Homework 4 Solutions

1. **Gradient Descent**

(a) The gradient descent update rule has the form:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \alpha \frac{\partial \mathcal{C}(\theta)}{\partial \theta_i}$$

We derive the gradient $\frac{\partial \mathcal{C}(\theta)}{\partial \theta_i}$ as follows:

$$\begin{aligned}
\frac{\partial \mathcal{C}(\theta)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left[\frac{a_1}{2}(\theta_1 - r_1)^2 + \cdots + \frac{a_i}{2}(\theta_i - r_i)^2 + \cdots + \frac{\theta_N}{2}(\theta_N - r_N)^2\right] \\
&= \frac{\partial}{\partial \theta_i}\left[\frac{a_i}{2}(\theta_i - r_i)^2\right] \\
&= \frac{a_i}{2}\frac{\partial}{\partial \theta_i}(\theta_i - r_i)^2 \\
&= \frac{a_i}{2}\left[2(\theta_i - r_i)\frac{\partial}{\partial \theta_i}(\theta_i - r_i)\right] \\
&= a_i(\theta_i - r_i)
\end{aligned}$$

Plugging this into the formula above, we obtain the final update rule:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \alpha a_i(\theta_i^{(t)} - r_i)$$

(b) From the definition of the error, we have:

$$e_i^{(t+1)} = \theta_i^{(t+1)} - r_i$$

Using the update rule derived in Part (a), we can write $\theta_i^{(t+1)}$ in terms of $\theta_i^{(t)}$:

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

We can re-arrange the terms to find two groups that match the definition of $e_i^{(t)}$:

$$e_i^{(t+1)} = \underbrace{\theta_i^{(t)} - r_i}_{e_i^{(t)}} - \alpha a_i(\underbrace{\theta_i^{(t)} - r_i}_{e_i^{(t)}})$$

Thus, we can express $e_i^{(t+1)}$ in terms of $e_i^{(t)}$ as follows:

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

(c) We observe a pattern by expanding the first few elements of the recursion:

t = 1:

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

t = 2:

$$\begin{aligned} e_i^{(2)} &= (1 - \alpha a_i) e_i^{(1)} \\ &= (1 - \alpha a_i) \left[(1 - \alpha a_i) e_i^{(0)} \right] \\ &= (1 - \alpha a_i)^2 e_i^{(0)} \end{aligned}$$

$$\vdots$$

t = N:

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

We can prove this relationship by induction as follows:

Proof.

- **Base case:** When $N = 1$, by definition $e_i^{(1)} = (1 - \alpha a_i) e_i^{(0)} = (1 - \alpha a_i)^1 e_i^{(0)}$.
- **Inductive step:** We assume that for $N = k$, $e_i^{(k)} = (1 - \alpha a_i)^k e_i^{(0)}$. Based on this inductive assumption, we show that the relationship holds for $N = k + 1$:

$$\begin{aligned} e_i^{(k+1)} &= (1 - \alpha a_i) e_i^{(k)} \\ &= (1 - \alpha a_i) \left[(1 - \alpha a_i)^k e_i^{(0)} \right] \\ &= (1 - \alpha a_i)^{k+1} e_i^{(0)} \end{aligned}$$

□

Thus, we have solved the recurrence to obtain a closed formula for $e_i^{(t)}$ in terms of the initial error $e_i^{(0)}$:

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

Now we can reason about the effects of choosing certain values of α and a_i. First, note that both $\alpha > 0$ and $a_i > 0$. If $\alpha a_i \in (0, 1]$ then $1 - \alpha a_i \in [0, 1)$ and $\lim_{N \to \infty} (1 - \alpha a_i)^N = 0$, so the error *decays over time*.

