Problem 9.1

$$\begin{aligned} p(\phi) &= \prod_{j=1}^{J} \mathcal{N}(\phi_j; 0, \sigma_\phi^2) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma_\phi^2}} \right)^J \exp\left\{ -\frac{\sum_{j=1}^{J} \phi_j^2}{2\sigma_\phi^2} \right\} \\ &= (2\pi\sigma_\phi^2)^{-\frac{J}{2}} \exp\left\{ -\frac{\phi^2}{2\sigma_\phi^2} \right\} \end{aligned}$$

By Bayes' theorem, the posterior distribution is:

$$\begin{aligned} P(\phi|x_i, y_i, \beta) &\propto P(\mathbf{y}|\mathbf{x}, \phi, \beta) P(\phi|\sigma_\phi^2), \\ P(\mathbf{y}|\mathbf{x}, \phi, \beta) &= \prod_{i=1}^{N} \mathcal{N}(y_i | f(\mathbf{x_i}, \phi), \beta^{-1}), \end{aligned}$$

and thus:

$$P(\phi|x_i, y_i, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i | f(\mathbf{x_i}, \phi))(2\pi\sigma_\phi^2)^{-\frac{J}{2}} \exp\left\{ -\frac{\phi^2}{2\sigma_\phi^2} \right\} \tag{1}$$

(Here we assume $t \sim \mathcal{N}(f(x; \phi), \beta^{-1})$).
Taking the negative logarithm of (1):

$$\begin{aligned} -\log P(\phi|x_i, y_i, \beta) &= \sum_{i=1}^{I} -\log \left(\prod_{i=1}^{N} \mathcal{N}(y_i | f(\mathbf{x_i}, \phi))(2\pi\sigma_\phi^2)^{-\frac{J}{2}} \exp\left\{ -\frac{\phi^2}{2\sigma_\phi^2} \right\} \right) \\ &= \frac{\beta}{2} \sum_{i=1}^{I} \{(y_i - f(x_i, \phi))^2\} \\ &+ \frac{1}{2\sigma_\phi^2} \phi^2 + (\text{terms unrelated to } \phi)(2) \end{aligned}$$

Maximizing the posterior is equivalent to minimizing (2):

$$\frac{\beta}{2} \sum_{i=1}^{N} \{(y_i - f(x_i, \phi))^2\} + \frac{1}{2\sigma_\phi^2} \phi^T \phi$$

Let $\lambda = \frac{1}{\sigma_\phi^2 \beta}$, which gives the squared error with L_2 regularization:

$$\frac{1}{2} \sum_{i=1}^{N} \{y_i - f(x_i, \phi)\}^2 + \frac{\lambda}{2} \phi^T \phi$$

Problem 9.2

$$L'(\phi) = \frac{1}{2}L(\phi) + \frac{\lambda}{2}\phi^2$$
$$\frac{\partial L'(\phi)}{\partial \phi} = \frac{1}{2}\frac{\partial L(\phi)}{\partial \phi} + \lambda\phi$$

The new gradient is modified to include a term $(\lambda\phi)$ to penalize large weights.

Problem 9.3

$$\begin{aligned}
\tilde{L} &= \sum_{i=1}^{n} \left(\phi_0 + \phi_1(x_i + \epsilon_i) - y_i\right)^2 \\
&= \sum_{i=1}^{n} \left((\phi_0 + \phi_1 x_i - y_i) + \phi_1\epsilon_i\right)^2 \\
&= \sum_{i=1}^{n} \left(\phi_0 + \phi_1 x_i - y_i\right)^2 + \phi_1^2 \sum_{i=1}^{n} \epsilon_i^2 + 2\phi_1 \sum_{i=1}^{n} (\phi_0 + \phi_1 x_i - y_i)\epsilon_i \\
\mathbb{E}(\tilde{L}) &= \sum_{i=1}^{n} \left(\phi_0 + \phi_1 x_i - y_i\right)^2 + \phi_1^2 \sum_{i=1}^{n} \sigma_x^2 \quad (\text{since } \epsilon_i \sim N(0, \sigma_x^2)) \\
&= L + (1 \cdot \sigma_x^2)\phi_1^2
\end{aligned}$$

This is equivalent to L_2 regularization with $\lambda = n\sigma_x^2$.

Problem 9.4

$$\Pr(y_i|x_i, \phi) = 0.9 \cdot \text{softmax}_{y_i}\left[f(x_i, \phi)\right] + \sum_{z \in \{1, \ldots, D_0\} \setminus y_i} \frac{0.1}{D_0 - 1} \cdot \text{softmax}_z\left[f(x_i, \phi)\right]$$

The loss function is the negative logarithm of the equation above.

Problem 9.5

The standard gradient update using L_2 regularization is:

$$\phi \leftarrow \phi - \alpha \frac{\partial \tilde{\mathcal{L}}}{\partial \phi}$$
$$\Leftrightarrow \phi \leftarrow \phi - \alpha \left\{ \frac{\partial \mathcal{L}}{\partial \phi} + \frac{\lambda}{2\alpha}(2\phi) \right\}$$
$$\Leftrightarrow \phi \leftarrow (1 - \lambda)\phi - \alpha \frac{\partial \mathcal{L}}{\partial \phi}$$

Problem 9.6

$$L_0 = |\phi_0|^0 + |\phi_1|^0 = \text{number of nonzero elements}$$

$$L_{\frac{1}{2}} = \sqrt{|\phi_0|} + \sqrt{|\phi_1|}$$

$$L_1 = |\phi_0| + |\phi_1|$$