Activation functions and their derivatives

Ten problems on the activations a network is built from: sigmoid, tanh, ReLU and leaky ReLU, softplus, SiLU and GELU, their derivatives, the same derivatives written in terms of the stored output, the backward pass through an elementwise layer and its gradient bounds, the behavior near zero, and the numerically stable cross-entropy from logits, with worked solutions and the mistakes that drop a product rule or differentiate the wrong variable.

BEFORE YOU START

Every layer of a network ends in an activation, and the backward pass through it is one elementwise multiplication by the activation's derivative. That is the whole of the math, but each activation has its own derivative, its own stable form and its own way of being written in terms of the output rather than the input, and those details are where hand-written backward passes go wrong. These ten problems differentiate the activations in common use, from the sigmoid and tanh of the first networks to the ReLU family, softplus, SiLU and GELU, rewrite each derivative in terms of the stored output where that is possible, push a gradient through an elementwise layer with the bounds that implies, look at every activation's behavior near zero, and end with the stable binary cross-entropy from logits that the same identities make safe. The five mistakes at the end are the ones that give a plausible-looking gradient: a derivative formula fed the pre-activation, a ReLU mask replaced by the ReLU's output, a gate treated as a constant, a stable formula with its linear term missing, and a chain rule stopped one factor short.

- An activation acts entry by entry: for $z \in \mathbb{R}^m$, $h = f(z)$ means $h_i = f(z_i)$ for a scalar function $f : \mathbb{R} \to \mathbb{R}$ with derivative f'. From the Jacobians page, $\partial h / \partial z = \operatorname{diag}(f'(z))$, and for a scalar loss L with $dh = \nabla_h L$ the gradient at z is $dz = dh \odot f'(z)$, with $\odot$ the elementwise product.
- The logistic sigmoid is $\sigma(t) = 1/(1 + e^{-t})$, with $\sigma'(t) = \sigma(t)(1 - \sigma(t))$ and $1 - \sigma(t) = \sigma(-t)$ (the regression page's Problem 4). $\tanh t = (e^t - e^{-t})/(e^t + e^{-t})$.
- $\operatorname{relu}(t) = \max(t, 0)$. The leaky ReLU with slope $0 < \alpha < 1$ is $f_\alpha(t) = t$ for $t \geq 0$ and αt for $t < 0$. The ELU with $\alpha > 0$ is t for $t \geq 0$ and $\alpha(e^t - 1)$ for $t < 0$.
- Softplus is $\zeta(t) = \log(1 + e^t)$. SiLU (also called Swish) is $t\,\sigma(t)$. GELU is $t\,\Phi(t)$, where Φ is the standard normal distribution function and $\varphi(t) = e^{-t^2/2}/\sqrt{2\pi}$ its density, with $\Phi' = \varphi$.
- $1[\cdot]$ is 1 when the condition holds and 0 otherwise. A derivative "in terms of the output" is a formula $f'(z) = g(h)$ using only $h = f(z)$, which matters because a framework keeps h for the backward pass and would otherwise have to keep z as well.
- Logarithms are natural, and $\log(1 + u)$ for small u is what `log1p` computes without losing precision.

Builds on: Jacobians and the chain rule

Problems

1. Show that $\zeta'(t) = \sigma(t)$, that $\log \sigma(t) = -\zeta(-t)$, and that $\dfrac{d}{dt}\log\sigma(t) = 1 - \sigma(t) = \sigma(-t)$. difficulty ·

2. Show that $\tanh'(t) = 1 - \tanh^2(t)$, that $\tanh t = 2\sigma(2t) - 1$, and hence that $\tanh'(t) = 4\sigma'(2t)$. What values can the two derivatives take? difficulty ·

3. Compute $\mathrm{relu}'(t)$ and $f'_\alpha(t)$ for $t \neq 0$, and say what happens at $t = 0$. Show that $\mathrm{relu}(ct) = c\,\mathrm{relu}(t)$ for $c > 0$ and what that says about the derivative. Then write both derivatives in terms of the output h. difficulty ··

4. SiLU is $f(t) = t\,\sigma(t)$. Compute $f'(t)$, write it in terms of the output $h = f(t)$ and $\sigma(t)$, and evaluate $f'(-2)$ to show that SiLU is not monotonic. difficulty ··

5. Compute $\mathrm{GELU}'(t)$ for $\mathrm{GELU}(t) = t\,\Phi(t)$. Then differentiate the tanh approximation $g(t) = \frac{1}{2}t\big(1 + \tanh u(t)\big)$ with $u(t) = \sqrt{2/\pi}\,(t + 0.044715\,t^3)$. difficulty ··

