Consider the gradient estimation for a standard MLP:

\textbf{Forward pass}
\[
x^{(0)} = x, \quad \forall l = 1, \ldots, L, \quad \left\{ \begin{array}{l}
s^{(l)} = w^{(l)} x^{(l-1)} + b^{(l)} \\
x^{(l)} = \sigma \left( s^{(l)} \right) \end{array} \right.
\]

\textbf{Backward pass}
\[
\left[ \frac{\partial \ell}{\partial x^{(L)}} \right] \text{ from the definition of } \ell \\
\left[ \begin{array}{c}
\frac{\partial \ell}{\partial w^{(l)}} \\
\frac{\partial \ell}{\partial s^{(l)}} \\
\frac{\partial \ell}{\partial \sigma} \\
\frac{\partial \ell}{\partial b^{(l)}} 
\end{array} \right] = \left( w^{(l+1)} \right)^\top \left[ \begin{array}{c}
\frac{\partial \ell}{\partial s^{(l+1)}} \\
\frac{\partial \ell}{\partial x^{(l-1)}} \\
\frac{\partial \ell}{s^{(l)}} \\
\frac{\partial \ell}{\sigma} \\
\frac{\partial \ell}{b^{(l-1)}}
\end{array} \right] \quad \sigma^\prime \left( s^{(l)} \right)
\]
We have

$$\left[ \frac{\partial \ell}{\partial x^{(l)}} \right] = \left( w^{(l+1)} \right)^T \left( \sigma' \left( s^{(l)} \right) \odot \left[ \frac{\partial \ell}{\partial x^{(l+1)}} \right] \right).$$

so the gradient “vanishes” exponentially with the depth if the $w$s are ill-conditioned or the activations are in the saturating domain of $\sigma$.

(Glorot and Bengio, 2010)
The design of the weight initialization aims at controlling

\[ \nabla \left( \frac{\partial \ell}{\partial w_{i,j}^{(l)}} \right) \quad \text{and} \quad \nabla \left( \frac{\partial \ell}{\partial b_{i}^{(l)}} \right) \]

so that **weights evolve at the same rate across layers during training**, and no layer reaches a saturation behavior before others.

We will use that, if \( A \) and \( B \) are independent

\[ \nabla (AB) = \nabla (A) \nabla (B) + \nabla (A) E(B)^2 + \nabla (B) E(A)^2 . \]

Notation in the coming slides will drop indexes when variances are identical for all activations or parameters in a layer.
In a standard layer

\[ x_i^{(l)} = \sigma \left( \sum_{j=1}^{N_{l-1}} w_{i,j} x_j^{(l-1)} + b_i^{(l)} \right) \]

where \( N_l \) is the number of units in layer \( l \), and \( \sigma \) is the activation function.

Assuming \( \sigma'(0) = 1 \), and we are in the linear regime

\[ x_i^{(l)} \simeq \sum_{j=1}^{N_{l-1}} w_{i,j} x_j^{(l-1)} + b_i^{(l)}. \]

From which, if both the \( w^{(l)} \)s and \( x^{(l-1)} \)s are centered, and biases set to zero:

\[ \mathbb{V}(x_i^{(l)}) \simeq \mathbb{V} \left( \sum_{j=1}^{N_{l-1}} w_{i,j} x_j^{(l-1)} \right) = \sum_{j=1}^{N_{l-1}} \mathbb{V}(w_{i,j}) \mathbb{V}(x_j^{(l-1)}) \]

and the \( x^{(l)} \)s are centered.

So if the \( w_{i,j}^{(l)} \) are sampled i.i.d in each layer, and all the activations have same variance, then

\[ \mathbb{V}(x_i^{(l)}) \simeq \sum_{j=1}^{N_{l-1}} \mathbb{V}(w_{i,j}) \mathbb{V}(x_j^{(l-1)}) = N_{l-1} \mathbb{V}(w^{(l)}) \mathbb{V}(x^{(l-1)}). \]

So we have for the variance of the activations:

\[ \mathbb{V}(x^{(l)}) \simeq \mathbb{V}(x^{(0)}) \prod_{q=1}^{l} N_{q-1} \mathbb{V}(w^{(q)}), \]

which leads to a first type of initialization to ensure

\[ \mathbb{V}(w^{(l)}) = \frac{1}{N_{l-1}}. \]
The standard PyTorch weight initialization for a linear layer

\[ f : \mathbb{R}^N \rightarrow \mathbb{R}^M \]

is

\[ w_{i,j} \sim \mathcal{U} \left[ -\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}} \right] \]

hence

\[ \mathbb{V}(w) = \frac{1}{3N}. \]

```python
>>> f = nn.Linear(5, 100000)
>>> f.weight.mean()
tensor(0.0007, grad_fn=<MeanBackward0>)
>>> f.weight.var()
tensor(0.0667, grad_fn=<VarBackward0>)
>>> torch.empty(1000000).uniform_(-1/math.sqrt(5), 1/math.sqrt(5)).var()
tensor(0.0667)
>>> 1./15.
0.06666666666666667
```

