Xavier and He initialization: the variance by hand

Ten problems on weight initialization: the variance of a product and of a dense layer's output, why a ReLU halves the second moment, He's 2/n_in and Xavier's 2/(n_in + n_out) derived forwards and backwards, the signal scale through depth, fan-in for convolutions, the leaky-ReLU gain and uniform bounds, and residual stacks, with worked solutions and the mistakes that lose a factor of 2, 3 or 9.

BEFORE YOU START

Before the first gradient step, a network's weights are random numbers, and their variance decides whether a signal that enters the first layer arrives at the last one at a usable size. Too small and every layer shrinks it, so the output and the gradients vanish; too large and they explode. The two standard answers, Xavier (Glorot and Bengio, 2010) and He (He et al., 2015), come from the same short calculation: the variance of a sum of products of independent random numbers. These ten problems do that calculation forwards and backwards, through a ReLU and a leaky ReLU, for convolutions, for uniform distributions and through residual stacks. The five mistakes at the end each change the variance by a constant factor per layer, which depth turns into an exponential: a variance used where a second moment belongs, Xavier trusted on a ReLU network, a uniform's half-width read as its standard deviation, a convolution's fan-in counted as its channels, and a residual block assumed to preserve what its branch preserves.

- $\mathbb{E}$ is expectation and $\operatorname{Var}(u) = \mathbb{E}[u^2] - (\mathbb{E}[u])^2$. The second moment $\mathbb{E}[u^2]$ equals the variance exactly when $\mathbb{E}[u] = 0$. If u and v are independent, $\mathbb{E}[uv] = \mathbb{E}[u]\,\mathbb{E}[v]$, and the same holds for any functions of them, such as u^2 and v^2.
- A dense layer is $y = Wx$ with $W \in \mathbb{R}^{n_{\text{out}} \times n_{\text{in}}}$, so $y_i = \sum_j W_{ij}\, x_j$. The fan-in n_{in} is the number of inputs each output sums; the fan-out n_{out} is the number of outputs each input feeds. Biases are initialized to 0 and left out.
- At initialization the W_{ij} are independent and identically distributed, independent of the layer's input, with mean 0 and variance σ_w^2, and symmetric: $-W_{ij}$ has the same distribution as W_{ij}. $\mathcal{N}(0, \sigma^2)$ (normal) and $U(-b, b)$ (uniform on $[-b, b]$) are both symmetric.
- A random variable z is symmetric about 0 if $-z$ has the same distribution as z; then $\mathbb{E}[z] = 0$. Every distribution on this page is continuous, so $P(z = 0) = 0$.
- $\operatorname{relu}(t) = \max(0, t)$, and $\operatorname{relu}'(t)$ is 1 for $t > 0$ and 0 for $t < 0$. The leaky ReLU with slope $\alpha \in [0, 1]$ is $f_\alpha(t) = t$ for $t \geq 0$ and αt for $t < 0$. Both act entry by entry.

- A deep network: $y^{(l)} = W^{(l)} h^{(l-1)}$ and $h^{(l)} = \mathrm{relu}(y^{(l)})$ for $l = 1, \dots, L$, with $h^{(0)} = x$ the input, $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ and weight variance σ_l^2. Entries of one layer are treated as identically distributed, so $\mathrm{Var}(y^{(l)})$ means the variance of any one entry.
- The backward pass uses the code-style names of the minibatch page: dy is $\nabla_y L$ for a scalar loss L. Through $y = Wx$, $dx = W^\top dy$, so $dx_j = \sum_i W_{ij}\, dy_i$; through $h = \mathrm{relu}(y)$, $dy = dh \odot \mathrm{relu}'(y)$, with $\odot$ the elementwise product.
- All of this describes the network at initialization, averaged over the random weights. The backward calculations also treat a layer's W as independent of the dy arriving from above, the simplification Glorot and Bengio made; it is not exact, but it predicts the scale well.

Builds on: One-hidden-layer backprop, the whole backward pass

Problems

1. Let w and x be independent with $\mathbb{E}[w] = 0$ and $\mathrm{Var}(w) = \sigma_w^2$. Show that $\mathrm{Var}(wx) = \sigma_w^2\, \mathbb{E}[x^2]$. Evaluate it for $\sigma_w^2 = 0.25$ and an x with mean 1 and variance 4, and compare it with $\mathrm{Var}(w)\,\mathrm{Var}(x)$. difficulty ·

2. A dense layer $y = Wx$ has weights as in Before you start, and inputs x_j with a common second moment $q = \mathbb{E}[x_j^2]$; the x_j need not be independent of each other. Show that $\mathrm{Var}(y_i) = n_{\text{in}}\, \sigma_w^2\, q$, and find the σ_w^2 that makes $\mathbb{E}[y_i^2] = q$. difficulty ··

3. Let z be symmetric about 0 with variance σ^2. Show that $\mathbb{E}[\mathrm{relu}(z)^2] = \sigma^2/2$. For $z \sim \mathcal{N}(0, \sigma^2)$, also find $\mathbb{E}[\mathrm{relu}(z)]$ and $\mathrm{Var}(\mathrm{relu}(z))$. difficulty ··