6. Show that $\zeta(t) - \mathrm{relu}(t) = \log(1 + e^{-|t|})$, which lies in $(0, \log 2]$, so that $\zeta(t) = \max(t, 0) + \log(1 + e^{-|t|})$. Show that $\zeta''(t) = \sigma(t)(1 - \sigma(t)) > 0$, so softplus is convex. difficulty ··

7. For the sigmoid, tanh, ReLU, leaky ReLU, softplus and ELU, write $f'(z)$ as a function of the output $h = f(z)$ alone. Do SiLU and GELU have such a formula? difficulty ··

8. A layer computes $z = Wx + b \in \mathbb{R}^m$ and $h = f(z)$ elementwise, and a scalar loss L has $dh = \nabla_h L$. Show that $dz = dh \odot f'(z)$. For $f = \tanh$, write $\nabla_W L$ and $\nabla_b L$ using only dh, h and x. Then show $\|dz\| \le \|dh\|$ for tanh and $\|dz\| \le \frac{1}{4}\|dh\|$ for the sigmoid. difficulty ···

9. For the sigmoid, tanh, softplus, SiLU and GELU, compute $f(0)$, $f'(0)$ and $f''(0)$, and write the quadratic approximation $f(t) \approx f(0) + f'(0)\,t + \frac{1}{2}f''(0)\,t^2$. difficulty ··

10. The binary cross-entropy of a logit t against a label $y \in \{0, 1\}$ is $\ell = -y \log \sigma(t) - (1 - y) \log(1 - \sigma(t))$. Show that $\ell = \zeta(t) - yt = \max(t, 0) - yt + \log(1 + e^{-|t|})$ and that $d\ell/dt = \sigma(t) - y$. Why is the last form the one libraries compute? difficulty ···

Answers

1. $\zeta'(t) = \sigma(t)$; $\log\sigma(t) = -\zeta(-t)$; $(\log\sigma)'(t) = 1 - \sigma(t) = \sigma(-t)$
2. $\tanh'(t) = 1 - \tanh^2(t) = 4\sigma'(2t)$; $\tanh' \in (0, 1]$ with its maximum 1 at $t = 0$, and $\sigma' \in (0, \frac{1}{4}]$ with its maximum at $t = 0$
3. $\mathrm{relu}'(t) = 1[t > 0] = 1[h > 0]$ and $f'_\alpha(t) = 1[t > 0] + \alpha 1[t < 0] = 1[h > 0] + \alpha 1[h < 0]$, with the value at 0 a convention; $\mathrm{relu}(ct) = c\,\mathrm{relu}(t)$ for $c > 0$, so the derivative is constant along each ray
4. $f'(t) = \sigma(t)\big(1 + t(1 - \sigma(t))\big) = h + \sigma(t)(1 - h)$; $f'(-2) \approx -0.091 < 0$, so SiLU decreases there
5. $\mathrm{GELU}'(t) = \Phi(t) + t\,\varphi(t)$ with $\varphi(t) = e^{-t^2/2}/\sqrt{2\pi}$; $g'(t) = \frac{1}{2}(1 + \tanh u) + \frac{1}{2}t\,(1 - \tanh^2 u)\sqrt{2/\pi}\,(1 + 0.134145\,t^2)$
6. $\zeta(t) = \max(t, 0) + \log(1 + e^{-|t|})$ with $0 < \zeta(t) - \mathrm{relu}(t) \le \log 2$; $\zeta''(t) = \sigma(t)(1 - \sigma(t)) > 0$, so softplus is convex
7. $\sigma' = h(1 - h)$; $\tanh' = 1 - h^2$; $\mathrm{relu}' = 1[h > 0]$; $f'_\alpha = 1[h > 0] + \alpha 1[h < 0]$; $\zeta' = 1 - e^{-h}$; $\mathrm{ELU}' = 1$ for $h > 0$ and $h + \alpha$ for $h < 0$; SiLU and GELU have no such formula, so their backward pass must keep z
8. $dz = dh \odot f'(z)$; for tanh, $\nabla_W L = \big(dh \odot (1 - h^2)\big)x^\top$ and $\nabla_b L = dh \odot (1 - h^2)$; $\|dz\| \le \|dh\|$ for tanh and $\|dz\| \le \frac{1}{4}\|dh\|$ for the sigmoid
9. $\sigma(t) \approx \frac{1}{2} + \frac{t}{4}$; $\tanh t \approx t$; $\zeta(t) \approx \log 2 + \frac{t}{2} + \frac{t^2}{8}$; $\mathrm{SiLU}(t) \approx \frac{t}{2} + \frac{t^2}{4}$; $\mathrm{GELU}(t) \approx \frac{t}{2} + \frac{t^2}{\sqrt{2\pi}}$
10. $\ell = \zeta(t) - yt = \max(t, 0) - yt + \log(1 + e^{-|t|})$; $\dfrac{d\ell}{dt} = \sigma(t) - y$

Worked solutions

Problem 1

> Show that $\zeta'(t) = \sigma(t)$, that $\log \sigma(t) = -\zeta(-t)$, and that $\dfrac{d}{dt} \log \sigma(t) = 1 - \sigma(t) = \sigma(-t)$.

