https://arxiv.org/pdf/1502.01852.pdf

https://stats.stackexchange.com/questions/15978/variance-of-product-of-dependent-variables

What are we trying to initialize?
https://datascience.stackexchange.com/questions/82917/why-do-we-want-the-variance-of-the-layers-to-remain-the-same-throughout-a-deep-n

> What exactly is helped by the variance being stable in this way?

Following the notation of the article, let's compute the gradient of the cost function w.r.t. the parameters in two consecutive layers (that we are going to call: layer i and layer $i+1$). In this layers, the quantity used to update their respective weight matrices, is given by:

$$\text{Layer } 1 \rightarrow \frac{\partial Cost}{\partial W^i} = \frac{\partial Cost}{\partial s^i}\frac{\partial s^i}{\partial W^i} = \frac{\partial Cost}{\partial s^i}(z^{i-1})^T$$

$$\text{Layer } 2 \rightarrow \frac{\partial Cost}{\partial W^{i+1}} = \frac{\partial Cost}{\partial s^{i+1}}\frac{\partial s^{i+1}}{\partial W^{i+1}} = \frac{\partial Cost}{\partial s^{i+1}}(z^{i})^T$$

There we can see that if we consider:

$$Var\left(\frac{\partial Cost}{\partial s^i}\right) = Var\left(\frac{\partial Cost}{\partial s^{i+1}}\right) \quad \leftrightarrow \quad Var(z^i) = Var(z^{i-1})$$

Then we would have:

$$Var\left(\frac{\partial Cost}{\partial W^i}\right) = Var\left(\frac{\partial Cost}{\partial W^{i+1}}\right)$$

This is a good thing because having the same variance in the updates of both layers means that the updates are globally spread in the same way, so assuming that the mean value of $\partial Cost/\partial W$ in both layers is the same, then this would mean that globally these **layers are learning at the same rythm**.

> Whatever the answer to (2) is, what is the proof or evidence that that's the case?

A good advantage of the previous reasoning is that if we could achieve this happening in the whole neural network, then all the layers in the NN would be learning at the same rythm! So problems like vanishing or exploding gradient would be avoided.

$$var(XY) = E\left((XY - E(XY))^2\right)$$
$$= E\left(X^2Y^2\right) - E(XY)^2$$

by definition
$$cov(X, Y) = E((X - E(X))(Y - E(Y)))$$
$$= E(XY) - E(X)E(Y)$$

and
$$cov\left(X^2, Y^2\right) = E\left(X^2Y^2\right) - E\left(X^2\right)E\left(Y^2\right)$$

$$so\ var(XY) = cov\left(X^2, Y^2\right) + E\left(X^2\right)E\left(Y^2\right) - (cov(X, Y) + E(X)E(Y))^2$$

$$if\ cov(X, Y) = cov\left(X^2, Y^2\right) = 0,$$
$$var(XY) = E\left(X^2\right)E\left(Y^2\right) - (E(X)E(Y))^2$$

$= \left(var(X) + E(X)^2\right)\left(var(Y) + E(Y)^2\right) - (E(X)E(Y))^2$
$= var(X)var(Y) + var(X)E(Y)^2 + var(Y)E(X)^2$

we want to keep the variance of gradient updates $\frac{\partial L}{\partial w} = X^t \frac{\partial L}{\partial y}$ *constant across every layer, so that the le*
can be kept the same across every layer
assuming X^t *and* $\frac{\partial L}{\partial y}$ *at each level are independent,*

$$var\left(\frac{\partial L}{\partial w}\right) = var\left(X^t \frac{\partial L}{\partial y}\right) = var\left(X^t\right) var\left(\frac{\partial L}{\partial y}\right) + var\left(X^t\right) E\left(\frac{\partial L}{\partial y}\right)^2 + var\left(\frac{\partial L}{\partial y}\right) E\left(X^t\right)^2$$

if we can keep $var(X)$ *and* $var\left(\frac{\partial L}{\partial y}\right)$ *constant across all the layers, then* $var\left(\frac{\partial L}{\partial w}\right)$ *will also be constant*!

$\mathbf{y} = \mathbf{Wx} + \mathbf{b}$
$var(\mathbf{y})$ = *trace of covariance matrix of* $y = var(\Sigma w_i x_i) = n\, var(w_i x_i)$
assuming $\mathbf{x}$ *and* $\mathbf{W}$ *are mutually independent*
and $E(w) = 0$
$var(w_i x_i) = var(w_i)var(x_i) + var(w_i)E(x_i)^2 + var(x_i)E(w_i)^2$
$= var(w_i)\left(var(x_i) + E(x_i)^2\right) = var(w_i)E\left(x_i^2\right)$

so $var(\mathbf{y}) = n_l var(w_i)E\left(x_i^2\right)$
if w^{l-1} *has a symmetric distribution around* 0, *then* $y^{l-1} \equiv f^{-1}\left(x^l\right)$ *has a symmetric distribution around* 0 *and mean* 0, *and if* f *is RELU*

then $E\left(x^2\right) = E\left(f(y)^2\right)$
$= E\left(0.5y^2\right) = 0.5E\left(y^2\right) = \frac{1}{2}var(y)$

so $var\left(\mathbf{y}^l\right) = \frac{1}{2} n_l var(w_i) var\left(y^{l-1}\right)$
this yields the recursive relationship

$$var\left(\mathbf{y}^l\right) = var\left(\mathbf{y}^1\right) \prod_{l=2}^{L} \frac{1}{2} n_l var\left(w^l\right)$$

A proper initialization method should avoid reducing or magnifying the magnitudes of input signals exponentially.

to keep $var(\mathbf{y}^l) \equiv var(\mathbf{y}^1)$, *we need to make* $\prod_{l=2}^{L} \frac{1}{2} n_l var(w^l) = 1 \rightarrow$ *so if we set set* $var(w^l) = 2/n$

for every level, we might be able to!

For the backpropagation, something similar, see the paper for details