4. In the deep network of Before you start, suppose the entries of $y^{(l-1)}$ are symmetric about 0. Show that $\mathrm{Var}(y^{(l)}) = \frac{1}{2}\, n_{l-1}\, \sigma_l^2\ \mathrm{Var}(y^{(l-1)})$, that the entries of $y^{(l)}$ are again symmetric, and find the σ_l^2 that keeps the variance fixed. difficulty ···

5. A ReLU network of constant width n uses $\sigma_l^2 = c/n$ in every layer. Find $\mathrm{Var}(y^{(L)}) / \mathrm{Var}(y^{(1)})$. Evaluate it for $c = 1$ and $L = 21$, which is Xavier's choice when $n_{\text{in}} = n_{\text{out}} = n$, and say what $c = 2.2$ does over the same depth. difficulty ··

6. In the backward pass through $y = Wx$, $dx = W^\top dy$. Suppose the dy_i share a second moment $\mathbb{E}[dy_i^2]$ and are independent of W. Show that $\mathrm{Var}(dx_j) = n_{\text{out}}\,\sigma_w^2\,\mathbb{E}[dy_i^2]$. Explain Xavier's $\sigma_w^2 = 2/(n_{\text{in}} + n_{\text{out}})$, and evaluate it and both scale factors for a layer from 784 inputs to 256 outputs. difficulty ··

7. In a ReLU network with He's fan-in initialization $\sigma_l^2 = 2/n_{l-1}$, the gradient goes back through $dy^{(l)} = dh^{(l)} \odot \mathrm{relu}'(y^{(l)})$ and $dh^{(l-1)} = W^{(l)\top} dy^{(l)}$. Treat $\mathrm{relu}'(y^{(l)})$ as independent of $dh^{(l)}$ and $W^{(l)}$ as independent of $dy^{(l)}$. Show that $\mathbb{E}[(dh^{(l-1)})^2] = \dfrac{n_l}{n_{l-1}}\,\mathbb{E}[(dh^{(l)})^2]$, and find the factor from $dh^{(4)}$ back to $dh^{(1)}$ for widths $n_1 = 1024$, $n_2 = 512$, $n_3 = 256$, $n_4 = 128$. Which variance would make each factor 1? difficulty ···

8. A 2-D convolution has $C_{\text{in}} = 64$ input channels, $C_{\text{out}} = 128$ output channels, a 3×3 kernel and stride 1: $y_{o,r,c} = \sum_{i,a,b} K_{o,i,a,b}\,x_{i,r+a,c+b}$. Find its fan-in and fan-out, and the He and Xavier weight variances. difficulty ·

9. For a leaky ReLU, show that $\mathbb{E}[f_\alpha(z)^2] = \frac{1}{2}(1+\alpha^2)\,\sigma_z^2$ for z symmetric about 0, and find the weight standard deviation $g/\sqrt{n_{\text{in}}}$ that preserves the variance, where g is the gain. Show that $\mathrm{Var}\left(U(-b,b)\right) = b^2/3$ and find the uniform bound b with the same variance. PyTorch's `nn.Linear` initializes its weight with `kaiming_uniform_(weight, a=math.sqrt(5))`, where `a` is the slope that enters the gain: what are its bound and variance? difficulty ···

10. A residual stack computes $h^{(l)} = h^{(l-1)} + \beta F_l(h^{(l-1)})$ for $l = 1, \dots, L$, with a scalar $\beta \geq 0$. Each branch is initialized so that $\mathbb{E}[F_l(h)_i\,h_i] = 0$ and $\mathbb{E}[F_l(h)_i^2] = \mathbb{E}[h_i^2]$ (for example $F_l(h) = Vh$ with $\mathrm{Var}(V_{ij}) = 1/n$). With $q_l = \mathbb{E}[(h_i^{(l)})^2]$, find q_L/q_0. Evaluate it for $\beta = 1$, $L = 10$, and for $\beta = 1/\sqrt{L}$ with $L = 10$; show that the second is below e for every L. difficulty ···