$\zeta'(t) = \dfrac{e^t}{1 + e^t}.$	Chain rule on $\log u$ with $u = 1 + e^t$: the derivative is u'/u.
$\dfrac{e^t}{1 + e^t} = \dfrac{1}{e^{-t} + 1} = \sigma(t).$	Divide numerator and denominator by e^t.
$\log \sigma(t) = -\log(1 + e^{-t}) = -\zeta(-t).$	The log of a reciprocal is minus the log, and $\zeta(-t) = \log(1 + e^{-t})$ by definition.
$\dfrac{d}{dt} \log \sigma(t) = \dfrac{d}{dt}\big[-\zeta(-t)\big] = \zeta'(-t) = \sigma(-t) = 1 - \sigma(t).$	The chain rule through $-t$ brings a factor -1 that cancels the minus sign; then step 2 at $-t$, and the identity from the regression page.
$\zeta'(t) = \sigma(t)$; $\log \sigma(t) = -\zeta(-t)$; $(\log \sigma)'(t) = 1 - \sigma(t) = \sigma(-t)$	The sigmoid is the slope of softplus, and the log-sigmoid is a negated softplus. Both identities return in Problem 10, where they make the cross-entropy loss safe to compute for large logits. The slope of $\log \sigma$ lies in $(0, 1)$: near 1 for very negative t, where $\log \sigma(t) \approx t$, and near 0 for large t, where $\sigma(t) \approx 1$.

Problem 2

> Show that $\tanh'(t) = 1 - \tanh^2(t)$, that $\tanh t = 2\sigma(2t) - 1$, and hence that $\tanh'(t) = 4\sigma'(2t)$. What values can the two derivatives take?

$\tanh t = s/c$ with $s = \sinh t = \frac{1}{2}(e^t - e^{-t})$ and $c = \cosh t = \frac{1}{2}(e^t + e^{-t})$, and $s' = c$, $c' = s$.	Differentiate the exponentials term by term; the minus sign in s turns into a plus and back.
$\tanh' = \dfrac{c \cdot c - s \cdot s}{c^2} = 1 - \dfrac{s^2}{c^2} = 1 - \tanh^2.$	Quotient rule, then split the fraction.

$2\sigma(2t) - 1 = \dfrac{2}{1 + e^{-2t}} - 1 = \dfrac{1 - e^{-2t}}{1 + e^{-2t}}.$	Put both terms over the common denominator.
$\dfrac{1 - e^{-2t}}{1 + e^{-2t}} = \dfrac{e^{t} - e^{-t}}{e^{t} + e^{-t}} = \tanh t.$	Multiply numerator and denominator by e^{t}, using $e^{t}e^{-2t} = e^{-t}$.
$\tanh'(t) = \dfrac{d}{dt}\left[2\sigma(2t) - 1\right] = 2 \cdot 2\,\sigma'(2t) = 4\sigma'(2t).$	Chain rule through the inner $2t$.
$\tanh'(t) = 1 - \tanh^2(t) = 4\sigma'(2t)$; $\tanh' \in (0, 1]$ **with its maximum** 1 **at** $t = 0$, **and** $\sigma' \in (0, \tfrac{1}{4}]$ **with its maximum at** $t = 0$	$\tanh$ is the sigmoid stretched vertically to $(-1, 1)$ and compressed horizontally by 2, so its slope is four times the sigmoid's at the matching point. Both derivatives are written in terms of the output in Problem 7, and the bounds are what Problem 8 uses: a backward step through either activation can shrink a gradient but never grow it.

Problem 3

> Compute $\mathrm{relu}'(t)$ and $f_\alpha'(t)$ for $t \neq 0$, and say what happens at $t = 0$. Show that $\mathrm{relu}(ct) = c\,\mathrm{relu}(t)$ for $c > 0$ and what that says about the derivative. Then write both derivatives in terms of the output h.