(d) The cost function $\mathcal{C}(\theta^{(t)})$ is:

$$\mathcal{C}(\theta^{(t)}) = \frac{a_1}{2} (\theta_1^{(t)} - r_1)^2 + \cdots + \frac{a_N}{2} (\theta_N^{(t)} - r_N)^2$$

We observe that for each $i \in 1, \ldots, N$, $\theta_i^{(t)} - r_i$ is the error $e_i^{(t)}$ that we worked with in Parts (b) and (c):

$$\mathcal{C}(\theta^{(t)}) = \frac{a_1}{2} (\underbrace{\theta_1^{(t)} - r_1}_{e_1^{(t)}})^2 + \cdots + \frac{a_N}{2} (\underbrace{\theta_N^{(t)} - r_N}_{e_N^{(t)}})^2$$

In Part (c), we found the following closed form for $e_i^{(t)}$:

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

Taking the square of this closed form, we have:

$$(e_i^{(t)})^2 = ((1 - \alpha a_i)^t (\theta_i^{(0)} - r_i))^2 = (1 - \alpha a_i)^{2t} (\theta_i^{(0)} - r_i)^2$$

Thus, we can write the formula for $\mathcal{C}(\theta^{(t)})$ as a function of the initial values $\theta^{(0)}$ as follows:

$$\begin{aligned}
\mathcal{C}(\theta^{(t)}) &= \frac{a_1}{2}(1 - \alpha a_1)^{2t}(\theta_1^{(0)} - r_1)^2 + \cdots + \frac{a_N}{2}(1 - \alpha a_N)^{2t}(\theta_N^{(0)} - r_N)^2 \\
&= \sum_{i=1}^{N} \frac{a_i}{2}(1 - \alpha a_i)^{2t}(\theta_i^{(0)} - r_i)^2
\end{aligned}$$

(e) The cost function has has the form:

$$\mathcal{C}(\theta) = \frac{1}{2}(\theta - \mathbf{r})^T \mathbf{A}(\theta - \mathbf{r})$$

where $\theta \in \mathbb{R}^N$, $\mathbf{r} \in \mathbb{R}^N$, and $\mathbf{A} \in \mathbb{R}^{N \times N}$.
First, we want to derive a vectorized gradient descent update rule. Similarly to Part (a), this rule will have the form:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\partial \mathcal{C}(\theta)}{\partial \theta}$$

Thus, we need to find the gradient of $\mathcal{C}(\theta)$ with respect to θ.
We find:

$$\frac{\partial \mathcal{C}(\theta)}{\partial \theta} = \frac{1}{2} \cdot 2\mathbf{A}(\theta - \mathbf{r}) = \mathbf{A}(\theta - \mathbf{r})$$

The gradient update rule for the parameter vector θ is therefore:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \mathbf{A}(\theta^{(t)} - \mathbf{r})$$

Recurrence for the Error $\mathbf{e} = \theta - \mathbf{r}$.

$$\mathbf{e}^{(t+1)} = \theta^{(t+1)} - \mathbf{r}$$

As we found in the previous part,

$$\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} - \alpha \mathbf{A}(\theta - \mathbf{r}) \\
&= \theta^{(t)} - \alpha \mathbf{A}\theta^{(t)} + \alpha \mathbf{A}\mathbf{r}
\end{aligned}$$

Pugging this into the equation for $\mathbf{e}^{(t+1)}$ yields:

$$\begin{aligned}
\mathbf{e}^{(t+1)} &= \theta^{(t)} - \alpha \mathbf{A}\theta^{(t)} + \alpha \mathbf{A}\mathbf{r} - \mathbf{r} \\
&= \underbrace{\theta^{(t)} - \mathbf{r}}_{\mathbf{e}^{(t)}} - \alpha \mathbf{A}(\underbrace{\theta^{(t)} - \mathbf{r}}_{\mathbf{e}^{(t)}}) \\
&= \mathbf{e}^{(t)} - \alpha \mathbf{A}\mathbf{e}^{(t)} \\
&= (\mathbf{I} - \alpha \mathbf{A})\mathbf{e}^{(t)}
\end{aligned}$$

