# Solutions to Problems 2–5

Deep Learning & Mathematics — Step-by-Step Derivations

---

## 1. Problem 2: Vector Calculus Review

Throughout, for a scalar $f(x)$ with $x \in \mathbb{R}^n$, the gradient $\frac{\partial f}{\partial x}$ is taken to be a *row* vector.

Let $x, c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$.

### (a) $f(x) = x^\top c$

Entry-wise, $\frac{\partial f}{\partial x_i} = c_i$. Hence

$$\frac{\partial f}{\partial x} = c^\top.$$

### (b) $g(x) = \|x\|_2^2 = \sum_{i=1}^n x_i^2$

Entry-wise, $\frac{\partial g}{\partial x_i} = 2x_i$. Hence

$$\frac{\partial g}{\partial x} = 2x^\top.$$

### (c) $h(x) = Ax$

The Jacobian has $i$-th row $\frac{\partial (a_i^\top x)}{\partial x} = a_i^\top$ (row $i$ of $A$). Thus

$$\frac{\partial (Ax)}{\partial x} = A.$$

### (d) $q(x) = x^\top Ax$

Write $q = \sum_{i,j} x_i A_{ij} x_j$. Then

$$\frac{\partial q}{\partial x_k} = \sum_j A_{kj} x_j + \sum_i x_i A_{ik} = (Ax)_k + (A^\top x)_k,$$

so

$$\frac{\partial q}{\partial x} = x^\top (A + A^\top).$$

### (e) When is this equal to $2x^\top A$?

Exactly when $A$ is symmetric: $A = A^\top$.

## 2. Problem 3: Least Squares and the Min-Norm Problem via SVD

Let $X \in \mathbb{R}^{m \times n}$ with SVD $X = U\Sigma V^\top$. Let $\Sigma^\dagger$ denote the pseudoinverse of $\Sigma$ obtained by inverting nonzero singular values.

### (a) Overdetermined least squares

Solve $\min_w \|Xw - y\|_2^2$. A least-squares minimizer is $w^\star = X^\dagger y$; when $X$ has full column rank, $w^\star = (X^\top X)^{-1} X^\top y$.

### (b) SVD form

Using $X^\top X = V\Sigma^\top \Sigma V^\top$,

$$w^\star = V(\Sigma^\top \Sigma)^{-1}\Sigma^\top U^\top y \;=\; V\Sigma^\dagger U^\top y.$$

### (c) Left-inverse property

Let $A := V\Sigma^\dagger U^\top$. Then

$$AX = V\Sigma^\dagger U^\top U\Sigma V^\top = VIV^\top = I_n,$$

so $A$ is a left inverse when $\mathrm{rank}(X) = n$.

### (d) Underdetermined minimum-norm solution

Solve $\min \|w\|_2^2$ subject to $Xw = y$. The minimum-norm solution is

$$w^\star = X^\top (XX^\top)^{-1} y = V\Sigma^\dagger U^\top y.$$

### (e) Same SVD expression

As above, the solution simplifies to $w^\star = V\Sigma^\dagger U^\top y$.

### (f) Right-inverse property

Let $B := V\Sigma^\dagger U^\top$. Then

$$XB = U\Sigma V^\top V\Sigma^\dagger U^\top = UIU^\top = I_m,$$

so $B$ is a right inverse when $\mathrm{rank}(X) = m$.

## 3. Problem 4: Five Interpretations of Ridge Regression

Given $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, and $\lambda > 0$.

### (a) Optimization viewpoint

$$\min_w \; \|y - Xw\|_2^2 + \lambda\|w\|_2^2 \quad\Longrightarrow\quad (X^\top X + \lambda I)w = X^\top y \quad\Longrightarrow\quad w = (X^\top X + \lambda I)^{-1} X^\top y.$$

### (b) Singular-value viewpoint

With $X = U\Sigma V^\top$,

$$w = V(\Sigma^\top \Sigma + \lambda I)^{-1}\Sigma^\top U^\top y.$$

Along singular direction $i$, the scalar gain is $\dfrac{\sigma_i}{\sigma_i^2 + \lambda}$, which damps small singular values.

### (c) MAP viewpoint

Assume prior $W \sim \mathcal{N}(0, I)$ and model $Y = XW + \sqrt{\lambda}\,N$ with $N \sim \mathcal{N}(0, I)$. Up to constants, the negative log posterior is

$$\frac{1}{2\lambda}\|y - Xw\|_2^2 + \frac{1}{2}\|w\|_2^2,$$

equivalent (after scaling) to ridge.

