Introduction and learning objectives

This tutorial connects two parts of neural-network learning: computing gradients with backpropagation and using those gradients to update the parameters. We first work backwards through an MLP, then study what gradient descent can guarantee.

By the end of the tutorial, you should be able to:

- Choose a suitable loss for binary classification and compute error signals backwards through an MLP using the chain rule.
- Write each weight gradient as error signal × input, and combine the gradients over the training examples.
- Use Taylor's theorem to derive a step-size condition under which gradient descent decreases the loss.
- Derive and interpret the $O(1/T)$ bound on the smallest squared gradient seen in the first T iterations.

Optional extension: derive the corresponding SGD bound and explain the term that remains when the step size is fixed.

1 Backpropagation with Multilayer Perceptrons

Throughout, $y_n = z_{n,1}^{(4)} = \sigma_4(a_{n,1}^{(4)})$ is the network's output for example n, and M_l is the number of units in layer l.

Problem 1. The target is binary and the sigmoid output lies in $(0, 1)$, so we interpret y_n as $p(t_n = 1|\mathbf{x}_n, \theta)$. Then t_n given $\mathbf{x}_n$ is Bernoulli,

$$p(t_n|\mathbf{x}_n, \theta) = y_n^{t_n}(1 - y_n)^{1-t_n},$$

and, assuming the examples are independent given their inputs, maximising the likelihood is the same as minimising the negative log-likelihood, a sum over examples. This gives the binary cross-entropy loss, the same loss as in logistic regression:

$$\boxed{\mathcal{L}(\theta; \mathcal{D}) = \sum_{n=1}^{N} \ell_n, \qquad \ell_n = -\big(t_n \log y_n + (1 - t_n) \log(1 - y_n)\big), \qquad y_n = \sigma_4\big(a_{n,1}^{(4)}\big).}$$

Problem 2.1. Apply the chain rule through y_n:

$$\delta_{n,1}^{(4)} = \frac{\partial \ell_n}{\partial a_{n,1}^{(4)}} = \frac{\partial \ell_n}{\partial y_n} \cdot \frac{\partial y_n}{\partial a_{n,1}^{(4)}}.$$

The first factor is

$$\frac{\partial \ell_n}{\partial y_n} = -\frac{t_n}{y_n} + \frac{1 - t_n}{1 - y_n} = \frac{y_n - t_n}{y_n(1 - y_n)}.$$

For the second factor, differentiate $\sigma(a) = (1 + e^{-a})^{-1}$:

$$\sigma'(a) = \frac{e^{-a}}{(1+e^{-a})^2} = \sigma(a)\big(1 - \sigma(a)\big), \qquad \text{so} \qquad \frac{\partial y_n}{\partial a^{(4)}_{n,1}} = y_n(1 - y_n).$$

Multiplying, the factor $y_n(1 - y_n)$ cancels:

$$\boxed{\delta^{(4)}_{n,1} = y_n - t_n.}$$

Problem 2.2. The weight $w^{(3)}_{1,k}$ enters the loss only through $a^{(4)}_{n,1} = \sum_{k'=1}^{M_3} w^{(3)}_{1,k'} z^{(3)}_{n,k'}$, and $\partial a^{(4)}_{n,1}/\partial w^{(3)}_{1,k} = z^{(3)}_{n,k}$. By the chain rule,

$$\boxed{\frac{\partial \ell_n}{\partial w^{(3)}_{1,k}} = \frac{\partial \ell_n}{\partial a^{(4)}_{n,1}} \cdot \frac{\partial a^{(4)}_{n,1}}{\partial w^{(3)}_{1,k}} = \delta^{(4)}_{n,1}\, z^{(3)}_{n,k}.}$$

This is the error signal of the receiving unit times the input carried by the weight.