The last line expresses a recurrence of the form $\mathbf{e}^{(t+1)} = \mathbf{B}\mathbf{e}^{(t)}$, where $\mathbf{B} = \mathbf{I} - \alpha\mathbf{A}$. Note that $\mathbf{B}$ is symmetric because: 1) if we multiply a symmetric matrix by a scalar, the result is a symmetric matrix, so $-\alpha\mathbf{A}$ is symmetric; 2) the sum of two symmetric matrices is symmetric, and both $\mathbf{I}$ and $\mathbf{I} - \alpha\mathbf{A}$ are symmetric, so $\mathbf{I} - \alpha\mathbf{A}$ is symmetric.

Now we wish to find an explicit form for $\mathbf{e}^{(t)}$ in terms of $\theta^{(0)}$. Note that:

$$\mathbf{e}^{(1)} = \mathbf{B}\mathbf{e}^{(0)}$$
$$\mathbf{e}^{(2)} = \mathbf{B}\mathbf{e}^{(1)} = \mathbf{B}(\mathbf{B}\mathbf{e}^{(0)}) = (\mathbf{B}\mathbf{B})\mathbf{e}^{(0)} = \mathbf{B}^2\mathbf{e}^{(0)}$$
$$\vdots$$
$$\mathbf{e}^{(t)} = \mathbf{B}^t\mathbf{e}^{(0)}$$

We can exploit the structure of the matrix $\mathbf{B}$ in order to be able to find powers $\mathbf{B}^t$ easily, rather than actually performing t matrix multiplications. Since $\mathbf{A}$ is a symmetric positive definite matrix, we can express it using its *eigendecomposition*, as $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q}$ is a matrix whose columns are the *eigenvectors* of $\mathbf{A}$, and $\boldsymbol{\Lambda}$ is a *diagonal* matrix containing the *eigenvalues* of $\mathbf{A}$ (that is, $\boldsymbol{\Lambda} = diag(\lambda)$, where λ is the vector of eigenvalues of $\mathbf{A}$).

Now, if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ, then:

$$\begin{aligned}\mathbf{B}\mathbf{x} &= (\mathbf{I} - \alpha\mathbf{A})\mathbf{x} \\ &= \mathbf{I}\mathbf{x} - \alpha\mathbf{A}\mathbf{x} \\ &= \mathbf{x} - \alpha(\mathbf{A}\mathbf{x}) \\ &= \mathbf{x} - \alpha\lambda\mathbf{x} \\ &= (1 - \alpha\lambda)\mathbf{x}\end{aligned}$$

So $\mathbf{x}$ is also an eigenvector of $\mathbf{B}$ with eigenvalue $1 - \alpha\lambda$. We see that the eigenvectors of $\mathbf{B}$ are the same as those of $\mathbf{A}$, and for each eigenvalue λ of $\mathbf{A}$, we have a corresponding eigenvalue $1 - \alpha\lambda$ of $\mathbf{B}$. Thus, the eigendecomposition of $\mathbf{B} = \mathbf{Q}(\mathbf{I} - \alpha\boldsymbol{\Lambda})\mathbf{Q}^T$.

We can find $\mathbf{e}^{(t)}$ more easily by expressing it in terms of the eigendecomposition of $\mathbf{B}$:

$$\begin{aligned}\mathbf{e}^{(t)} &= \mathbf{B}^t\mathbf{e}^{(0)} \\ &= \mathbf{Q}(\mathbf{I} - \alpha\boldsymbol{\Lambda})^t\mathbf{Q}^T\mathbf{e}^{(0)}\end{aligned}$$