For $t > 0$, $\mathrm{relu}(t) = t$ and $\mathrm{relu}'(t) = 1$; for $t < 0$, $\mathrm{relu}(t) = 0$ and $\mathrm{relu}'(t) = 0$.	On each side of 0 the function is a straight line, with slopes 1 and 0.
$f_\alpha'(t) = 1$ for $t > 0$ and α for $t < 0$.	The same, with slope α on the left.
At $t = 0$ the one-sided slopes differ, so neither function is differentiable there; for the ReLU every value in $[0, 1]$ is a subgradient, and frameworks use $\mathrm{relu}'(0) = 0$.	A convex function's subgradients at a kink are the slopes between the two one-sided slopes; for a continuous pre-activation $P(t = 0) = 0$, so the choice has no effect on training.

For $c > 0$, $\text{relu}(ct) = \max(ct, 0) = c\max(t, 0) = c\,\text{relu}(t)$; differentiating in t gives $c\,\text{relu}'(ct) = c\,\text{relu}'(t)$, so $\text{relu}'(ct) = \text{relu}'(t)$.	A positive factor comes out of a maximum; the chain rule on the left and the constant c on the right.
For the ReLU, $h > 0$ exactly when $t > 0$, so $\text{relu}'(t) = 1[h > 0]$; for the leaky ReLU, h has the sign of t because $\alpha > 0$, so $f'_\alpha(t) = 1[h > 0] + \alpha 1[h < 0]$.	Each derivative depends on t only through its sign, and the output carries the sign.
$\text{relu}'(t) = 1[t > 0] = 1[h > 0]$ **and** $f'_\alpha(t) = 1[t > 0] + \alpha 1[t < 0] = 1[h > 0] + \alpha 1[h < 0]$, **with the value at** 0 **a convention;** $\text{relu}(ct) = c\,\text{relu}(t)$ **for** $c > 0$, **so the derivative is constant along each ray**	The derivative is a mask, so a ReLU backward pass needs only one bit per unit. Positive homogeneity is why a ReLU network's output is piecewise linear in its input and why the scale of one layer's weights can be moved into the next layer's. A unit whose pre-activation is negative for every input in the data has mask 0 everywhere, receives no gradient, and never recovers: the dead ReLU that the leaky slope α exists to prevent.

Problem 4

> SiLU is $f(t) = t\,\sigma(t)$. Compute $f'(t)$, write it in terms of the output $h = f(t)$ and $\sigma(t)$, and evaluate $f'(-2)$ to show that SiLU is not monotonic.

$f'(t) = \sigma(t) + t\,\sigma'(t)$.	Product rule.
$f'(t) = \sigma(t)\big(1 + t(1 - \sigma(t))\big)$.	$\sigma' = \sigma(1 - \sigma)$ from the regression page's Problem 4, then factor out $\sigma(t)$.
$t\,\sigma(t)(1 - \sigma(t)) = h(1 - \sigma(t))$, so $f'(t) = \sigma(t) + h - h\sigma(t) = h + \sigma(t)(1 - h)$.	Replace $t\,\sigma(t)$ by h in the second term of step 1 and collect.

At $t=-2$: $\sigma(-2)=1/(1+e^{2})\approx 0.1192$, so $1-\sigma(-2)\approx 0.8808$ and $f'(-2)\approx 0.1192\,(1-2\cdot 0.8808)=0.1192\cdot(-0.7616)\approx -0.0908$.	Step 2 with $e^{2}\approx 7.389$.
$f'(t)=\sigma(t)\big(1+t(1-\sigma(t))\big)=h+\sigma(t)(1-h)$; $f'(-2)\approx -0.091<0$, **so SiLU decreases there**	SiLU has a minimum of about -0.278 near $t\approx -1.28$ and rises back toward 0 as $t\to-\infty$, so its slope is negative on the far left. For positive t the derivative exceeds 1 (at $t=2$ it is about 1.09), so unlike the sigmoid and $\tanh$ it can enlarge a gradient. The output form still contains $\sigma(t)$, which is not a function of h alone; Problem 7 says why.

Problem 5

> Compute $\mathrm{GELU}'(t)$ for $\mathrm{GELU}(t)=t\,\Phi(t)$. Then differentiate the tanh approximation $g(t)=\frac{1}{2}t\big(1+\tanh u(t)\big)$ with $u(t)=\sqrt{2/\pi}\,(t+0.044715\,t^{3})$.

$\mathrm{GELU}'(t)=\Phi(t)+t\,\Phi'(t)=\Phi(t)+t\,\varphi(t).$	Product rule; the derivative of a distribution function is its density.
$g'(t)=\frac{1}{2}\big(1+\tanh u\big)+\frac{1}{2}t\,(1-\tanh^{2}u)\,u'(t).$	Product rule on t times the bracket, then the chain rule through u, with $\tanh'=1-\tanh^{2}$ from Problem 2.
$u'(t)=\sqrt{2/\pi}\,(1+3\cdot 0.044715\,t^{2})=\sqrt{2/\pi}\,(1+0.134145\,t^{2}).$	Power rule on the cubic.