Answers

1. $\mathrm{Var}(wx) = \sigma_w^2\,\mathbb{E}[x^2] = 0.25 \times 5 = 1.25$, while $\mathrm{Var}(w)\,\mathrm{Var}(x) = 1$
2. $\mathrm{Var}(y_i) = n_{\text{in}}\,\sigma_w^2\,q$, so $\sigma_w^2 = 1/n_{\text{in}}$ keeps $\mathbb{E}[y_i^2] = q$
3. $\mathbb{E}[\mathrm{relu}(z)^2] = \sigma^2/2$; for normal z, $\mathbb{E}[\mathrm{relu}(z)] = \sigma/\sqrt{2\pi}$ and $\mathrm{Var}(\mathrm{relu}(z)) = \sigma^2\left(\frac{1}{2} - \frac{1}{2\pi}\right) \approx 0.341\,\sigma^2$
4. $\mathrm{Var}(y^{(l)}) = \frac{1}{2}\,n_{l-1}\sigma_l^2\,\mathrm{Var}(y^{(l-1)})$; He initialization $\sigma_l^2 = 2/n_{l-1}$ makes the factor exactly 1
5. $\mathrm{Var}(y^{(L)})/\,\mathrm{Var}(y^{(1)}) = (c/2)^{L-1}$; for $c = 1$, $L = 21$ it is $2^{-20} \approx 9.5 \times 10^{-7}$, a standard deviation 1024 times smaller
6. $\mathrm{Var}(dx_j) = n_{\text{out}}\,\sigma_w^2\,\mathbb{E}[dy_i^2]$; for $784 \to 256$, Xavier gives $\sigma_w^2 = 2/1040 = 1/520$, a forward factor $784/520 \approx 1.51$ and a backward factor $256/520 \approx 0.49$
7. $\mathbb{E}[(dh^{(l-1)})^2] = (n_l/n_{l-1})\,\mathbb{E}[(dh^{(l)})^2]$; from layer 4 back to layer 1 the factor is $n_4/n_1 = 1/8$; $\sigma_l^2 = 2/n_l$ (fan-out) makes each factor 1
8. fan-in $= 576$, fan-out $= 1152$; He $\sigma_w^2 = 2/576 = 1/288$ (standard deviation ≈ 0.0589); Xavier $\sigma_w^2 = 2/(576 + 1152) = 1/864$
9. $\mathbb{E}[f_\alpha(z)^2] = \frac{1}{2}(1+\alpha^2)\sigma_z^2$, $g = \sqrt{2/(1+\alpha^2)}$, uniform bound $b = g\sqrt{3/n_{\text{in}}}$; with slope $\sqrt{5}$ the bound is $1/\sqrt{n_{\text{in}}}$ and the variance $1/(3n_{\text{in}})$
10. $q_L/q_0 = (1+\beta^2)^L$: 1024 for $\beta = 1$, $L = 10$, and $(1 + 1/L)^L \approx 2.594 < e$ for $\beta = 1/\sqrt{L}$, $L = 10$

Worked solutions

Problem 1

> Let w and x be independent with $\mathbb{E}[w] = 0$ and $\operatorname{Var}(w) = \sigma_w^2$. Show that $\operatorname{Var}(wx) = \sigma_w^2\,\mathbb{E}[x^2]$. Evaluate it for $\sigma_w^2 = 0.25$ and an x with mean 1 and variance 4, and compare it with $\operatorname{Var}(w)\operatorname{Var}(x)$.

$\mathbb{E}[wx] = \mathbb{E}[w]\,\mathbb{E}[x] = 0.$	Independence factorizes the expectation of the product, and $\mathbb{E}[w] = 0$ kills it whatever x's mean is.
$\operatorname{Var}(wx) = \mathbb{E}[w^2x^2] - 0^2 = \mathbb{E}[w^2]\,\mathbb{E}[x^2].$	The definition of variance with step 1's mean; w^2 and x^2 are functions of independent variables, so they are independent too.
$\mathbb{E}[w^2] = \operatorname{Var}(w) + (\mathbb{E}[w])^2 = \sigma_w^2.$	The definition of variance rearranged, with a zero mean.
$\mathbb{E}[x^2] = \operatorname{Var}(x) + (\mathbb{E}[x])^2 = 4 + 1 = 5.$	The same identity for x, whose mean is not zero, so its square stays.
$\operatorname{Var}(wx) = \sigma_w^2\,\mathbb{E}[x^2] = 0.25 \times 5 = 1.25,$ **while** $\operatorname{Var}(w)\operatorname{Var}(x) = 1$	The product of the variances misses $\sigma_w^2(\mathbb{E}[x])^2$: because w's sign is random, even a constant input contributes spread to wx. The two agree only when $\mathbb{E}[x] = 0$, which a ReLU output never satisfies (Problem 3).

Problem 2

> A dense layer $y = Wx$ has weights as in Before you start, and inputs x_j with a common second moment $q = \mathbb{E}[x_j^2]$; the x_j need not be independent of each other. Show that $\operatorname{Var}(y_i) = n_{\text{in}}\,\sigma_w^2\,q$, and find the σ_w^2 that makes $\mathbb{E}[y_i^2] = q$.

$\mathbb{E}[y_i] = \sum_j \mathbb{E}[W_{ij}]\,\mathbb{E}[x_j] = 0.$	Linearity of expectation, then independence of W from x, then the zero mean of the weights.
$\mathbb{E}[y_i^2] = \sum_j \sum_k \mathbb{E}[W_{ij}W_{ik}\,x_jx_k].$	Square the sum and use linearity; there are n_{in}^2 terms.