Problem 2.3. $a^{(3)}_{n,k}$ affects the loss only through $z^{(3)}_{n,k}$, which feeds into the single output unit through the weight $w^{(3)}_{1,k}$. The chain rule follows this path:

$$\delta^{(3)}_{n,k} = \frac{\partial \ell_n}{\partial a^{(3)}_{n,k}} = \frac{\partial \ell_n}{\partial a^{(4)}_{n,1}} \cdot \frac{\partial a^{(4)}_{n,1}}{\partial z^{(3)}_{n,k}} \cdot \frac{\partial z^{(3)}_{n,k}}{\partial a^{(3)}_{n,k}}.$$

Here $\partial a^{(4)}_{n,1}/\partial z^{(3)}_{n,k} = w^{(3)}_{1,k}$, and with the ReLU convention in the question, $\partial z^{(3)}_{n,k}/\partial a^{(3)}_{n,k} = \mathbb{I}[a^{(3)}_{n,k} \geq 0]$. Therefore

$$\boxed{\delta^{(3)}_{n,k} = \delta^{(4)}_{n,1}\, w^{(3)}_{1,k}\, \mathbb{I}[a^{(3)}_{n,k} \geq 0].}$$

Problem 2.4. $a^{(2)}_{n,j}$ affects the loss through $z^{(2)}_{n,j}$, which feeds into *every* unit of layer 3, through the weights $w^{(2)}_{k,j}$ for $k = 1, \ldots, M_3$. The chain rule sums over these M_3 paths:

$$\delta^{(2)}_{n,j} = \frac{\partial \ell_n}{\partial a^{(2)}_{n,j}} = \sum_{k=1}^{M_3} \frac{\partial \ell_n}{\partial a^{(3)}_{n,k}} \cdot \frac{\partial a^{(3)}_{n,k}}{\partial z^{(2)}_{n,j}} \cdot \frac{\partial z^{(2)}_{n,j}}{\partial a^{(2)}_{n,j}}.$$

As in 2.3, $\partial a^{(3)}_{n,k}/\partial z^{(2)}_{n,j} = w^{(2)}_{k,j}$ and $\partial z^{(2)}_{n,j}/\partial a^{(2)}_{n,j} = \mathbb{I}[a^{(2)}_{n,j} \geq 0]$. The last factor does not depend on k, so it moves out of the sum:

$$\boxed{\delta^{(2)}_{n,j} = \mathbb{I}[a^{(2)}_{n,j} \geq 0] \sum_{k=1}^{M_3} w^{(2)}_{k,j}\, \delta^{(3)}_{n,k}.}$$

Problem 2.5. As in 2.2, $w^{(1)}_{j,i}$ enters ℓ_n only through $a^{(2)}_{n,j} = \sum_{i'=1}^{M_1} w^{(1)}_{j,i'} z^{(1)}_{n,i'}$, and the input to the first layer is the data, $z^{(1)}_{n,i} = x_{n,i}$. So

$$\frac{\partial \ell_n}{\partial w^{(1)}_{j,i}} = \frac{\partial \ell_n}{\partial a^{(2)}_{n,j}} \cdot \frac{\partial a^{(2)}_{n,j}}{\partial w^{(1)}_{j,i}} = \delta^{(2)}_{n,j}\, x_{n,i}.$$

The weights are shared by all examples and $\mathcal{L} = \sum_n \ell_n$, so the gradient is the sum of the per-example gradients:

$$\boxed{\frac{\partial \mathcal{L}}{\partial w_{j,i}^{(1)}} = \sum_{n=1}^{N} \delta_{n,j}^{(2)}\, x_{n,i}.}$$

The general pattern. Every weight gradient has the form

$$\frac{\partial \mathcal{L}}{\partial w_{j,i}^{(l)}} = \sum_{n=1}^{N} \underbrace{\delta_{n,j}^{(l+1)}}_{\text{error signal}} \cdot \underbrace{z_{n,i}^{(l)}}_{\text{input}},$$

and the error signals are computed backwards, from the output layer to the first hidden layer:

$$\delta_{n,i}^{(l)} = \sigma_l'\big(a_{n,i}^{(l)}\big) \sum_{j=1}^{M_{l+1}} w_{j,i}^{(l)}\, \delta_{n,j}^{(l+1)}.$$

Problems 2.3 and 2.4 are this recursion with σ_3' and σ_2' the ReLU derivative. For example, the gradient for the middle layer, which the question did not ask for, is $\partial \mathcal{L}/\partial w_{k,j}^{(2)} = \sum_n \delta_{n,k}^{(3)}\, z_{n,j}^{(2)}$.

