1. **Gradient Descent**

   (a) Derive gradient descent update rule for each θ_i with learning rate α.

$$\begin{aligned}\frac{\partial \mathcal{C}}{\partial \theta_i} &= a_i(\theta_i - r_i)\\ \theta_i^{(t+1)} &= \theta_i^{(t)} - \alpha \frac{\partial \mathcal{C}}{\partial \theta_i^{(t)}}\\ &= (1-\alpha a_i)\theta_i^{(t)} + \alpha a_i r_i\end{aligned}$$

   (b) Rewrite update rule in terms of error $e_i^{(t)} = \theta_i^{(t)} - r_i$

$$\begin{aligned}e_i^{(t+1)} &= (1-\alpha a_i)\theta_i^{(t)} + \alpha a_i r_i - r_i\\ &= (1-\alpha a_i)\theta_i^{(t)} - (1-\alpha a_i) r_i\\ &= (1-\alpha a_i)(\theta_i^{(t)} - r_i)\\ &= (1-\alpha a_i)e_i^{(t)}\end{aligned}$$

   (c) Solve recurrence to obtain explicit formula for $e_i^{(t)}$ in terms of initial error $e_i^{(0)}$

$$e_i^{(t)} = (1-\alpha a_i)^t e_i^{(0)}$$

   Error decays over time if $0 < 1 - \alpha a_i < 1$, i.e. $0 < \alpha < 1/a_i$ and similarly error grows over time if $\alpha > 1/a_i$

   (d) Write an explicit formula for the cost $\mathcal{C}(\boldsymbol{\theta}^{(t)})$ as a function of initial value $\boldsymbol{\theta}^{(0)}$

$$\mathcal{C}(\boldsymbol{\theta}^{(t)}) = \frac{1}{2}\sum_{i=1}^{N} a_i \left(e_i^{(t)}\right)^2 = \frac{1}{2}\sum_{i=1}^{N} a_i \left((1-\alpha a_i)^t e_i^{(0)}\right)^2 = \frac{1}{2}\sum_{i=1}^{N} a_i \left(1-\alpha a_i\right)^{2t}\left(\theta_i^{(0)} - r_i\right)^2$$

   As $t \to \infty$, the term whose $(1-\alpha a_i)^{2t}$ is largest starts to dominate.

2. **Dropout**

   (a) Find expressions for $\mathbb{E}\{y\}$ and $var\{y\}$ for a given $\mathbf{x}$ and $\mathbf{w}$
   By linearity of expectation and variance for independent random variables

$$\begin{aligned}\mathbb{E}\{y\} &= \sum_j w_j x_j \mathbb{E}\{m_j\} = \frac{1}{2}\sum_j w_j x_j = \frac{1}{2}\mathbf{w}^T\mathbf{x}\\ var\{y\} &= \sum_j w_j^2 x_j^2 var\{m_j\} = \frac{1}{4}\sum_j w_j^2 x_j^2 = \frac{1}{4}(\mathbf{w}^T\mathbf{x})^2\end{aligned}$$

(b) Determine $\tilde{w}_j$ as as a function of w_j such that $\mathbb{E}\{y\} = \tilde{y} = \sum_j \tilde{w}_j x_j$, where $\tilde{y}$ is a deterministic prediction

$$\frac{1}{2}\sum_j w_j x_j = \mathbb{E}\{y\} = \sum_j \tilde{w}_j x_j$$

So we have $\tilde{w}_j = \frac{1}{2}w_j$. So $\mathbf{w} = 2\tilde{\mathbf{w}}$

(c) Show cost can be rewritten to another form

$$\begin{aligned}
\mathcal{E} &= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left\{(y^{(i)} - t^{(i)})^2\right\} \\
&= \frac{1}{2N}\sum_{i=1}^{N}\left\{\mathbb{E}\left\{y^{(i)2}\right\} - 2t\mathbb{E}\left\{y^{(i)}\right\} + t^{(i)2}\right\} \\
&= \frac{1}{2N}\sum_{i=1}^{N}\left\{var\left\{y^{(i)}\right\} + \left(\mathbb{E}\left\{y^{(i)}\right\}\right)^2 - 2t\mathbb{E}\left\{y^{(i)}\right\} + t^{(i)2}\right\} \\
&= \frac{1}{2N}\sum_{i=1}^{N}(\mathbb{E}\{y\}^{(i)} - t^{(i)})^2 + \frac{1}{8N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x})^2 && (var\{y\} = \frac{1}{4}(\mathbf{w}^T\mathbf{x})^2) \\
&= \frac{1}{2N}\sum_{i=1}^{N}(\tilde{y}^{(i)} - t^{(i)})^2 + \frac{1}{8N}\sum_{i=1}^{N}(2\tilde{\mathbf{w}}^T\mathbf{x})^2 && (\mathbf{w} = 2\tilde{\mathbf{w}}) \\
&= \frac{1}{2N}\sum_{i=1}^{N}(\tilde{y}^{(i)} - t^{(i)})^2 + \frac{1}{2}(\tilde{\mathbf{w}}^T\mathbf{x})^2 \\
&= \frac{1}{2N}\sum_{i=1}^{N}(\tilde{y}^{(i)} - t^{(i)})^2 + \mathcal{R}(\tilde{w}_1, \cdots, \tilde{w}_D)
\end{aligned}$$

where

$$\mathcal{R}(\tilde{w}_1, \cdots, \tilde{w}_D) = \frac{1}{2}(\tilde{\mathbf{w}}^T\mathbf{x})^2 = \frac{1}{2}\sum_{j=1}^{D}(\tilde{w}_j x_j)^2$$