For $j \neq k$: $\mathbb{E}[W_{ij} W_{ik}\, x_j x_k] = \mathbb{E}[W_{ij}]\,\mathbb{E}[W_{ik}]\,\mathbb{E}[x_j x_k] = 0$.	W_{ij}, W_{ik} and the input are mutually independent. The cross terms vanish because of the weights' zero means, so correlated inputs do no harm.
For $j = k$: $\mathbb{E}[W_{ij}^2\, x_j^2] = \sigma_w^2\, q$.	Problem 1, step 2, with $\mathbb{E}[W_{ij}^2] = \sigma_w^2$.
$\mathrm{Var}(y_i) = n_{\text{in}}\, \sigma_w^2\, q$, **so** $\sigma_w^2 = 1/n_{\text{in}}$ **keeps** $\mathbb{E}[y_i^2] = q$	Only the n_{in} diagonal terms survive, and with a zero mean the second moment is the variance. This is LeCun's initialization. It preserves the scale of a linear network, and of a tanh network near 0, where $\tanh t \approx t$. For $n_{\text{in}} = 256$ it is a standard deviation of $1/16$.

Problem 3

> Let z be symmetric about 0 with variance σ^2. Show that $\mathbb{E}[\mathrm{relu}(z)^2] = \sigma^2/2$. For $z \sim \mathcal{N}(0, \sigma^2)$, also find $\mathbb{E}[\mathrm{relu}(z)]$ and $\mathrm{Var}(\mathrm{relu}(z))$.

$\mathrm{relu}(z)^2 = z^2$ when $z > 0$ and 0 when $z \leq 0$.	relu passes positive values and zeroes the rest.
$\mathbb{E}[z^2;\, z > 0] = \mathbb{E}[z^2;\, z < 0]$, where $\mathbb{E}[u;\, A]$ is the expectation of u times the indicator of the event A.	$-z$ has the same distribution as z, and $z \mapsto -z$ swaps the two events while leaving z^2 unchanged.
$\mathbb{E}[z^2;\, z > 0] + \mathbb{E}[z^2;\, z < 0] = \mathbb{E}[z^2] = \sigma^2$.	The two events cover everything except $z = 0$, where $z^2 = 0$; and $\mathbb{E}[z] = 0$, so the second moment is the variance.
$\mathbb{E}[\mathrm{relu}(z)^2] = \sigma^2/2$.	Steps 1 to 3: the left side is one of two equal halves of σ^2.
For $z \sim \mathcal{N}(0, \sigma^2)$: $\mathbb{E}[\mathrm{relu}(z)] = \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_0^\infty t\, e^{-t^2/(2\sigma^2)}\, dt = \frac{\sigma^2}{\sigma\sqrt{2\pi}} = \frac{\sigma}{\sqrt{2\pi}}$.	The integrand is the normal density times t on $t > 0$, and $-\sigma^2 e^{-t^2/(2\sigma^2)}$ is an antiderivative of $t\, e^{-t^2/(2\sigma^2)}$.

$\mathbb{E}[\mathrm{relu}(z)^2] = \sigma^2/2$; **for normal** z, $\mathbb{E}[\mathrm{relu}(z)] = \sigma/\sqrt{2\pi}$ **and** $\mathrm{Var}(\mathrm{relu}(z)) = \sigma^2\left(\frac{1}{2} - \frac{1}{2\pi}\right) \approx 0.341\,\sigma^2$	The variance is step 4 minus step 5 squared. The next layer needs the second moment, $\sigma^2/2$, not this variance (Problem 2), and the $\frac{1}{2}$ is what He's factor 2 undoes (Problem 4).

Problem 4

> In the deep network of Before you start, suppose the entries of $y^{(l-1)}$ are symmetric about 0. Show that $\mathrm{Var}(y^{(l)}) = \frac{1}{2}\, n_{l-1}\, \sigma_l^2\ \mathrm{Var}(y^{(l-1)})$, that the entries of $y^{(l)}$ are again symmetric, and find the σ_l^2 that keeps the variance fixed.

$\mathrm{Var}(y_i^{(l)}) = n_{l-1}\,\sigma_l^2\,\mathbb{E}\big[(h_j^{(l-1)})^2\big]$.	Problem 2 with $x = h^{(l-1)}$, which is computed from earlier layers only and so is independent of $W^{(l)}$.
$\mathbb{E}\big[(h_j^{(l-1)})^2\big] = \frac{1}{2}\,\mathrm{Var}(y_j^{(l-1)})$.	Problem 3 applied to $z = y_j^{(l-1)}$, which is symmetric by hypothesis.
$\mathrm{Var}(y_i^{(l)}) = \frac{1}{2}\, n_{l-1}\, \sigma_l^2\ \mathrm{Var}(y_j^{(l-1)})$.	Substitute step 2 into step 1.
Negating row i of $W^{(l)}$ negates $y_i^{(l)}$ and leaves the joint distribution of the weights and $h^{(l-1)}$ unchanged, so $y_i^{(l)}$ is symmetric.	The weights are symmetric and independent of everything below them. The first layer's output is symmetric for the same reason, whatever the input, so the hypothesis holds at every layer from $l = 2$ on.
$\mathrm{Var}(y^{(l)}) = \frac{1}{2}\, n_{l-1}\sigma_l^2\, \mathrm{Var}(y^{(l-1)})$; **He initialization** $\sigma_l^2 = 2/n_{l-1}$ **makes the factor exactly** 1	The standard deviation is $\sqrt{2/n_{\text{in}}}$, which is $1/16$ for $n_{\text{in}} = 512$. PyTorch's `kaiming_normal_` with `nonlinearity='relu'` uses this, as gain/$\sqrt{\text{fan_in}}$ with gain $\sqrt{2}$; its default mode is `'fan_in'`.