$\mathrm{GELU}'(t) = \Phi(t) + t\,\varphi(t)$ **with** $\varphi(t) = e^{-t^2/2}/\sqrt{2\pi}$; $g'(t) = \frac{1}{2}(1 + \tanh u) + \frac{1}{2} t\,(1 - \tanh^2 u)\sqrt{2/\pi}\,(1 + 0.134145\,t^2)$	Both give $\frac{1}{2}$ at $t = 0$, since $\Phi(0) = \frac{1}{2}$ and $\tanh 0 = 0$. GELU′ tends to 1 as $t \to \infty$ and to 0 as $t \to -\infty$, and like SiLU it is negative on part of the left: at $t = -1.5$, $\Phi(-1.5) \approx 0.067$ while $1.5\,\varphi(1.5) \approx 0.194$, so the derivative is about -0.13. The approximation tracks the exact function to within a few thousandths on $[-3, 3]$, but its derivative needs the inner factor u' (Mistake 5).

Problem 6

> Show that $\zeta(t) - \mathrm{relu}(t) = \log(1 + e^{-|t|})$, which lies in $(0, \log 2]$, so that $\zeta(t) = \max(t, 0) + \log(1 + e^{-|t|})$. Show that $\zeta''(t) = \sigma(t)(1 - \sigma(t)) > 0$, so softplus is convex.

For $t \geq 0$: $\zeta(t) = \log\left(e^t(1 + e^{-t})\right) = t + \log(1 + e^{-t}) = \mathrm{relu}(t) + \log(1 + e^{-\lvert t\rvert})$.	Factor e^t out of $1 + e^t$ inside the log; the log of a product is a sum; $\lvert t\rvert = t$ here.
For $t < 0$: $\zeta(t) = \log(1 + e^t) = \mathrm{relu}(t) + \log(1 + e^{-\lvert t\rvert})$.	$\mathrm{relu}(t) = 0$ and $e^t = e^{-\lvert t\rvert}$.
$0 < \log(1 + e^{-\lvert t\rvert}) \leq \log 2$, with the maximum at $t = 0$ and the value tending to 0 as $\lvert t\rvert \to \infty$.	$e^{-\lvert t\rvert} \in (0, 1]$, and $\log(1 + u)$ is increasing in u with $\log 1 = 0$.
$\zeta'' = \sigma' = \sigma(1 - \sigma) > 0$.	Problem 1 gives $\zeta' = \sigma$, and $\sigma \in (0, 1)$ makes both factors positive.

$\zeta(t) = \max(t, 0) + \log(1 + e^{-\lvert t \rvert})$ **with** $0 < \zeta(t) - \mathrm{relu}(t) \le \log 2$; $\zeta''(t) = \sigma(t)(1 - \sigma(t)) > 0$, **so softplus is convex**	Softplus is the ReLU with its kink rounded off, never more than $\log 2 \approx 0.693$ above it. Computed as $\log(1 + e^t)$, it overflows in double precision once t exceeds about 709, and for t near -40 the e^t vanishes next to the 1; in the stable form $e^{-\lvert t \rvert} \le 1$ never overflows and `log1p` keeps the small values exact. Convexity is the statement that the slope σ only ever increases, from 0 to 1, where the ReLU's jumps.

Problem 7

> For the sigmoid, tanh, ReLU, leaky ReLU, softplus and ELU, write $f'(z)$ as a function of the output $h = f(z)$ alone. Do SiLU and GELU have such a formula?

Sigmoid: $\sigma'(z) = \sigma(z)(1 - \sigma(z)) = h(1 - h)$; tanh: $1 - \tanh^2(z) = 1 - h^2$.	The regression page's Problem 4 and Problem 2 here, with the output substituted.
ReLU: $1[h > 0]$; leaky ReLU: $1[h > 0] + \alpha 1[h < 0]$.	Problem 3.
Softplus: $h = \log(1 + e^z)$ gives $e^h = 1 + e^z$, so $e^z = e^h - 1$ and $\zeta'(z) = \sigma(z) = \dfrac{e^z}{1 + e^z} = \dfrac{e^h - 1}{e^h} = 1 - e^{-h}$.	Invert the softplus by exponentiating, then write σ with e^z in the numerator (Problem 1, step 2 read backwards) and substitute.
ELU: for $z > 0$, $h = z > 0$ and $f' = 1$; for $z < 0$, $h = \alpha(e^z - 1) < 0$ and $f' = \alpha e^z = h + \alpha$.	Differentiate each branch; on the left, $\alpha e^z = \alpha(e^z - 1) + \alpha$. The sign of h says which branch applies. At $z = 0$ the slopes are α from the left and 1 from the right, so the derivative exists only when $\alpha = 1$; otherwise the backward pass picks one by convention, as for ReLU.

