. In all experiments we train use an initial learning rate of $2 \cdot 10^{-4}$ with halving after MSE encounters plateaus lasting 40 epochs.

Low-variance initializations for SIREN of the type used by Sitzmann et al. do however remain relevant because of their good performance before learning rate reduction.

Forward propagation in SIRENs

In relation to the case where the pre-activations are uniformly distributed it has been claimed that when $X \sim \mathcal{U}(-1,1)$ $\sin(aX + b)$ will be Arcsine$(-1,1)$ distributed irrespective of b. This is only approximately true. The variance of a Arcsine$(-1,1)$ distribution is $1/2$. There is a change in behaviour at $\pi/2$, where the variance finally reaches $1/2$ and for values greater than $\pi/2$ the variance remains close to $1/2$, but it is in fact not $1/2$ in general and the difference is not altogether small, unless c is large, as can be seen from the graph below. The blue line shows $y = 1/2$. This can be resolved either by scaling the initial uniform distribution, thus causing the input to be the $\mathcal{U}[-\pi/2, \pi/2]$ distributed instead of $\mathcal{U}[-1,1]$, by using a precisely scaled activation function, $\sin(a\cdot)$ where $a = \frac{\pi}{2}$ instead of $\sin(\cdot)$, or by using a heavily scaled activation function $\sin(a\cdot)$ where a might be 30. This last approach is what has been proposed by Sitzmann et al. for use in practice, while they use $a = \pi/2$ in a theoretical analysis.



The normal distribution of the pre-act ivations that occurs in the layers following the first arises as follows:

The activations from the previous layer are assumed to be Arcsine$(-1, 1)$ distributed and because of this they have mean zero and variance 1/2.

The synaptic weights are independent of the activations of the previous layer and are $\mathcal{U}(-c/\sqrt{n}, c/\sqrt{n})$ distributed. This distribution has mean zero and variance $\frac{1}{12}(2c/\sqrt{n})^2 = \frac{1}{3}c^2/n$.

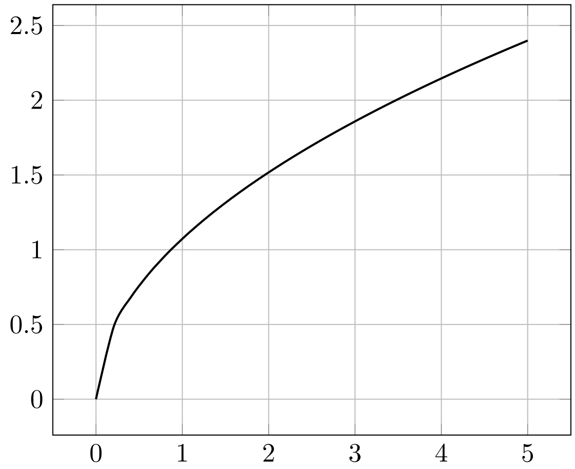
The pre-activations for the new layer are the dot product of the previous activations and the synaptic weights of the new layer. The mean of the this is zero and the variance of the product of a single activation X_i and its corresponding synaptic weight W_i is

$$\begin{aligned}
\operatorname{Var}[W_iX_i] &= \mathbb{E}\left[(W_iX_i - E[W_iX_i])^2\right] = \\
&= \mathbb{E}\left[W_i^2X_i^2 - 2W_iX_i\mathbb{E}\left[W_iX_i\right] + E[W_iX_i]^2\right] = \\
&= \mathbb{E}\left[W_i^2\right]\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[W_i\right]^2\mathbb{E}\left[X_i\right]^2 = \\
&= \frac{1}{3}c^2/n \cdot \frac{1}{2} - 0 \cdot 0 = \frac{1}{6}c^2/n
\end{aligned}$$

We know now that $\sqrt{n}W_iX_i$ are independent, identically distributed random variables with mean zero and variance $\frac{1}{6}c^2$ and thus, by the central limit theorem $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sqrt{n}W_iX_i \xrightarrow{d} N\left(0, \frac{1}{6}c^2\right)$. Thus $\sum_{i=1}^{n} W_iX \xrightarrow{d} N(0, \frac{1}{6}c^2)$.