Problem 5

A ReLU network of constant width n uses $\sigma_l^2 = c/n$ in every layer. Find $\mathrm{Var}(y^{(L)})/\,\mathrm{Var}(y^{(1)})$. Evaluate it for $c = 1$ and $L = 21$, which is Xavier's choice when $n_{\text{in}} = n_{\text{out}} = n$, and say what $c = 2.2$ does over the same depth.

$\mathrm{Var}(y^{(l)}) = \frac{c}{2}\,\mathrm{Var}(y^{(l-1)})$ for $l \geq 2$.	Problem 4, step 3, with $n_{l-1}\sigma_l^2 = c$.
$\mathrm{Var}(y^{(L)}) = \left(\frac{c}{2}\right)^{L-1}\mathrm{Var}(y^{(1)})$.	Apply step 1 for $l = 2, \ldots, L$: that is $L - 1$ factors. The first layer reads the raw input, not a ReLU output, so it is not one of them.
For $c = 1$, $L = 21$: $2^{-20} = 1/1048576 \approx 9.5 \times 10^{-7}$.	$2^{10} = 1024$, so $2^{20} = 1024^2$.
For $c = 2.2$: $1.1^{20} \approx 6.7$.	The same formula with $c/2 = 1.1$.
$\mathrm{Var}(y^{(L)})/\,\mathrm{Var}(y^{(1)}) = (c/2)^{L-1}$; **for** $c = 1$, $L = 21$ **it is** $2^{-20} \approx 9.5 \times 10^{-7}$, **a standard deviation** 1024 **times smaller**	The standard deviation scales by the square root, 2^{-10}. A constant factor per layer becomes exponential in depth, so only $c = 2$ neither vanishes nor explodes; 10% too much variance per layer already multiplies it by 6.7 over twenty layers.

Problem 6

In the backward pass through $y = Wx$, $dx = W^\top dy$. Suppose the dy_i share a second moment $\mathbb{E}[dy_i^2]$ and are independent of W. Show that $\mathrm{Var}(dx_j) = n_{\text{out}}\,\sigma_w^2\,\mathbb{E}[dy_i^2]$. Explain Xavier's $\sigma_w^2 = 2/(n_{\text{in}} + n_{\text{out}})$, and evaluate it and both scale factors for a layer from 784 inputs to 256 outputs.

$dx_j = \sum_{i=1}^{n_{\text{out}}} W_{ij}\,dy_i$.	Entry j of $W^\top dy$ combines column j of W with dy: one term per output.
$\mathbb{E}[dx_j] = 0$ and $\mathbb{E}[dx_j^2] = n_{\text{out}}\,\sigma_w^2\,\mathbb{E}[dy_i^2]$.	Problem 2's argument with the roles swapped: the cross terms $\mathbb{E}[W_{ij}W_{kj}\,dy_i\,dy_k]$ vanish for $i \neq k$ because W_{ij} and W_{kj} are independent with mean 0, and the sum now runs over n_{out} terms.

The forward pass keeps its scale when $n_{\text{in}}\sigma_w^2 = 1$ (Problem 2), the backward pass when $n_{\text{out}}\sigma_w^2 = 1$.	Both hold only if $n_{\text{in}} = n_{\text{out}}$.
$\sigma_w^2 = \dfrac{2}{n_{\text{in}} + n_{\text{out}}}$ gives $\frac{1}{2}\big(n_{\text{in}}\sigma_w^2 + n_{\text{out}}\sigma_w^2\big) = 1$.	Glorot and Bengio's compromise: the variance is the reciprocal of the average fan, so the two factors average to 1.
$\operatorname{Var}(dx_j) = n_{\text{out}}\,\sigma_w^2\,\mathbb{E}[dy_i^2]$; **for** $784 \to 256$, **Xavier gives** $\sigma_w^2 = 2/1040 = 1/520$, **a forward factor** $784/520 \approx 1.51$ **and a backward factor** $256/520 \approx 0.49$	Neither direction is preserved exactly; each is off by the ratio of one fan to the average. The derivation assumes linear units, so on a ReLU network both factors are halved again (Problems 5 and 7).

Problem 7

> In a ReLU network with He's fan-in initialization $\sigma_l^2 = 2/n_{l-1}$, the gradient goes back through $dy^{(l)} = dh^{(l)} \odot \operatorname{relu}'(y^{(l)})$ and $dh^{(l-1)} = W^{(l)\top} dy^{(l)}$. Treat $\operatorname{relu}'(y^{(l)})$ as independent of $dh^{(l)}$ and $W^{(l)}$ as independent of $dy^{(l)}$. Show that $\mathbb{E}[(dh^{(l-1)})^2] = \dfrac{n_l}{n_{l-1}}\,\mathbb{E}[(dh^{(l)})^2]$, and find the factor from $dh^{(4)}$ back to $dh^{(1)}$ for widths $n_1 = 1024$, $n_2 = 512$, $n_3 = 256$, $n_4 = 128$. Which variance would make each factor 1?

