Source task and conventions

Answer all three parts, including the derivative convention at zero and the optimizer assumptions.

Conventional solution and proof

Write $a_i(x) = b_i + \sum_{j=1}^{D} w_{ij} x_j$ and $z_i = \max(0, a_i)$. (a) A linear function on $\{0,1\}^D$ is maximized by setting $x_j = 1$ for $w_{ij} > 0$ and zero for $w_{ij} < 0$. Hence

$$\boxed{z_i \text{ is zero for every input } \iff b_i + \sum_j \max(w_{ij}, 0) \le 0.}$$

The same maximum holds on the full cube $[0,1]^D$.

(b) Let the outgoing weight be $v_i = \theta_i^{(z \to y)}$ and $\partial L / \partial y = 1$. At $a_i \neq 0$, the chain rule gives

$$\frac{\partial L}{\partial b_i} = v_i \mathbf{1}\{a_i > 0\}, \qquad \frac{\partial L}{\partial w_{ij}} = v_i x_j \mathbf{1}\{a_i > 0\}, \qquad \frac{\partial L}{\partial v_i} = z_i.$$

At $a_i = 0$, ReLU has no classical derivative. A chosen subgradient $\eta \in [0,1]$ replaces the indicator; the common convention is $\eta = 0$.

(c) Under that zero-at-zero convention, a dead unit has zero incoming-weight, bias and outgoing-weight data gradients on every permitted input. Plain gradient descent or stochastic gradient descent without an additional update term therefore leaves all these parameters unchanged. Induction over optimization steps proves it remains dead. With strictly negative maximum preactivation, all input points are differentiable inactive points. If equality occurs in (a) and a nonzero subgradient is chosen at zero, the permanence statement is not automatic. Stored momentum, injected noise or other parameter updates also fall outside the pure zero-gradient argument.