It has been claimed that $\sin(W^TX)$ will be Arcsin$(-1, 1)$ if $c > \sqrt{6}$. This is not the case: $\operatorname{Var}[\sin(kZ)] = \frac{1}{2}(1 - e^{-2k^2})$. Consequently, when $c = \sqrt{6}$ $W^TX \sim N(0, 1)$, giving $\operatorname{Var}[\sin(Z)] = \frac{1}{2}(1 - e^{-2}) \approx 0.43$, but if $\sin(Z)$ had been Arcsin$(-1, 1)$ it would have variance 1/2.

The k at which $\frac{1}{2}(1 - e^{-2k^2}) = 0.5 - \epsilon$ does not grow quickly when ϵ is made small. The k which produces $\epsilon = 10^{-p}$ is $k = \sqrt{p\log(10)/2}$.



This gives a reason to choose higher values of c than $\sqrt{6}$ or $\frac{\pi}{2\sqrt{6}}$. Another will come from the following analysis of backward propagation in SIRENs.

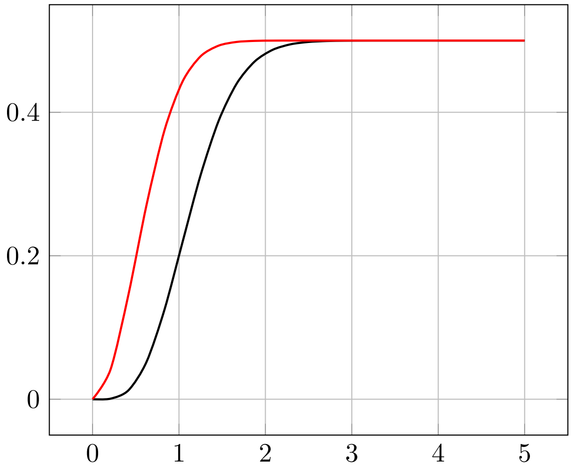
Backward propagation in SIRENs

Consider the elementwise nonlinearity of a SIREN neuron and the set of synapses which receive input from it. This is a function $f : \mathbb{R} \to \mathbb{R}^n$ assigning to an input x the vector $(w_1 \sin(x), ..., w_n \sin(x))^T$. Consider the case when $\frac{\partial E}{\partial f(x)_k}$ are known. Then $\frac{\partial E}{\partial x} = \frac{\partial E}{\partial f(x)_k} \frac{\partial f(x)_k}{\partial x} = \sum_k \frac{\partial E}{\partial f(x)_k} w_k \cos(x)$.

Treating the gradients $\frac{\partial E}{\partial f(x)_k}$ as inputs and the gradient $\frac{\partial E}{\partial x}$ we obtain a dual neuron with activation function $\cos(\cdot)$ and weights w_k. The weight distribution of $(w_k)_{k=1}^n$ will be the same as for forward layers.

Over many fully connected layers their pre-activations, i.e. gradients before the dual activation function is applied will become approximately normal distributed: when the input distribution has high variance it will be approximately uniformly distributed on a wide interval, and such a distribution transformed by the cosine will be approximately $\text{Arcsine}(-1, 1)$ distributed. Consequently this will lead to approximately normal distributed pre-activations in the dual, backwards network.

However, for the cosine of zero-centred normal distribution to be approximately $\text{Arcsine}(-1, 1)$ this distribution must have higher variance than for the sine of the same distribution to be $\text{Arcsine}(-1, 1)$ distributed. This can be seen by comparing the variance $\text{Var}[\cos(kZ)] = \frac{1}{2} + \frac{1}{2}e^{-2k^2} - e^{-k^2}$ to the variance $\text{Var}[\sin(kZ)] = \frac{1}{2}(1 - e^{-2k^2})$.



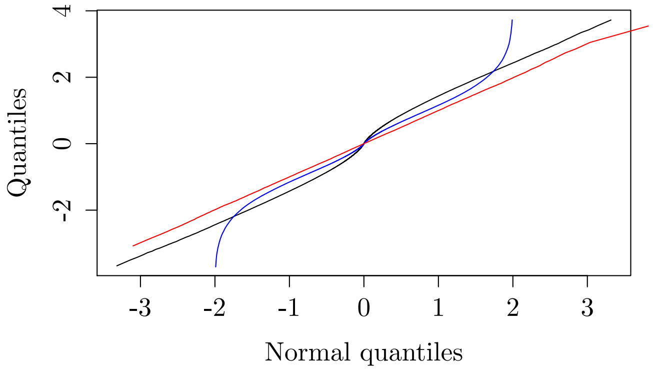
The black curve shows the variance of a cosine transformed zero-centred normal distribution and the red that of a sine-transformed zero-centred normal distribution as a function of the standard deviation of that distribution.