$\operatorname{relu}'(y_i^{(l)})$ is 1 with probability $\frac{1}{2}$ and 0 otherwise.	$y_i^{(l)}$ is symmetric about 0 (Problem 4, step 4) and continuous, so it is positive half the time.
$\mathbb{E}[(dy_i^{(l)})^2] = \mathbb{E}[\operatorname{relu}'(y_i^{(l)})^2]\,\mathbb{E}[(dh_i^{(l)})^2] = \frac{1}{2}\,\mathbb{E}[(dh_i^{(l)})^2]$.	The two factors are treated as independent, and $0^2 = 0$, $1^2 = 1$, so the square of the mask is the mask.
$\mathbb{E}[(dh_j^{(l-1)})^2] = n_l\,\sigma_l^2\,\mathbb{E}[(dy_i^{(l)})^2]$.	Problem 6, step 2: $W^{(l)}$ has n_l rows, so the backward sum has n_l terms.
$\mathbb{E}[(dh_j^{(l-1)})^2] = n_l \cdot \dfrac{2}{n_{l-1}} \cdot \frac{1}{2}\,\mathbb{E}[(dh_i^{(l)})^2] = \dfrac{n_l}{n_{l-1}}\,\mathbb{E}[(dh_i^{(l)})^2]$.	Steps 2 and 3 with $\sigma_l^2 = 2/n_{l-1}$: the ReLU's $\frac{1}{2}$ cancels the 2.

From $dh^{(4)}$ to $dh^{(1)}$: $\dfrac{n_4}{n_3}\cdot\dfrac{n_3}{n_2}\cdot\dfrac{n_2}{n_1}=\dfrac{n_4}{n_1}=\dfrac{128}{1024}=\dfrac{1}{8}$.	Step 4 for $l=4,3,2$; the product telescopes.
$\mathbb{E}[(dh^{(l-1)})^2]=(n_l/n_{l-1})\,\mathbb{E}[(dh^{(l)})^2]$; **from layer** 4 **back to layer** 1 **the factor is** $n_4/n_1=1/8$; $\sigma_l^2=2/n_l$ **(fan-out) makes each factor** 1	With fan-in initialization the backward pass is off only by the ratio of two widths, not by an exponential in depth, which is why the fan-in mode is a safe default. PyTorch exposes the other choice as `mode='fan_out'`.

Problem 8

A 2-D convolution has $C_{\text{in}}=64$ input channels, $C_{\text{out}}=128$ output channels, a 3×3 kernel and stride 1: $y_{o,r,c}=\sum_{i,a,b}K_{o,i,a,b}\,x_{i,r+a,c+b}$. Find its fan-in and fan-out, and the He and Xavier weight variances.

Each output $y_{o,r,c}$ is a sum over $i=1,\dots,64$ and $a,b=0,1,2$: $64\times 3\times 3=576$ products.	Fan-in counts the products summed into one output, and a convolution sums over every input channel and every kernel position.
$\operatorname{Var}(y_{o,r,c})=576\,\sigma_w^2\,\mathbb{E}[x^2]$.	Problem 2 with $n_{\text{in}}=576$: the 576 weights in one output's sum are distinct and independent. The same weights are reused at other positions, which correlates different outputs but does not change the variance of one.
An input $x_{i,r,c}$ away from the border appears in $128\times 3\times 3=1152$ outputs, each through a different weight.	It sits at each of the 9 kernel offsets of some window, for each of the 128 output channels; the backward sum of Problem 6 has that many terms.
fan-in $=576$, **fan-out** $=1152$; **He** $\sigma_w^2=2/576=1/288$ **(standard deviation** ≈ 0.0589**)**; **Xavier** $\sigma_w^2=2/(576+1152)=1/864$	Both are channels times the receptive-field size k_hk_w, which is how PyTorch computes the fans of a convolution weight.

Problem 9

> For a leaky ReLU, show that $\mathbb{E}[f_\alpha(z)^2] = \frac{1}{2}(1+\alpha^2)\,\sigma_z^2$ for z symmetric about 0, and find the weight standard deviation $g/\sqrt{n_{\text{in}}}$ that preserves the variance, where g is the gain. Show that $\operatorname{Var}\big(U(-b,b)\big) = b^2/3$ and find the uniform bound b with the same variance. PyTorch's `nn.Linear` initializes its weight with `kaiming_uniform_(weight, a=math.sqrt(5))`, where `a` is the slope that enters the gain: what are its bound and variance?