We can look at the variance of the gradient w.r.t. the activations. Since

\[ \frac{\partial \ell}{\partial x_i^{(l)}} \approx \sum_{h=1}^{N_{l+1}} \frac{\partial \ell}{\partial x_h^{(l+1)}} w_{h,i}^{(l+1)} \]

we get

\[ \mathbb{V} \left( \frac{\partial \ell}{\partial x_i^{(l)}} \right) \approx N_{l+1} \mathbb{V} \left( \frac{\partial \ell}{\partial x_h^{(l+1)}} \right) \mathbb{V} \left( w_{(l+1)} \right). \]

So we have for the variance of the gradient w.r.t. the activations:

\[ \mathbb{V} \left( \frac{\partial \ell}{\partial x_i^{(l)}} \right) \approx \mathbb{V} \left( \frac{\partial \ell}{\partial x_{(l)}} \right) \prod_{q=l+1}^{L} N_q \mathbb{V} \left( w^{(q)} \right). \]
Since
\[ x_i^{(l)} \simeq \sum_{j=1}^{N_{l-1}} w_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)} \]
we have
\[ \frac{\partial \ell}{\partial w_{i,j}^{(l)}} \simeq \frac{\partial \ell}{\partial x_i^{(l)}} x_j^{(l-1)} \]
and we get the variance of the gradient w.r.t. the weights
\[
\text{Var}\left( \frac{\partial \ell}{\partial w^{(l)}} \right) \simeq \text{Var}\left( \frac{\partial \ell}{\partial x^{(l)}} \right) \text{Var}\left( x^{(l-1)} \right)
= \text{Var}\left( \frac{\partial \ell}{\partial x^{(L)}} \right) \prod_{q=l+1}^{L} N_q \text{Var}\left( w^{(q)} \right) \text{Var}\left( x^{(0)} \right) \prod_{q=1}^{l} N_{q-1} \text{Var}\left( w^{(q)} \right)
= \frac{N_0}{N_l} \left( \prod_{q=1}^{L} N_q \text{Var}\left( w^{(q)} \right) \right) \text{Var}\left( x^{(0)} \right) \text{Var}\left( \frac{\partial \ell}{\partial x^{(L)}} \right).
\]

Similarly, since
\[ x_i^{(l)} \simeq \sum_{j=1}^{N_{l-1}} w_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)} \]
we have
\[ \frac{\partial \ell}{\partial b_i^{(l)}} \simeq \frac{\partial \ell}{\partial x_i^{(l)}} \]
so we get the variance of the gradient w.r.t. the biases
\[
\text{Var}\left( \frac{\partial \ell}{\partial b^{(l)}} \right) \simeq \text{Var}\left( \frac{\partial \ell}{\partial x^{(l)}} \right).
\]
So the variance of the gradient w.r.t. the weights is the same in all layers.

To control the variance of activations, we need

$$V(w^{(l)}) = \frac{1}{N_{l-1}}.$$ 

and to control the variance of the gradient w.r.t. activations, and through it the variance of the gradient w.r.t. the biases

$$V(w^{(l)}) = \frac{1}{N_{l}}.$$ 

From which we get as a compromise the “Xavier initialization”

$$V(w^{(l)}) = \frac{1}{N_{l-1}+N_{l}} = \frac{2}{N_{l-1} + N_{l}}.$$ 

(Glorot and Bengio, 2010)

In torch/nn/init.py

```python
def xavier_normal_(tensor, gain = 1):
    fan_in, fan_out = _calculate_fan_in_and_fan_out(tensor)
    std = gain * math.sqrt(2.0 / (fan_in + fan_out))
    with torch.no_grad():
        return tensor.normal_(0, std)
```
4.2.2 Gradient Propagation Study

To empirically validate the above theoretical ideas, we have plotted some normalized histograms of activation values, 

\[
\text{Back-propagated gradients}
\]

These results illustrate the effect of the choice of activation and initialization. As a reference we include in Figure 7, the back-propagated gradients normalized histograms with hyperbolic tangent activation, with standard (top) vs normalized (bottom) initialization.

4.3 Back-propagated Gradients During Learning

As first noted by Bradley (2009), we observe (Figure 7) that the variance of the back-propagated gradients at the beginning of training, after the standard initialization, gets smaller as it is propagated downwards. However we find that this trend is reversed very quickly during learning. The back-propagated gradients of very different magnitudes at different layers are obtained with the other datasets.

The dynamic of learning in such networks is complex and increases back-propagated gradients (bottom of Figure 7). Using our normalized initialization we do not see such deepening of the error as training progresses and asymptotes. Figure 11 illustrates with error curves showing the evolution of test errors for all the datasets studied (Glorot and Bengio, 2010).