Putting this all together in the context of the cost function, we have:

$$\begin{aligned}
\mathcal{C}(\theta^{(t)}) &= \frac{1}{2}(\theta^{(t)} - \mathbf{r})^T \mathbf{A}(\theta^{(t)} - \mathbf{r}) \\
&= \frac{1}{2}(\mathbf{e}^{(t)})^T \mathbf{A}(\mathbf{e}^{(t)}) \\
&= \frac{1}{2}\left[\mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T \mathbf{e}^{(0)}\right]^T \mathbf{A} \left[\mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T \mathbf{e}^{(0)}\right] \\
&= \frac{1}{2}\left[(\mathbf{e}^{(0)})^T \mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T\right] \mathbf{A} \left[\mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T \mathbf{e}^{(0)}\right] \\
&= \frac{1}{2}(\mathbf{e}^{(0)})^T \mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T \left[\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\right] \mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T \mathbf{e}^{(0)} \\
&= \frac{1}{2}(\mathbf{e}^{(0)})^T \mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t (\mathbf{Q}^T\mathbf{Q})\mathbf{\Lambda}(\mathbf{Q}^T\mathbf{Q})(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T \mathbf{e}^{(0)} \\
&= \frac{1}{2}(\mathbf{e}^{(0)})^T \mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{\Lambda}(\mathbf{I} - \alpha\mathbf{\Lambda})^t \mathbf{Q}^T \mathbf{e}^{(0)}
\end{aligned}$$

Since $(\mathbf{I} - \alpha\mathbf{\Lambda})^t$ and $\mathbf{\Lambda}$ are both diagonal matrices, we can commute their order, and combine the two instances of $(\mathbf{I} - \alpha\mathbf{\Lambda})^t$ to yield $(\mathbf{I} - \alpha\mathbf{\Lambda})^{2t}$:

$$\mathcal{C}(\theta^{(t)}) = \frac{1}{2}(\mathbf{e}^{(0)})^T \mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^{2t} \mathbf{\Lambda}\mathbf{Q}^T \mathbf{e}^{(0)}$$

Replacing $\mathbf{e}^{(0)}$ with $\theta^{(0)} - \mathbf{r}$, we obtain:

$$\mathcal{C}(\theta^{(t)}) = \frac{1}{2}(\theta^{(0)} - \mathbf{r})^T \mathbf{Q}(\mathbf{I} - \alpha\mathbf{\Lambda})^{2t} \mathbf{\Lambda}\mathbf{Q}^T (\theta^{(0)} - \mathbf{r})$$

2. **Dropout**

(a) Find expressions for $\mathbb{E}[y]$ and $\text{Var}[y]$ for a given data point.

We can determine $\mathbb{E}[y]$ and $\text{Var}[y]$ using the properties of expectation and variance.

$$\begin{aligned}
\mathbb{E}[y] &= \mathbb{E}\left[\sum_j m_j w_j x_j\right] \\
&= \sum_j w_j x_j \mathbb{E}[m_j] && \text{by linearity of expectation} \\
&= \frac{1}{2}\sum_j w_j x_j && \text{by the expectation formula for a Bernoulli r.v.} \\
\text{Var}[y] &= \text{Var}\left[\sum_j m_j w_j x_j\right] \\
&= \sum_j \text{Var}\left[m_j w_j x_j\right] && \text{by independence} \\
&= \sum_j w_j^2 x_j^2 \text{Var}[m_j] && \text{by the scalar multiplication rule for variance} \\
&= \frac{1}{4}\sum_j w_j^2 x_j^2 && \text{by the variance formula for a Bernoulli r.v.}
\end{aligned}$$

(b) Determine $\tilde{w}_j$ as a function of w_j such that

$$\mathbb{E}[y] = \tilde{y} = \sum_j \tilde{w}_j x_j$$

Based on the expectation derived in Part (a), we have:

$$\begin{aligned}\mathbb{E}[y] &= \frac{1}{2}\sum_j w_j x_j^{(i)} \\ &= \sum_j (\frac{1}{2}w_j)x_j^{(i)}\end{aligned}$$