The convergence of the pre-activations to normal distribution

With uniformly distributed input the product of a weight and an input has finite variance and mean and with correct normalization their sums convergence to a normal distribution.

However, sums of variables that are themselves normal distributed need no reference to the central limit theorem, and are inherently normal distributed; furthermore, the product of a uniform random variable and any sine-transformed random variable has finite support, with no tails. Consequently they must converge to a normal distribution unnecessarily slowly.

This gives reason to consider weight distributions whose products with $\sin(kZ)$ where Z is a normal distribution are zero-centred normal distributons, however, it straightforward to consider simply normal distributed weights, which are sufficient to make the product of a weight and a pre-activation have tails like those of a normal distributed random variable.



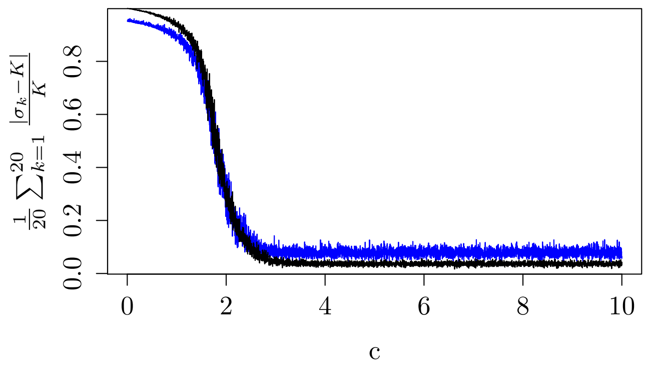
Th red curve shows samples from a normal distribution plotted against samples from a normal distribution, the black samples from a product of a sine-transformed normal distribution and a normal distribution against a normal distribution and the blue curve shows samples from the product of a uniform distribution and a sine-transformed normal random variable against a normal distribution. It's apparent that while the uniform distribution leads to good fit in the centre it fits the tails badly.

Normal distributed weights will be found to lead to higher final accuracy in experiments.

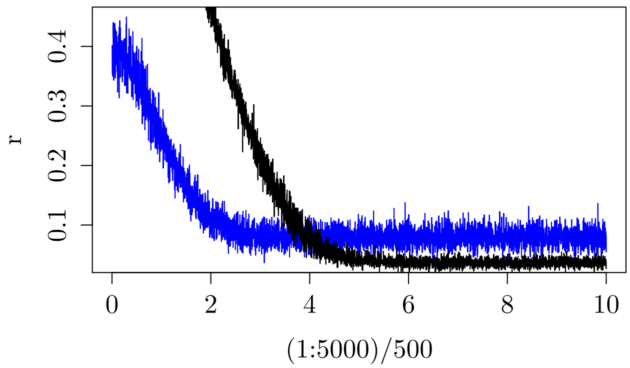
Simulations

The need for higher variance pre-activations becomes apparent in simulations. We consider a SIREN neural network with 20 layers and width 64 which receives $\mathcal{U}[-\pi/2, \pi/2]$ at initialization, for initializations with different choices of c.

The blue and black curves show, respectively, the mean of the absolute deviation of the pre-activations and of the activations from their intended values, $c/\sqrt{6}$ for the pre-activations and $\sqrt{1/2}$ for the activations, divided by those intended values, as a function of c.



In this we show a dual network corresponding to gradient propagation, unrealistically receiving $\mathcal{U}[-\pi/2, \pi/2]$ distributed input.



This gives reason to consider values of c as large as 5. It will be found that such networks can be fitted to images to substantially higher accuracy than has previously been possible.

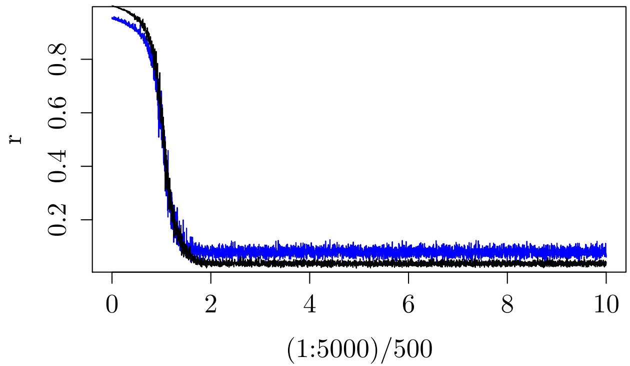
Because of the high values of c are motivated by the variance of activations and pre-activations from its intended values in the network with cosine activation this also gives reason to consider a shifted sine, which places the forward and backward propagation on equal terms, i.e. to use $\sin(\cdot + \pi/4)$ activation.