When consecutive layers have the same dimension, the average singular value corresponds to the average ratio of in-finital activation variances going from \( z_i \) to \( z_{i+1} \). With our normalized initialization, this ratio is constant across layers, as shown on Figure 8. However, this is not anymore independent of the activation values and the weight gradients and of the back-propagated gradients at layer \( i \). Remember that if \( A \) and \( B \) are independent, we have

\[
\text{var}(AB) = \text{var}(A) \text{var}(B) + \text{var}(A) \text{E}(B)^2 + \text{var}(B) \text{E}(A)^2
\]

\[
= \text{var}(A) \text{E}(B^2) + \text{var}(B) \text{E}(A)^2.
\]

The weights can also be scaled to account for the activation functions.

\[
\text{The weights can also be scaled to account for the activation functions.}
\]

Remember that if \( A \) and \( B \) are independent, we have

\[
\text{var}(AB) = \text{var}(A) \text{var}(B) + \text{var}(A) \text{E}(B)^2 + \text{var}(B) \text{E}(A)^2
\]

\[
= \text{var}(A) \text{E}(B^2) + \text{var}(B) \text{E}(A)^2.
\]
For the forward pass, if

\[ s_i^{(l)} = \sum_{j=1}^{N_{l-1}} w_{i,j}^{(l)} \sigma \left( s_j^{(l-1)} \right) + b_i^{(l)} \]

\[ x_i^{(l)} = \sigma \left( s_i^{(l)} \right), \]

and \( E \left( w^{(l)} \right) = 0 \), \( s^{(l-1)} \) is symmetric, and \( \sigma \) is ReLU, we have

\[ V \left( s_i^{(l)} \right) = N_{l-1} V \left( w^{(l)} \sigma \left( s^{(l-1)} \right) \right) \]

\[ = N_{l-1} V \left( w^{(l)} \right) E \left( \sigma \left( s^{(l-1)} \right)^2 \right) \]

\[ = N_{l-1} V \left( w^{(l)} \right) \frac{1}{2} E \left( \left( s^{(l-1)} \right)^2 \right) \]

\[ = \frac{1}{2} N_{l-1} V \left( w^{(l)} \right) V \left( s^{(l-1)} \right). \]

For the backward

\[ V \left( \frac{\partial \ell}{\partial x_i^{(l)}} \right) = \sum_{h=1}^{N_{l+1}} V \left( \sigma' \left( s_h^{(l+1)} \right) \frac{\partial \ell}{\partial x_h^{(l+1)}} w_{h,i}^{(l+1)} \right) \]

\[ = \sum_{h=1}^{N_{l+1}} E \left( \sigma' \left( s_h^{(l+1)} \right) \left( \frac{\partial \ell}{\partial x_h^{(l+1)}} w_{h,i}^{(l+1)} \right)^2 \right) \]

\[ = \sum_{h=1}^{N_{l+1}} \frac{1}{2} E \left( \left( \frac{\partial \ell}{\partial x_h^{(l+1)}} w_{h,i}^{(l+1)} \right)^2 \right) \]

\[ = \frac{1}{2} N_{l+1} \sum_{h=1}^{N_{l+1}} V \left( \frac{\partial \ell}{\partial x_h^{(l+1)}} \right) V \left( w_{h,i}^{(l+1)} \right). \]
So ReLU impacts the forward and backward pass as if the weights had half their variances, which motivates multiplying them by a corrective gain of $\sqrt{2}$.

(He et al., 2015)

The same type of reasoning can be applied to other activation functions.

In `torch/nn/init.py`

```python
def calculate_gain(nonlinearity, param=None):
    linear_fns = ['linear', 'conv1d', 'conv2d', 'conv3d',
                  'conv_transpose1d', 'conv_transpose2d', 'conv_transpose3d']
    if nonlinearity in linear_fns or nonlinearity == 'sigmoid':
        return 1
    elif nonlinearity == 'tanh':
        return 5.0 / 3
    elif nonlinearity == 'relu':
        return math.sqrt(2.0)
    /.../
```

Data normalization
The analysis for the weight initialization relies on keeping the activation variance constant.

For this to be true, not only the variance has to remained unchanged through layers, but it has to be correct for the input too.

\[ \mathcal{V}(x^{(0)}) = 1. \]

This can be done in several ways. Under the assumption that all the input components share the same statistics, we can do

```python
mu, std = train_input.mean(), train_input.std()
train_input.sub_(mu).div_(std)
test_input.sub_(mu).div_(std)
```

Thanks to the magic of broadcasting we can normalize component-wise with

```python
mu, std = train_input.mean(0), train_input.std(0)
train_input.sub_(mu).div_(std)
test_input.sub_(mu).div_(std)
```

To go one step further, some techniques initialize the weights explicitly so that the empirical moments of the activations are as desired.

As such, they take into account the statistics of the network activation induced by the statistics of the data.
References