A formula in h alone exists only when each output comes from inputs with the same slope. SiLU and GELU are not monotonic (Problems 4 and 5): SiLU takes the value -0.2 at $t \approx -0.55$, where its slope is about 0.24, and again at $t \approx -2.4$, where its slope is about -0.10.	One value of h, two different derivatives, so no function of h can give both.
$\sigma' = h(1-h)$; $\tanh' = 1 - h^2$; $\mathrm{relu}' = 1[h > 0]$; $f'_\alpha = 1[h > 0] + \alpha 1[h < 0]$; $\zeta' = 1 - e^{-h}$; $\mathrm{ELU}' = 1$ **for** $h > 0$ **and** $h + \alpha$ **for** $h < 0$; **SiLU and GELU have no such formula, so their backward pass must keep** z	This is why a framework's sigmoid, tanh and ReLU save their output and free the input, while SiLU and GELU save the input. Each formula is checked numerically against finite differences of the forward function, which is the test to run on a hand-written backward pass as well.

Problem 8

> A layer computes $z = Wx + b \in \mathbb{R}^m$ and $h = f(z)$ elementwise, and a scalar loss L has $dh = \nabla_h L$. Show that $dz = dh \odot f'(z)$. For $f = \tanh$, write $\nabla_W L$ and $\nabla_b L$ using only dh, h and x. Then show $\|dz\| \le \|dh\|$ for tanh and $\|dz\| \le \frac{1}{4}\|dh\|$ for the sigmoid.

$h_i = f(z_i)$ depends on z_i alone, so $\partial h_i / \partial z_j = f'(z_i)$ for $j = i$ and 0 otherwise: $\partial h / \partial z = \mathrm{diag}(f'(z))$.	An elementwise map has a diagonal Jacobian, as on the Jacobians page.
$dz = (\partial h / \partial z)^\top dh = \mathrm{diag}(f'(z))\, dh = dh \odot f'(z)$.	The chain rule for a scalar loss applies the transposed Jacobian to dh; a diagonal matrix is its own transpose, and applying it scales entry i by $f'(z_i)$.
For tanh, $f'(z) = 1 - h^2$, so $dz = dh \odot (1 - h^2)$.	Problem 7.

$z_i = \sum_j W_{ij} x_j + b_i$, so $\partial L/\partial W_{ij} = dz_i\, x_j$ and $\partial L/\partial b_i = dz_i$: $\nabla_W L = dz\, x^\top$ and $\nabla_b L = dz$.	W_{ij} and b_i appear in z_i only, with coefficients x_j and 1; the outer product collects the entries in the shape of W.
$\|dz\|^2 = \sum_i dh_i^2\, f'(z_i)^2 \le \big(\max_i f'(z_i)^2\big)\|dh\|^2$.	Each term is bounded by the largest factor times the same dh_i^2.
$dz = dh \odot f'(z)$; **for** tanh, $\nabla_W L = \big(dh \odot (1 - h^2)\big)x^\top$ **and** $\nabla_b L = dh \odot (1 - h^2)$; $\|dz\| \le \|dh\|$ **for** tanh **and** $\|dz\| \le \frac{1}{4}\|dh\|$ **for the sigmoid**	Problem 2's bounds $\tanh' \le 1$ and $\sigma' \le \frac{1}{4}$ in step 5. Through L sigmoid layers the activations alone shrink a gradient by at least 4^{-L} unless the weights compensate, which is the vanishing-gradient problem the initialization page quantifies; tanh at best preserves it, near 0, and a ReLU passes each surviving entry unchanged and zeroes the rest. SiLU and GELU, whose derivatives exceed 1 on part of the line, obey no such bound.

Problem 9

> For the sigmoid, tanh, softplus, SiLU and GELU, compute $f(0)$, $f'(0)$ and $f''(0)$, and write the quadratic approximation $f(t) \approx f(0) + f'(0)\,t + \frac{1}{2}f''(0)\,t^2$.