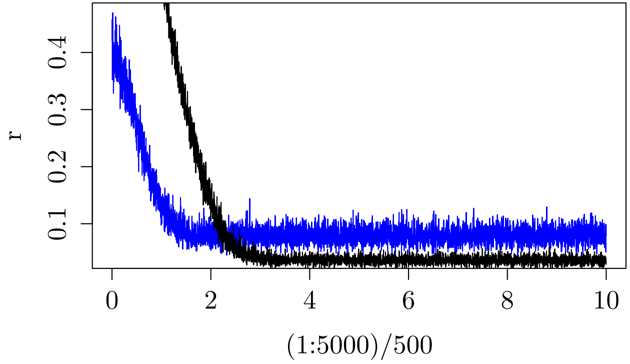
Experiments

We consider four situations: SIRENs with uniform initialization, SIRENs with normal initialization, SIRENS with uniform initialization and $\pi/4$-shifted activation function and SIRENS with normal initialization and $\pi/4$-shifted activation functions. For each of these we consider the results of exploratory simulations of in order to determine a reasonable choice for the variance of the initialization, of the type that we have already shown for SIRENs with uniform initialization.

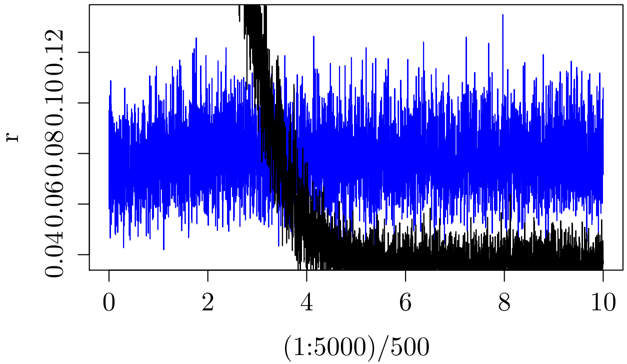
As before the The blue and black curves show, respectively, the mean of the absolute deviation of the pre-activations and of the activations from their intended values, for the pre-activations and for the activations, divided by those intended values, as a function of an initialization parameter.

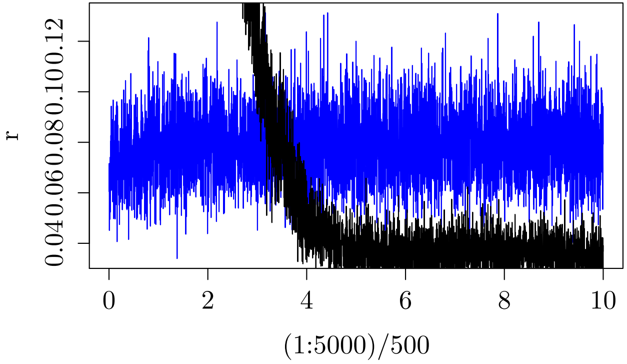
These first two graphs relate to the case where the initialization is normal with mean zero and and standard deviation $c/\sqrt{n}$ where n is the layer width. The first graph, showing standard deviation deviations during forward propagation, indicates that a good choice of c should be $c > 2$, the second showing backward standard deviation deviations indicate that a good choice of c should be $c > 3$.