### (d) Fake-data (row augmentation)

Let

$$\hat{X} = \begin{bmatrix} X \\ \sqrt{\lambda}\,I_d \end{bmatrix}, \qquad \hat{y} = \begin{bmatrix} y \\ 0_d \end{bmatrix}.$$

Then $\|\hat{y} - \hat{X}w\|_2^2 = \|y - Xw\|_2^2 + \lambda\|w\|_2^2$, so OLS on $(\hat{X}, \hat{y})$ gives ridge.

### (e) Fake-features (column augmentation) & min-norm

Set $X^\vee = [\,X \ \ \sqrt{\lambda}\,I_n\,] \in \mathbb{R}^{n\times(d+n)}$ and solve the minimum-norm problem $\min_\eta \|\eta\|_2^2$ s.t. $X^\vee \eta = y$. The minimum-norm solution has first $d$ entries

$$w^\star = X^\top (XX^\top + \lambda I)^{-1} y,$$

the kernel form of ridge.

### (f) Equivalence identity

The identity

$$(X^\top X + \lambda I)^{-1} X^\top = X^\top (XX^\top + \lambda I)^{-1}$$

implies the kernel and primal forms are equal.

### (g) Limits

As $\lambda \to \infty$, $w \to 0$ (complete shrinkage). As $\lambda \to 0$, we recover OLS in the full-column-rank case and $X^\dagger y$ otherwise.

## 4. Problem 5: ReLU Elbow Update under SGD

Consider a scalar ReLU

$$\phi(x) = \max\{0,\, wx + b\}, \qquad \ell(x, y) = \tfrac{1}{2}\big(\phi(x) - y\big)^2,$$

with the subgradient at 0 fixed to 0.

### (a) Basics

The elbow is $e = -\frac{b}{w}$ $(w \neq 0)$. Also

$$\frac{d\ell}{d\phi} = \phi(x) - y, \qquad \frac{\partial \phi}{\partial w} = x\,\mathbf{1}_{\{wx+b>0\}}, \qquad \frac{\partial \phi}{\partial b} = \mathbf{1}_{\{wx+b>0\}},$$

so

$$\frac{\partial \ell}{\partial w} = (\phi(x) - y)\,x\,\mathbf{1}_{\{wx+b>0\}}, \qquad \frac{\partial \ell}{\partial b} = (\phi(x) - y)\,\mathbf{1}_{\{wx+b>0\}}.$$

### (b) One GD step with residual $\phi(x) - y = 1$

With learning rate $\lambda > 0$,

$$w' = w - \lambda x\, \mathbf{1}_{\{wx+b>0\}}, \qquad b' = b - \lambda\, \mathbf{1}_{\{wx+b>0\}}.$$

**Inactive** ($\phi(x) = 0$): no change.
**Active, $w > 0$, $x > 0$:** $w$ decreases, $b$ decreases; slope flattens and elbow typically moves right.
**Active, $w > 0$, $x < 0$:** $w$ increases, $b$ decreases; slope increases, elbow shift ambiguous.
**Active, $w < 0$, $x > 0$:** $w$ becomes more negative, $b$ decreases; elbow tends to move left, magnitude of slope increases.

### (c) One-hidden-layer network

Let $x \in \mathbb{R}$, hidden units $i = 1, \ldots, d$ with preactivations $z_i = w_i x + b_i$, activations $a_i = \max\{0, z_i\}$, and output $\hat{f}(x) = \sum_{i=1}^{d} v_i a_i$. The elbow of unit $i$ is

$$e_i = -\frac{b_i}{w_i} \quad (w_i \neq 0).$$

### (d) One SGD step and new elbow

Let $r = \hat{f}(x) - y$. The gradients are

$$\frac{\partial \ell}{\partial v_i} = r\, a_i, \qquad \frac{\partial \ell}{\partial w_i} = r\, v_i\, \mathbf{1}_{\{z_i>0\}}\, x, \qquad \frac{\partial \ell}{\partial b_i} = r\, v_i\, \mathbf{1}_{\{z_i>0\}}.$$

Updates:

$$v_i' = v_i - \lambda r a_i, \qquad w_i' = w_i - \lambda r v_i\, \mathbf{1}_{\{z_i>0\}}\, x, \qquad b_i' = b_i - \lambda r v_i\, \mathbf{1}_{\{z_i>0\}}.$$

Thus the new elbow is

$$e_i' = -\frac{b_i'}{w_i'} = -\frac{b_i - \lambda r v_i\, \mathbf{1}_{\{z_i>0\}}}{w_i - \lambda r v_i\, \mathbf{1}_{\{z_i>0\}}\, x}.$$

If unit $i$ is inactive at $x$, then $w_i, b_i$ (hence $e_i$) are unchanged for that step.