$\sigma(0) = \frac{1}{2}$, $\sigma'(0) = \frac{1}{4}$, and $\sigma'' = \sigma'(1 - 2\sigma)$, so $\sigma''(0) = 0$.	Product rule on $\sigma' = \sigma(1 - \sigma)$: $\sigma'' = \sigma'(1 - \sigma) - \sigma\sigma'$.
$\tanh 0 = 0$, $\tanh'(0) = 1$, and $\tanh'' = -2\tanh \cdot \tanh'$, so $\tanh''(0) = 0$.	Chain rule on $1 - \tanh^2$.
$\zeta(0) = \log 2$, $\zeta'(0) = \sigma(0) = \frac{1}{2}$, $\zeta''(0) = \sigma'(0) = \frac{1}{4}$.	Problems 1 and 6.
SiLU: $f(0) = 0$, $f'(0) = \sigma(0) + 0 = \frac{1}{2}$, and $f'' = 2\sigma' + t\sigma''$, so $f''(0) = 2 \cdot \frac{1}{4} = \frac{1}{2}$.	Differentiate $f' = \sigma + t\sigma'$ once more with the product rule.

GELU: $f(0) = 0$, $f'(0) = \Phi(0) = \frac{1}{2}$, and $f'' = 2\varphi + t\varphi'$, so $f''(0) = 2\varphi(0) = 2/\sqrt{2\pi} = \sqrt{2/\pi} \approx 0.798$.	Differentiate $f' = \Phi + t\varphi$; $\varphi' = -t\varphi$ vanishes at 0.
$\sigma(t) \approx \frac{1}{2} + \frac{t}{4}$; $\tanh t \approx t$; $\zeta(t) \approx \log 2 + \frac{t}{2} + \frac{t^2}{8}$; $\mathrm{SiLU}(t) \approx \frac{t}{2} + \frac{t^2}{4}$; $\mathrm{GELU}(t) \approx \frac{t}{2} + \frac{t^2}{\sqrt{2\pi}}$	The sigmoid and $\tanh$ are odd about their center, so they have no t^2 term and are linear maps of gain $\frac{1}{4}$ and 1 near 0 (the stretch of Problem 2). Softplus, SiLU and GELU all have gain $\frac{1}{2}$ at 0, and their t^2 terms are what bends them. The error of each approximation is of order t^3, as on the Taylor-series page; a network whose pre-activations stay small is close to a linear network, which is one reason initialization scale matters.

Problem 10

> The binary cross-entropy of a logit t against a label $y \in \{0, 1\}$ is $\ell = -y \log \sigma(t) - (1 - y) \log(1 - \sigma(t))$. Show that $\ell = \zeta(t) - yt = \max(t, 0) - yt + \log(1 + e^{-|t|})$ and that $d\ell/dt = \sigma(t) - y$. Why is the last form the one libraries compute?

$-\log \sigma(t) = \zeta(-t)$ and $-\log(1 - \sigma(t)) = -\log \sigma(-t) = \zeta(t)$.	Problem 1 twice, with $1 - \sigma(t) = \sigma(-t)$.
$\ell = y\,\zeta(-t) + (1 - y)\,\zeta(t)$.	Substitute step 1 into the definition.
$\zeta(-t) = \zeta(t) - t$.	$\log(1 + e^{-t}) = \log\left(e^{-t}(e^{t} + 1)\right) = -t + \log(1 + e^{t})$.
$\ell = y\big(\zeta(t) - t\big) + (1 - y)\zeta(t) = \zeta(t) - yt$.	Collect the $\zeta(t)$ terms: $y + (1 - y) = 1$.
$\ell = \max(t, 0) - yt + \log(1 + e^{-\lvert t \rvert})$.	Problem 6's stable form of ζ.
$\dfrac{d\ell}{dt} = \zeta'(t) - y = \sigma(t) - y$.	Problem 1.

$\ell = \zeta(t) - yt = \max(t, 0) - yt + \log(1 + e^{-\lvert t \rvert})$; $\dfrac{d\ell}{dt} = \sigma(t) - y$	The definition evaluates $\sigma(t)$ first: at $t = -1000$ it underflows to 0 and $\log 0$ is $-\infty$, so a confident wrong prediction gives an infinite loss and a `nan` gradient. The stable form gives $\ell = 1000$ exactly for $y = 1$ there, with gradient $\sigma(-1000) - 1 = -1$. This is PyTorch's `binary_cross_entropy_with_logits` and Keras's `from_logits=True`; the gradient $\sigma(t) - y$ is the regression page's prediction error, which the one-hidden-layer page takes as its starting δ.

Where this goes wrong

1. Sigmoid derivative computed from the pre-activation

The formula $\sigma' = h(1 - h)$ is remembered with a variable name, and in a backward pass both z and h are to hand.