$f_\alpha(z)^2 = z^2$ for $z > 0$ and $\alpha^2 z^2$ for $z < 0$.	The leaky ReLU multiplies negative inputs by α, so their squares by α^2.
$\mathbb{E}[f_\alpha(z)^2] = \frac{1}{2}\sigma_z^2 + \alpha^2 \cdot \frac{1}{2}\sigma_z^2 = \frac{1}{2}(1+\alpha^2)\,\sigma_z^2$.	Problem 3, steps 2 and 3: each sign carries half of σ_z^2.
$\operatorname{Var}(y^{(l)}) = \frac{1}{2}(1+\alpha^2)\,n_{\text{in}}\,\sigma_w^2\operatorname{Var}(y^{(l-1)})$, so $\sigma_w^2 = \dfrac{2}{(1+\alpha^2)\,n_{\text{in}}}$ and $g = \sqrt{2/(1+\alpha^2)}$.	Problem 4, step 3, with step 2 in place of the ReLU's $\frac{1}{2}$. $\alpha = 0$ gives He's $\sqrt{2}$ and $\alpha = 1$, a linear unit, gives LeCun's 1.
$\operatorname{Var}\big(U(-b,b)\big) = \displaystyle\int_{-b}^{b} \frac{t^2}{2b}\,dt = \frac{b^2}{3}$.	The density is $1/(2b)$ on the interval and the mean is 0; $\int_{-b}^{b} t^2\,dt = 2b^3/3$.
$b^2/3 = g^2/n_{\text{in}}$ gives $b = g\sqrt{3/n_{\text{in}}}$.	Match the uniform's variance to the normal one's.
Slope $\sqrt{5}$ in step 3's gain: $g^2 = 2/(1+5) = 1/3$, so $b = \sqrt{1/3}\,\sqrt{3/n_{\text{in}}} = 1/\sqrt{n_{\text{in}}}$.	`kaiming_uniform_` computes its bound as gain times $\sqrt{3/\text{fan_in}}$, with the leaky-ReLU gain of step 3 evaluated at its `a`; a slope above 1 is not a sensible activation, only a way to set the scale.
$\mathbb{E}[f_\alpha(z)^2] = \frac{1}{2}(1+\alpha^2)\sigma_z^2$, $g = \sqrt{2/(1+\alpha^2)}$, **uniform bound** $b = g\sqrt{3/n_{\text{in}}}$; **with slope** $\sqrt{5}$ **the bound is** $1/\sqrt{n_{\text{in}}}$ **and the variance** $1/(3n_{\text{in}})$	That variance is a sixth of He's $2/n_{\text{in}}$, so a plain stack of default `nn.Linear` layers with ReLUs shrinks the variance by $1/6$ per layer (Problem 5 with $c = 1/3$). The same step 5 gives He uniform $\sqrt{6/n_{\text{in}}}$ and Xavier uniform $\sqrt{6/(n_{\text{in}}+n_{\text{out}})}$.

Problem 10

> A residual stack computes $h^{(l)} = h^{(l-1)} + \beta F_l(h^{(l-1)})$ for $l = 1, \dots, L$, with a scalar $\beta \geq 0$. Each branch is initialized so that $\mathbb{E}[F_l(h)_i\,h_i] = 0$ and $\mathbb{E}[F_l(h)_i^2] = \mathbb{E}[h_i^2]$ (for example $F_l(h) = Vh$

with $\mathrm{Var}(V_{ij}) = 1/n$). With $q_l = \mathbb{E}[(h_i^{(l)})^2]$, find q_L/q_0. Evaluate it for $\beta = 1$, $L = 10$, and for $\beta = 1/\sqrt{L}$ with $L = 10$; show that the second is below e for every L.

$q_l = \mathbb{E}[(h_i + \beta F_i)^2] = q_{l-1} + 2\beta\,\mathbb{E}[h_i F_i] + \beta^2\,\mathbb{E}[F_i^2]$, with $h = h^{(l-1)}$ and $F = F_l(h)$.	Expand the square and use linearity of expectation.
$q_l = (1 + \beta^2)\, q_{l-1}$.	The cross term is 0 and $\mathbb{E}[F_i^2] = q_{l-1}$, by the hypotheses on the branch.
$q_L = (1 + \beta^2)^L\, q_0$.	Apply step 2 for each of the L blocks.
$\beta = 1$, $L = 10$: $2^{10} = 1024$.	The skip and the branch each carry the full second moment, and they add.
$\beta = 1/\sqrt{L}$: $(1 + 1/L)^L$, which for $L = 10$ is $1.1^{10} \approx 2.594$.	Step 3 with $\beta^2 = 1/L$.
$L \log(1 + 1/L) < L \cdot \frac{1}{L} = 1$, so $(1 + 1/L)^L < e$.	$\log(1 + t) < t$ for $t > 0$, because log is concave and t is its tangent line at $t = 0$.
$q_L/q_0 = (1 + \beta^2)^L$: 1024 **for** $\beta = 1$, $L = 10$, **and** $(1 + 1/L)^L \approx 2.594 < e$ **for** $\beta = 1/\sqrt{L}$, $L = 10$	Uncorrelated signals add their second moments, so a stack of variance-preserving branches still grows exponentially with depth. Scaling the branches by $1/\sqrt{L}$ caps the growth at e however deep the stack; some ResNet recipes instead start each branch's last layer at zero, which is $\beta = 0$ at initialization.

Where this goes wrong

1. He variance computed from Var(h) instead of E[h²]

Initialization is described as keeping "the variance" fixed, so it is natural to put the previous layer's variance into the formula.