Thus,

$$\tilde{w}_j = \frac{1}{2}w_j$$

(c) Using the model from the previous section, show that the cost $\mathcal{E}$ can be written as:

$$\mathcal{E} = \frac{1}{2N}\sum_{i=1}^{N}(\tilde{y}^{(i)} - t^{(i)})^2 + \mathcal{R}(\tilde{w}_1, \ldots, \tilde{w}_D)$$

Equation 1 in the homework states:

$$\mathcal{E} = \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}[(y^{(i)} - t^{(i)})^2]$$

Using the fact that the expectation is a linear operation, we can expand it as follows:

$$\mathbb{E}[(y^{(i)} - t^{(i)})^2] = \mathbb{E}[(y^{(i)})^2] - 2\mathbb{E}[y^{(i)}t^{(i)}] + \mathbb{E}[(t^{(i)})^2]$$

We can express $\mathbb{E}[(y^{(i)})^2]$ in terms of the variance as follows:

$$\mathbb{E}[(y^{(i)})^2] = \mathrm{Var}[y^{(i)}] + \mathbb{E}[y^{(i)}]^2$$

Since $\tilde{y}^{(i)} = \mathbb{E}[y^{(i)}]$, we have:

$$\mathbb{E}[(y^{(i)})^2] = \mathrm{Var}[y^{(i)}] + (\tilde{y}^{(i)})^2$$

Since $t^{(i)}$ is not a function of the $m_j^{(i)}$'s, $t^{(i)}$ is treated as a constant in the expectation $\mathbb{E}[y^{(i)}t^{(i)}]$, so we have:

$$\begin{aligned}\mathbb{E}[y^{(i)}t^{(i)}] &= t^{(i)}\mathbb{E}[y^{(i)}] \\ &= t^{(i)}\tilde{y}^{(i)}\end{aligned}$$

Similarly, since $t^{(i)}$ is not a function of the $m_j^{(i)}$'s, the expectation of $(t^{(i)})^2$ with respect to the $m_j^{(i)}$'s is $(t^{(i)})^2$:

$$\mathbb{E}[(t^{(i)})^2] = (t^{(i)})^2$$

Putting these terms together, we have:

$$\begin{aligned}\mathbb{E}[(y^{(i)} - t^{(i)})^2] &= \mathrm{Var}[y^{(i)}] + (\tilde{y}^{(i)})^2 - 2t^{(i)}(\tilde{y}^{(i)})^2 + (t^{(i)})^2 \\ &= (\tilde{y}^{(i)} - t^{(i)})^2 + \mathrm{Var}[y^{(i)}]\end{aligned}$$

Plugging this derivation of $\mathbb{E}[(y^{(i)} - t^{(i)})^2]$ into the original expression for $\mathcal{E}$ yields:

$$\begin{aligned}\mathcal{E} &= \frac{1}{2N}\sum_{i=1}^{N}\Big((\tilde{y}^{(i)} - t^{(i)})^2 + \mathrm{Var}[y^{(i)}]\Big) \\ &= \frac{1}{2N}\sum_{i=1}^{N}(\tilde{y}^{(i)} - t^{(i)})^2 + \frac{1}{2N}\sum_{i=1}^{N}\mathrm{Var}[y^{(i)}]\end{aligned}$$

Finally, we can substitute the expression for the variance that we derived in Part (a) to obtain a regularization term that does not involve any expectations:

$$\begin{aligned}\mathcal{E} &= \frac{1}{2N}\sum_{i=1}^{N}(\tilde{y}^{(i)} - t^{(i)})^2 + \frac{1}{2N}\sum_{i=1}^{N}\frac{1}{4}\sum_{j} w_j^2 (x_j^{(i)})^2 \\ &= \frac{1}{2N}\sum_{i=1}^{N}(\tilde{y}^{(i)} - t^{(i)})^2 + \frac{1}{8N}\sum_{i=1}^{N}\sum_{j} w_j^2 (x_j^{(i)})^2\end{aligned}$$