2 Gradient Descent

Assumptions. We use Taylor's theorem for a twice continuously differentiable f with the question's global upper bound $f''(w) < L$. From Problem 3 onward, assume the stated minimum f^* is finite. The gradient-norm bounds in Problems 3–5 also hold with any finite lower bound in place of f^*. The later convex objective-gap bounds use $f^* = f(w^*)$ at a minimizer. Step sizes are positive, and the stopping tolerance η below is positive.

Problem 1. Apply Taylor's Theorem with $h = f$, $x_0 = w_t$ and $x = w_{t+1}$. By the update rule, $x - x_0 = w_{t+1} - w_t = -\alpha f'(w_t)$. So for some ξ (between w_t and w_{t+1}),

$$\begin{aligned} f(w_{t+1}) &= f(w_t) + f'(w_t)\big(-\alpha f'(w_t)\big) + \frac{f''(\xi)}{2}\big(-\alpha f'(w_t)\big)^2 \\ &= f(w_t) - \alpha f'(w_t)^2 + \frac{f''(\xi)}{2}\alpha^2 f'(w_t)^2 \\ &\le f(w_t) - \alpha f'(w_t)^2 + \frac{L}{2}\alpha^2 f'(w_t)^2 \\ &= f(w_t) - \alpha\left(1 - \frac{\alpha L}{2}\right) f'(w_t)^2. \end{aligned}$$

The inequality uses $f''(\xi) < L$ and $\alpha^2 f'(w_t)^2 \ge 0$.

Problem 2. If $0 < \alpha < 1/L$, then $\alpha L/2 < 1/2$, so $1 - \alpha L/2 > 1/2$. Rearranging Problem 1,

$$f(w_t) - f(w_{t+1}) \ge \alpha\left(1 - \frac{\alpha L}{2}\right) f'(w_t)^2 \ge \frac{\alpha}{2} f'(w_t)^2,$$

where the second inequality uses $1 - \alpha L/2 > 1/2$ and $f'(w_t)^2 \ge 0$. Each step decreases f by at least $\frac{\alpha}{2} f'(w_t)^2$: the loss never increases, and it strictly decreases unless $f'(w_t) = 0$. (Problem 1 gives a decrease for any $0 < \alpha < 2/L$; the condition $\alpha < 1/L$ gives the constant $1/2$.)

Problem 3. Sum the inequality from Problem 2 over $t = 0, \dots, T-1$. The right-hand side telescopes:

$$\begin{aligned} \frac{\alpha}{2}\sum_{t=0}^{T-1} f'(w_t)^2 &\le \sum_{t=0}^{T-1}\big(f(w_t) - f(w_{t+1})\big) \\ &= f(w_0) - f(w_T) \\ &\le f(w_0) - f^*, \end{aligned}$$

where the last inequality holds because $f^* \le f(w)$ for every w, in particular for $w = w_T$.

Problem 4. The minimum of T numbers is at most their average, since each number is at least the minimum. Using Problem 3,

$$\min_{t\in\{0,\dots,T-1\}} f'(w_t)^2 \le \frac{1}{T}\sum_{t=0}^{T-1} f'(w_t)^2 \le \frac{1}{T}\cdot\frac{2}{\alpha}\big(f(w_0) - f^*\big) = \boxed{\frac{2(f(w_0) - f^*)}{\alpha T}.}$$

The bound is $O(1/T)$: to guarantee an iterate with $f'(w_t)^2 \le \epsilon$, it is enough to run $T \ge 2(f(w_0)-f^*)/(\alpha\epsilon)$ iterations. The guarantee concerns the gradient only. Since f is not assumed to be convex, a small gradient can also occur at a local minimum, a local maximum, a stationary inflection point (such as $w=0$ for $f(w)=w^3e^{-w^2}$) or a flat region, not only at a global minimum. This example has a bounded second derivative and an attained finite minimum, so it satisfies the assumptions above for a sufficiently large L.