$dz = dh \odot \sigma'(z)$	Right so far: Problem 8.
"$\sigma'(z) = z(1 - z)$."	The habit that causes the mistake: reading $\sigma(1 - \sigma)$ as a rule about whatever variable is written, when the factor in it is the output $\sigma(z)$.
$dz = dh \odot z \odot (1 - z)$ ✗	$z(1 - z)$ is not the derivative of anything here: at $z = 3$ it gives -6 where $\sigma'(3) \approx 0.045$, and it is negative whenever z lies outside $[0, 1]$. The factor is $h(1 - h)$ with $h = \sigma(z)$, always in $(0, \frac{1}{4}]$ (Problem 7). The error hides as long as the pre-activations happen to sit in $(0, 1)$; a finite-difference check at any z outside that interval exposes it at once.

2. ReLU backward multiplied by the output instead of its mask

The output of a ReLU is 0 where the gradient should be blocked and positive where it should pass, which looks like a mask.

$dz = dh \odot \mathrm{relu}'(z)$	Right so far: Problem 8.
"h is zero where the unit is off and positive where it is on, so it is the mask."	The shortcut that causes the mistake: h has the right zeros, but it is not an indicator.
$dz = dh \odot h$ ✗	The derivative is $1[z > 0]$, which is 1 on every active unit; h instead scales each active gradient by the unit's own value, so a unit at $z = 0.01$ passes almost nothing and a unit at $z = 50$ passes fifty times its gradient. The line is the backward pass of $\frac{1}{2}\,\mathrm{relu}(z)^2$, not of $\mathrm{relu}(z)$. The right mask is $1[h > 0]$ (Problem 3).

3. SiLU differentiated with the sigmoid held constant

SiLU is described as the input gated by a sigmoid, and a gate is easy to treat as fixed.

$f(t) = t\,\sigma(t)$	Right so far: the definition.
"The gate $\sigma(t)$ multiplies t, so the derivative is the gate."	The analogy that causes the mistake: a gate computed from a different variable, as in a recurrent cell, is a constant with respect to t; this one is a function of t.
$f'(t) = \sigma(t)$ ✗	The product rule's second term $t\,\sigma(t)(1 - \sigma(t))$ is missing, and it is not small: at $t = 2$ it is about 0.21, and at $t = -2$ it makes the true derivative -0.091 (Problem 4) where $\sigma(-2) \approx 0.12$ is positive. The wrong formula says SiLU is increasing everywhere, which it is not.

4. Softplus stabilized without the max(t, 0)

The stable identity ends in $\log(1 + e^{-|t|})$, and the $e^{-|t|}$ is visibly the part that stops the overflow.

$\zeta(t) = \max(t, 0) + \log(1 + e^{-\lvert t \rvert})$	Right so far: Problem 6.
"The overflow came from e^t, so swapping it for $e^{-\lvert t \rvert}$ is the whole fix."	The shortcut that causes the mistake: the $\lvert t \rvert$ only appears after e^t is factored out, and the factor's logarithm is the $\max(t, 0)$ term.
$\zeta(t) = \log(1 + e^{-\lvert t \rvert})$ ✗	For $t > 0$ this is $\zeta(-t) = \zeta(t) - t$ (Problem 10, step 3): it decreases toward 0 for large t instead of growing like t, giving 0.0067 at $t = 5$ in place of 5.0067. It agrees with softplus for every $t \le 0$, where the dropped term is 0, so it passes any test that only uses negative inputs.

5. GELU's tanh approximation differentiated without the inner derivative

The approximation is $\frac{1}{2}t(1 + \tanh u)$ with u a short expression that is easy to treat as the variable itself.

$g(t) = \frac{1}{2}t\,(1 + \tanh u)$ with $u = \sqrt{2/\pi}\,(t + 0.044715\,t^3)$	Right so far: the definition.
"Differentiate as $t \tanh t$: product rule, with $\tanh' = 1 - \tanh^2$."	The habit that causes the mistake: $u \approx 0.8t$ for small t looks close enough to t to be differentiated as if it were t.
$g'(t) = \frac{1}{2}(1 + \tanh u) + \frac{1}{2}t\,(1 - \tanh^2 u)$ ✗	The chain rule puts the factor $u'(t) = \sqrt{2/\pi}\,(1 + 0.134145\,t^2)$ on the second term (Problem 5). That factor is 0.80 at $t = 0$ and 1.76 at $t = 3$, so the term is off by 20% near zero and by three-quarters further out, and the gradients of every GELU layer in the network inherit the error.

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 4 of *Deep Learning Math by Hand*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/activation-functions-and-their-derivatives