$\mathrm{Var}(y_i^{(l)}) = n\,\sigma^2\,\mathbb{E}\big[(h_j^{(l-1)})^2\big]$	Right so far: Problem 4, step 1, for width n.

"For independent factors, $\mathrm{Var}(wx) = \mathrm{Var}(w)\,\mathrm{Var}(x)$, so the input enters through its variance."	The shortcut that causes the mistake: that product rule needs both means to be 0, and a ReLU output has mean $\sigma/\sqrt{2\pi}$ (Problem 3).
$\mathrm{Var}(y^{(l)}) = n\sigma^2\left(\frac{1}{2} - \frac{1}{2\pi}\right)\mathrm{Var}(y^{(l-1)})$, so $\sigma^2 = 1/\left(\left(\frac{1}{2} - \frac{1}{2\pi}\right)n\right) \approx 2.93/n$ ✗	The second moment, $\frac{1}{2}\,\mathrm{Var}(y^{(l-1)})$, is what reaches the next layer (Problem 1), giving $2/n$. With $2.93/n$ each layer multiplies the variance by about 1.47, and twenty layers by about 2100.

2. Xavier assumed to keep a ReLU network's variance constant

Xavier initialization is a common library default (Keras's `Dense` layer uses `glorot_uniform`), and it was derived to keep the forward variance fixed.

$\sigma^2 = 2/(n_{\text{in}} + n_{\text{out}}) = 1/n$ when $n_{\text{in}} = n_{\text{out}} = n$	Right so far: Problem 6, step 4, at equal widths.
"With $n\sigma^2 = 1$, Problem 2 says each layer passes its variance on unchanged."	The shortcut that causes the mistake: Problem 2 preserves the second moment of the layer's input, and Xavier's derivation assumes that input is the previous layer's output unchanged, as in a linear network.
$\mathrm{Var}(y^{(l)}) = n\sigma^2\,\mathrm{Var}(y^{(l-1)}) = \mathrm{Var}(y^{(l-1)})$ ✗	A ReLU sits between the layers and halves the second moment (Problem 3), so the factor is $\frac{1}{2}$ per layer: after 21 layers the variance is 2^{-20} of the first layer's (Problem 5), and the gradients shrink the same way.

3. He uniform bound set to the standard deviation √(2/n_in)

The normal and the uniform versions of He initialization are both described by the scale $\sqrt{2/n_{\text{in}}}$, and that number gets used as the uniform's limit.

He initialization wants $\mathrm{Var}(W_{ij}) = 2/n_{\text{in}}$	Right so far: Problem 4.

“Sample $W_{ij} \sim U\big(-\sqrt{2/n_{\text{in}}},\ \sqrt{2/n_{\text{in}}}\big)$.”	The shortcut that causes the mistake: treating a uniform's half-width as its standard deviation.
$\text{Var}(W_{ij}) = 2/n_{\text{in}}$ ✗	$\text{Var}\left(U(-b,b)\right) = b^2/3$ (Problem 9), so this gives $2/(3n_{\text{in}})$, a third of the target, and the variance falls by a factor of 3 per ReLU layer. The bound must be $\sqrt{3}$ times larger: $\sqrt{6/n_{\text{in}}}$.

4. Convolution fan-in counted as the input channels

For a dense layer the fan-in is the number of input features, and a convolution's input features look like its channels.

$\text{Var}(y) = \text{fan-in} \times \sigma^2\, \mathbb{E}[x^2]$, where the fan-in is the number of products summed into one output	Right so far: Problem 2.
“The layer has $C_{\text{in}} = 64$ input channels, so its fan-in is 64.”	The analogy that causes the mistake: a convolution sums over the kernel's positions as well as the channels.
$\sigma^2 = 2/C_{\text{in}} = 2/64 = 1/32$ ✗	Each output sums $C_{\text{in}} k_h k_w = 576$ products (Problem 8), so with ReLUs each layer multiplies the variance by $\frac{1}{2} \times 576/32 = 9$, the kernel area: five such layers multiply it by $9^5 = 59049$. The correct variance is $2/576$.

5. Residual block assumed to preserve variance because its branch does

Each branch of a residual network is initialized to preserve the scale of its input, so the whole block looks as if it does too.

$\mathbb{E}[F_l(h)_i^2] = \mathbb{E}[h_i^2]$ and $\mathbb{E}[F_l(h)_i\, h_i] = 0$	Right so far: the branch hypotheses of Problem 10.

“The branch preserves the variance and the skip connection passes h through untouched, so their sum keeps it too.”	The analogy that causes the mistake: treating the sum of two signals of the same size as if it had that size.
$\mathbb{E}[(h_i^{(l)})^2] = \mathbb{E}[(h_i^{(l-1)})^2]$ ✗	Uncorrelated terms add their second moments, so each block doubles it (Problem 10): 2^L over L blocks, about 1.1×10^{15} for $L = 50$. Scale the branches by $1/\sqrt{L}$, or start them at zero.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 20 of *Deep Learning Math by Hand*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/xavier-and-he-initialization