Problem 5 (Optional). *Derivation.* Apply Taylor's Theorem as in Problem 1, now with $w_{t+1}-w_t = -\alpha g_t$:

$$f(w_{t+1}) = f(w_t) - \alpha f'(w_t)g_t + \frac{f''(\xi)}{2}\alpha^2 g_t^2 \le f(w_t) - \alpha f'(w_t)g_t + \frac{L}{2}\alpha^2 g_t^2.$$

Take the expectation conditional on w_t. Given w_t, the quantities $f(w_t)$ and $f'(w_t)$ are fixed, so by the two assumptions on g_t,

$$\begin{aligned}\mathbb{E}[f(w_{t+1})|w_t] &\le f(w_t) - \alpha f'(w_t)\,\mathbb{E}[g_t|w_t] + \frac{L}{2}\alpha^2\,\mathbb{E}[g_t^2|w_t] \\ &\le f(w_t) - \alpha f'(w_t)^2 + \frac{L}{2}\alpha^2\sigma^2.\end{aligned}$$

Take the expectation over w_t (law of total expectation) and rearrange:

$$\alpha\,\mathbb{E}[f'(w_t)^2] \le \mathbb{E}[f(w_t)] - \mathbb{E}[f(w_{t+1})] + \frac{L}{2}\alpha^2\sigma^2.$$

Sum over $t=0,\dots,T-1$. The first two terms telescope; take w_0 fixed (not random), and use $\mathbb{E}[f(w_T)] \ge f^*$:

$$\alpha\sum_{t=0}^{T-1}\mathbb{E}[f'(w_t)^2] \le f(w_0) - \mathbb{E}[f(w_T)] + \frac{L}{2}\alpha^2\sigma^2 T \le f(w_0) - f^* + \frac{L}{2}\alpha^2\sigma^2 T.$$

Divide by αT and bound the minimum by the average, as in Problem 4:

$$\boxed{\min_{t\in\{0,\dots,T-1\}}\mathbb{E}[f'(w_t)^2] \le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f'(w_t)^2] \le \frac{f(w_0)-f^*}{\alpha T} + \frac{\alpha\sigma^2 L}{2}.}$$

This derivation does not need $\alpha < 1/L$, because the second-order term is bounded with σ^2 instead of being compared with the first-order term.

Where the second term comes from. In both derivations, the second-order Taylor term is at most $\frac{L}{2}(w_{t+1}-w_t)^2$.

- For gradient descent, $w_{t+1}-w_t = -\alpha f'(w_t)$, so the bound is $\frac{L}{2}\alpha^2 f'(w_t)^2$. It is proportional to the squared gradient, so it shrinks as the gradient shrinks, and for $\alpha < 1/L$ the first-order decrease $\alpha f'(w_t)^2$ outweighs it (Problem 2).
- For SGD, $w_{t+1}-w_t = -\alpha g_t$, so the bound is $\frac{L}{2}\alpha^2 g_t^2$, with conditional mean $\frac{L}{2}\alpha^2\mathbb{E}[g_t^2|w_t]$. Since $\mathbb{E}[g_t|w_t] = f'(w_t)$, we have $\mathbb{E}[g_t^2|w_t] = f'(w_t)^2 + \mathrm{Var}(g_t|w_t)$, and the variance part need not shrink when the gradient does. The assumption bounds $\mathbb{E}[g_t^2|w_t]$ by the constant σ^2, so in the bound each step adds at most $\frac{L}{2}\alpha^2\sigma^2$ to the expected loss, whatever the size of the gradient.

