| | $\uparrow C$ | $\downarrow C$ |
|---|---|---|
| P1. Model capacity | Large | Low |
| Over / under. | Overfitting | underfitting |
| Bias Var | Low bias & high Var | high bias & low Var. |

P2. (a) $J = \frac{1}{2M} \|Xw - t\|^2 + \lambda \|w\|^2$

$= \frac{1}{2M} (Xw - t)^T (Xw - t) + \lambda w^T w.$

$\frac{\partial}{\partial w} J(w) = \frac{1}{M} X^T (Xw - t) + 2\lambda w = 0.$

$w = \left(\frac{1}{M} \|X\|^2 + 2\lambda\right)^{-1} \frac{1}{M} X^T t.$

(b). Initialize gradient vector $\nabla J$ to 0s for $w_i$.

Initialize $J \to 0$.

For each training sample. m from i to M:

    prediction = $\Sigma(w_i \cdot x_i^{(m)})$.

    Error = prediction − $t^{(m)}$

    $J += \frac{1}{2} \times \text{error}^2$

    for all i:

        grad_wi += error × $x_i^{(m)}$.

for all i:

    $J += (\frac{\lambda}{2}) \times w_i^2$

for all i:

    grad_wi += $\lambda \times w_i$

return gradicent vector $\nabla J$ for all weights $w_i$

P3: $p(w|D) \propto p(w)\, p(D|w)$.

$$\underset{w}{\operatorname{argmax}} \log p(w|D) \propto \underset{w}{\operatorname{argmax}} (\log p(w) + \log p(D|w))$$

$$p(D|w) = \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{\wedge}\left(-\frac{1}{2\sigma^2}(t^i - w^T x^i)^2\right)$$

$$\log p(D|w) = -\frac{1}{2\sigma^2}(t^i - w^T x^i)^2 + \text{const}$$

$$p(w) = \prod_{i=1}^{m} \frac{1}{2b} e^{-\frac{|w_i|}{b}}$$

$$\log p(w) = -\frac{1}{b}\sum_{i=1}^{m} |w_i| + C$$

$$= -\frac{1}{b}\|w\| + C$$

$$\log p(w) + \log p(D|w) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(t^i - w^T x^i)^2 - \frac{1}{b}\|w\| + C$$

result $\Leftrightarrow L_1$ regularization.