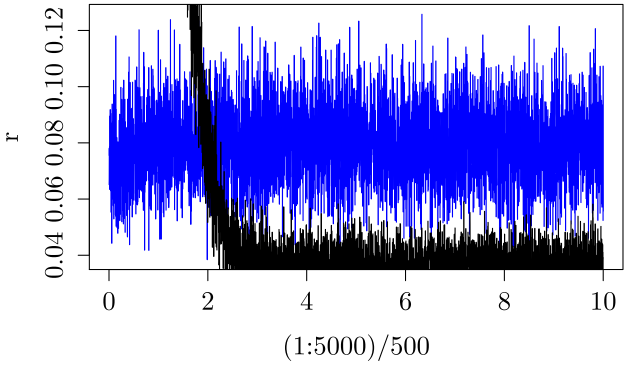


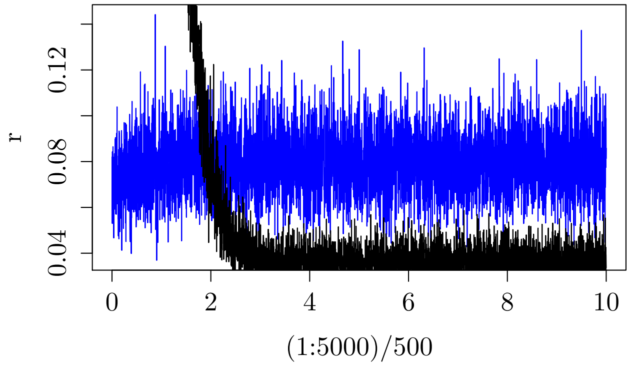
These second two graphs relate to the case where the initialization is $\mathcal{U}[-c/\sqrt{n}, c/\sqrt{n}]$-distributed where n is the layer width and where the activation function is $\sin(\cdot + \pi/4)$. Because of the symmetry produced by the use of the shifted activation function both graphs are the same and and indicate that a good choice of c is $c > 5$.



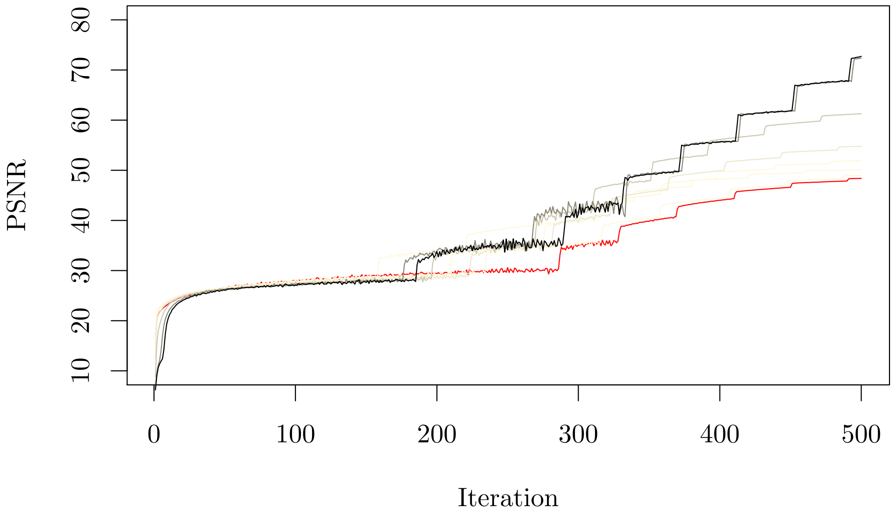


These third pair of graphs relate to the case where the initialization is $N(0, c^2/n)$ where n is the layer width and where the activation function is $\sin(\cdot + \pi/4)$. Again because of the symmetry produced by the use of a shifted activation function both graphs are the same and indicate that a good choice of c is $c > 2.7$. Observe also that the deviation of the pre-activations is low even for very small values of c: $\mathbb{E}[\sin(W + \pi/4)^2] = 1/\sqrt{2}$ for any proability distribution that is symmetric about zero.