Over T steps these per-step terms sum to $\frac{L}{2}\alpha^2\sigma^2 T$. Dividing by αT leaves $\frac{\alpha\sigma^2 L}{2}$, which does not depend on T. With a fixed step size, the bound therefore does not go to zero as $T \to \infty$. A smaller α lowers this term but slows the decrease of the first term. If the number of steps T is fixed in advance, choosing $\alpha = c/\sqrt{T}$ for a constant $c > 0$ makes both terms $O(1/\sqrt{T})$:

$$\frac{f(w_0) - f^*}{c\sqrt{T}} + \frac{c\,\sigma^2 L}{2\sqrt{T}}.$$

Further remarks (beyond the question sheet). *A stopping rule.* Take $0 < \alpha < 1/L$ as in Problems 2–4, and stop gradient descent as soon as a step is small, $|w_{t+1} - w_t| \le \eta$. By Problem 4 with $T_0 = \max\{1, \lceil 2\alpha(f(w_0) - f^*)/\eta^2 \rceil\}$, some $t \in \{0, \dots, T_0 - 1\}$ satisfies

$$|w_{t+1} - w_t| = \alpha|f'(w_t)| \le \alpha\sqrt{\frac{2(f(w_0) - f^*)}{\alpha T_0}} \le \eta,$$

so the rule stops within T_0 steps. The last inequality uses $T_0 \ge 2\alpha(f(w_0) - f^*)/\eta^2$.

Convex functions. A differentiable f is convex if and only if $f(y) \ge f(x) + f'(x)(y - x)$ for all x, y: the function lies on or above each of its tangent lines. Let w^* be a minimiser. Taking $x = w_t$ and $y = w^*$ gives $f(w_t) - f(w^*) \le f'(w_t)(w_t - w^*)$. Expanding the update,

$$(w_t - w^*)^2 - (w_{t+1} - w^*)^2 = (w_t - w^*)^2 - (w_t - w^* - \alpha f'(w_t))^2 = 2\alpha f'(w_t)(w_t - w^*) - \alpha^2 f'(w_t)^2,$$

so

$$f'(w_t)(w_t - w^*) = \frac{(w_t - w^*)^2 - (w_{t+1} - w^*)^2}{2\alpha} + \frac{\alpha f'(w_t)^2}{2}.$$

Average over $t = 0, \dots, T - 1$; the first part telescopes and $(w_T - w^*)^2 \ge 0$:

$$\min_{t \in \{0,\dots,T-1\}} f(w_t) - f^* \le \frac{1}{T}\sum_{t=0}^{T-1}\big(f(w_t) - f^*\big) \le \frac{1}{T}\sum_{t=0}^{T-1} f'(w_t)(w_t - w^*) \le \frac{(w_0 - w^*)^2}{2\alpha T} + \frac{\alpha}{2T}\sum_{t=0}^{T-1} f'(w_t)^2.$$

For this convex bound, suppose $|f'(w)| \le K$ for all w, with $K > 0$. The preceding distance calculation holds for every $\alpha > 0$, without a curvature bound or the restriction $\alpha < 1/L$. The last sum is at most TK^2, and for $w_0 \ne w^*$ the step size $\alpha = |w_0 - w^*|/(K\sqrt{T})$ gives

$$\min_{t \in \{0,\dots,T-1\}} f(w_t) - f^* \le \frac{(w_0 - w^*)^2}{2\alpha T} + \frac{\alpha K^2}{2} = \frac{K|w_0 - w^*|}{\sqrt{T}}.$$

So for convex f with $|f'| \le K$ and this step size, the best loss value found in the first T steps is within $O(1/\sqrt{T})$ of the minimum f^*. This is a statement about the loss value, not the gradient as in Problem 4. The step size uses $|w_0 - w^*|$; any $R \ge |w_0 - w^*|$ in its place gives the bound $KR/\sqrt{T}$. Under the assumptions of Problems 1–2 instead ($f'' < L$ and $0 < \alpha < 1/L$), convexity gives the faster rate $f(w_T) - f^* \le (w_0 - w^*)^2/(2\alpha T)$, which is $O(1/T)$ (a standard result, not derived here).