A6(a)

We have regularized negative log-likelihood function:

$$J(w,b) = \frac{1}{n}\sum_{i=1}^{n} \log\Big(1 + \exp\big(-\,y_i\big(b + x_i^T w\big)\big)\Big) \;+\; \lambda\,\|w\|_2^2.$$

Gradient with respect to w

For a single example (x_i, y_i), define:

$$\ell_i(w,b) = \log\Big(1 + \exp\big(-\,y_i\,(b + x_i^T w)\big)\Big).$$

Let

$$z_i = -\,y_i\,(b + x_i^T w).$$

Then

$$\ell_i(w,b) = \log\big(1 + \exp(z_i)\big).$$

Step 1: Differentiate $\log(1 + \exp(z_i))$ **w.r.t.** w Using the chain rule:

$$\frac{d}{dw}\,\log\big(1 + \exp(z_i)\big) = \frac{1}{1 + \exp(z_i)} \cdot \frac{d}{dw}\,\exp(z_i).$$

Since

$$z_i = -\,y_i\big(b + x_i^T w\big),$$

we have

$$\frac{dz_i}{dw} = -\,y_i\,\frac{d}{dw}\big(b + x_i^T w\big) = -\,y_i\,x_i.$$

Thus,

$$\frac{d}{dw}\,\exp(z_i) = \exp(z_i)\,\big(-\,y_i\,x_i\big).$$

Substituting back:

$$\frac{d}{dw}\,\log\big(1 + \exp(z_i)\big) = \frac{1}{1 + \exp(z_i)}\,\exp(z_i)\,\big(-\,y_i\,x_i\big).$$

Rewriting in terms of the original variable,

$$\frac{d}{dw}\,\log\Big(1 + \exp\big(-\,y_i\,(b + x_i^T w)\big)\Big) = -\,y_i\,x_i\,\frac{\exp\big(-y_i\,(b + x_i^T w)\big)}{1 + \exp\big(-y_i\,(b + x_i^T w)\big)}.$$

Step 2: We observe that

$$\frac{\exp\bigl(-y_i\,(b+x_i^T w)\bigr)}{1+\exp\bigl(-y_i\,(b+x_i^T w)\bigr)} = \frac{\bigl(1+\exp\bigl(-y_i(b+x_i^T w)\bigr)\bigr)-1}{1+\exp\bigl(-y_i(b+x_i^T w)\bigr)} = 1 - \frac{1}{1+\exp\bigl(-y_i(b+x_i^T w)\bigr)}.$$

Since $\mu_i(w,b)$ is defined as

$$\mu_i(w,b) = \frac{1}{1+\exp\bigl(-\,y_i\,(b+x_i^T w)\bigr)},$$

it follows that

$$\frac{\exp\bigl(-y_i\,(b+x_i^T w)\bigr)}{1+\exp\bigl(-y_i\,(b+x_i^T w)\bigr)} = 1-\mu_i(w,b).$$

Hence,

$$-\,y_i\,x_i\,\frac{\exp\bigl(-y_i(b+x_i^T w)\bigr)}{1+\exp\bigl(-y_i(b+x_i^T w)\bigr)} = -\,y_i\,x_i\,\bigl[\,1-\mu_i(w,b)\bigr].$$

Step 3: Sum over all i and include regularization The total cost is

$$J(w,b) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(w,b) + \lambda\,\|w\|_2^2.$$

Therefore,

$$\frac{\partial}{\partial w}\Bigl[\tfrac{1}{n}\sum_{i=1}^{n}\ell_i(w,b)\Bigr] = \frac{1}{n}\sum_{i=1}^{n}\Bigl[-\,y_i\,x_i\bigl(1-\mu_i(w,b)\bigr)\Bigr].$$

The derivative of $\lambda\,\|w\|_2^2$ with respect to w is $2\,\lambda\,w$. Combining these gives

$$\boxed{\nabla_w J(w,b) = \frac{1}{n}\sum_{i=1}^{n}\Bigl[-\,y_i\,x_i\,\bigl(1-\mu_i(w,b)\bigr)\Bigr] \;+\; 2\,\lambda\,w.}$$

■

Gradient with respect to b

Step 1: Differentiate w.r.t. b Again let

$$z_i = -\,y_i\,(b+x_i^T w).$$

Then

$$\frac{dz_i}{db} = -\,y_i,$$

which implies

$$\frac{d}{db}\exp(z_i) = \exp(z_i)\,\bigl(-\,y_i\bigr).$$

Thus,

$$\frac{d}{db}\log\bigl(1+\exp(z_i)\bigr) = \frac{1}{1+\exp(z_i)}\;\exp(z_i)\;\bigl(-\,y_i\bigr).$$

Rewriting in original variables,

$$\frac{d}{db}\;\log\Bigl(1+\exp\bigl(-\,y_i(b+x_i^T w)\bigr)\Bigr) = -\,y_i\;\frac{\exp\bigl(-y_i(b+x_i^T w)\bigr)}{1+\exp\bigl(-y_i(b+x_i^T w)\bigr)}.$$

Using the same observation as before,

$$= -\,y_i\;\bigl[\,1-\mu_i(w,b)\bigr].$$

Step 2: Sum over i and simplify Hence,

$$\frac{\partial}{\partial b}\Bigl[\tfrac{1}{n}\sum_{i=1}^{n}\ell_i(w,b)\Bigr] = -\,\frac{1}{n}\;\sum_{i=1}^{n} y_i\;\bigl[\,1-\mu_i(w,b)\bigr].$$

Since $\lambda\,\|w\|_2^2$ does not involve b,

$$\nabla_b J(w,b) = -\,\frac{1}{n}\;\sum_{i=1}^{n} y_i\;\bigl[\,1-\mu_i(w,b)\bigr].$$

Moving the negative sign inside the summation,

$$\boxed{\nabla_b J(w,b) = \frac{1}{n}\;\sum_{i=1}^{n}\bigl(\mu_i(w,b)-1\bigr)y_i.}$$

■