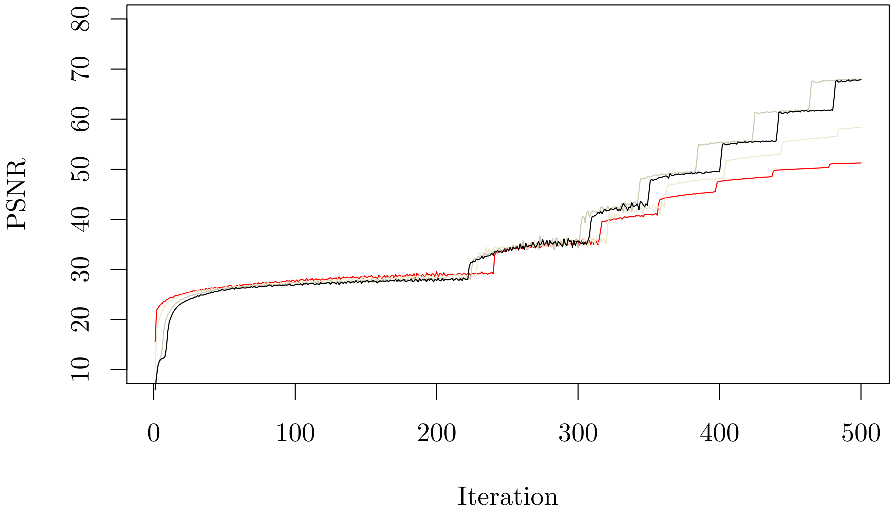




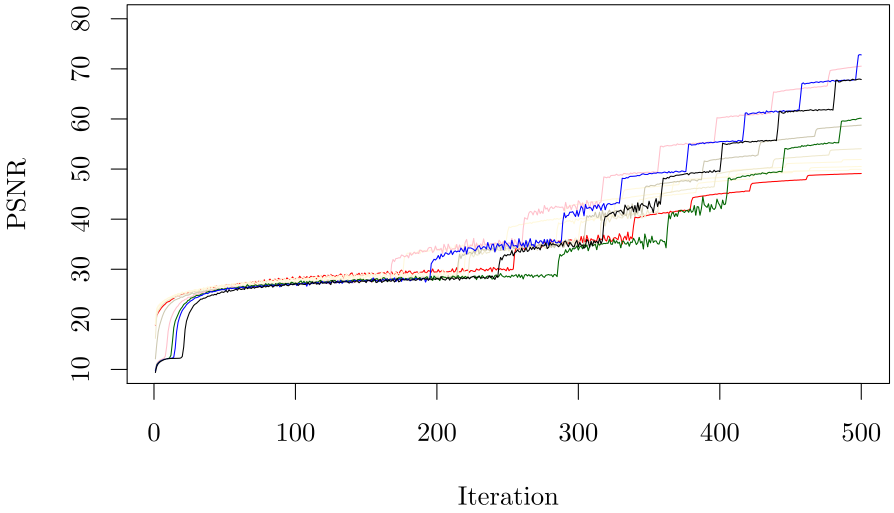
The graph below shows training runs with uniform weight initialization. The training run shown in black uses $c = 5.1$, that in red $c = 2.0$, the intermediary curves show intermediary values $c = 2.5, 3.0, 3.5, 4.0, 4.5$ and produce in turn curves that have higher final PSNR. The c proposed by Sitzmann et al. is $c = \sqrt{6} = 2.44$.



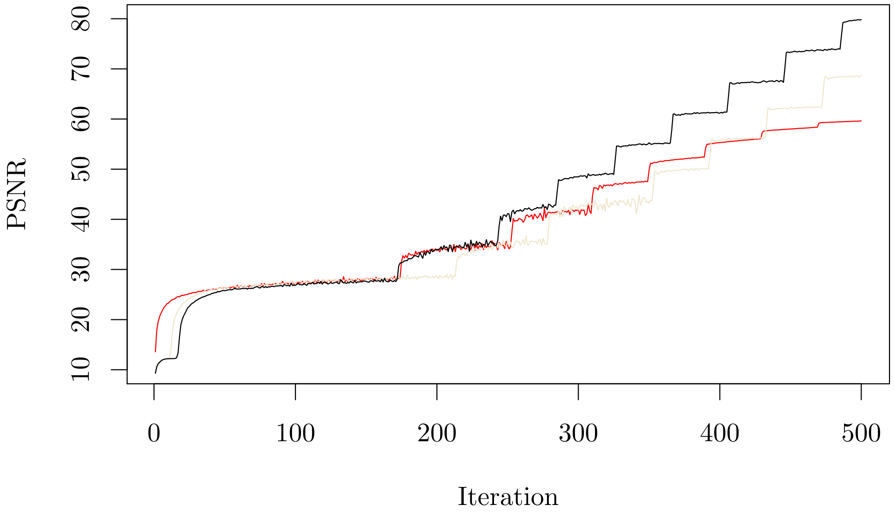
The graph below shows training runs with $N(0, c^2/n)$ weight initialization. The training run shown in black uses $c = 3.1$, that in red $c = 2.0$. Th remainder use intermediary values $c = 2.5, 3.0$, with the higher values producing higher final PSNR.



The graph below shows training runs with $\mathcal{U}[-c/\sqrt{n}, c/\sqrt{n}]$ initializations with the shifted activation function $\sin(\cdot + \pi/4)$. c corresponds to the training runs as follows: black: $c = 5.3$, blue $c = 5.2$, green, $c = 5.1$, pink $c = 5.0$. The remaining runs correspond to $c = 2.0, 2.5, 3.0, 3.5, 4.0$ and $c = 4.5$, with the higher values producing higher final PSNR.



The graph below shows training runs with $N(0, c^2/n)$ weight initialization and shifted activation function $\sin(\cdot + \pi/4)$. The black curve corresponds to $c = 3.1$, the red to $c = 2.5$ and the intermediary curve to $c = 3.0$.



P.S. **Derivation of the formula for $\boldsymbol{E}[\sin(\boldsymbol{kZ})^2]$**

We give the derivation of the formula for $\mathbb{E}[\sin(kZ)^2]$. Consider a mean of a particular transformed Wiener process

$$f(t) = \mathbb{E}\left[\sin^2(cW_t)\right].$$

Knowing the derivatives

$$\frac{d}{dx}\sin^2(cx) = 2c\sin(cx)\cos(cx) = c\sin(2cx)$$

.

$$\frac{d^2}{dx^2}\sin(x)^2 = 2c^2\cos(2cx),$$

we may apply Itô's lemma

$$\begin{aligned}\sin^2(cW_t) &= \int_0^t c\sin(2cW_s)dW_s + \frac{1}{2}\int_0^t 2c^2\cos(2cW_s)ds = \\ &= \int_0^t c\sin(2cW_s)dW_s + \int_0^t c^2\cos(2cW_s)ds.\end{aligned}$$

Consequently

$$f(t) = c^2\int_0^t \mathbb{E}\left[\cos(2cW_s)\right]ds.$$

Now, consider

$$g(t,u) = \mathbb{E}\left[\cos(uW_t)\right].$$

Applying Itô's lemma we obtain

$$\begin{aligned} g(t,u) = \mathbb{E}\left[\cos(u(W_s))\right] = \mathbb{E}\left[\frac{1}{2}\int_0^t -u^2\cos(uW_s)ds\right] = \\ = -\frac{1}{2}u^2\int_0^t \mathbb{E}\left[\cos(uW_s)\right]ds = -\frac{1}{2}u^2\int_0^t g(s,u)ds. \end{aligned}$$

Thus

$$\frac{\partial g}{\partial t}\left(t,u\right) = -\frac{1}{2}u^2 g\left(t,u\right)$$

and

$$g(t,u) = C(u)e^{-\frac{1}{2}u^2 t}$$

.

Since

$$g(0,u) = 1$$

We have

$$C(u) = 1$$

and thus that

$$g(t,u) = e^{-\frac{1}{2}u^2 t}.$$

Thus

$$\begin{aligned} f(t) = c^2 \int_0^t \mathbb{E}\left[\cos(2cW_s)\right] ds = c^2 \int_0^t g(t,2c) = \\ = c^2 \int_0^t e^{-\frac{1}{2}4c^2 t} = c^2 \int_0^t e^{-2c^2 t} = \\ = c^2 \frac{e^{-2tc^2} - 1}{-2c^2} = \frac{1 - e^{-2c^2 t}}{2}. \end{aligned}$$

We can conclude that

$$\mathbb{E}\left[\sin(cZ)^2\right] = \frac{1 - e^{-2c^2}}{2}.$$

P.P.S. **Derivation of the formula for Var $[\cos(cZ)]$**

In the calculation of $\mathbb{E}[\sin(cZ)^2]$ we obtained two results that are relevant also for this calculation, that

$$\mathbb{E}[\cos(uW_t)] = e^{-\frac{1}{2}u^2 t}$$

and the conclusion of the previous calculation, that

$$\mathbb{E}[\sin(cZ)^2] = \frac{1 - e^{-2c^2}}{2}.$$

Because

$$\mathbb{E}\left[\sin(cZ)^2 + \cos(cZ)^2\right] = 1$$

it immediately follows that

$$\begin{aligned}\mathbb{E}\left[\cos(cZ)^2\right] &= 1 - \mathbb{E}\left[\sin(cZ)^2\right] = \\ &= 1 - \frac{1 - e^{-2c^2}}{2} = \frac{1}{2} + \frac{1}{2}e^{-2c^2}.\end{aligned}$$

Using that

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

we obtain

$$\begin{aligned}\mathrm{Var}\left[\cos\left(cZ\right)\right] &= \frac{1}{2} + \frac{1}{2}e^{-2c^2} - \left(e^{-\frac{1}{2}c^2}\right)^2 = \\ &= \frac{1}{2} + \frac{1}{2}e^{-2c^2} - e^{-c^2}\end{aligned}$$