Problem Set #1

Joe Michail
DATA SCI 423 – Machine Learning
Northwestern University

April 14, 2020

Question 3.1

1. $g(w) = w \log w + (1-w)\log(1-w), w \in (0,1)$

$$\begin{aligned}
\frac{dg}{dw} &= \log w \ \frac{dw}{dw} + w\frac{d \log w}{dw} + \log(1-w)\frac{d\ (1-w)}{dw} + (1-w)\frac{d\ \log(1-w)}{dw} \\
&= \log w + \frac{w}{w} + \log(1-w) \cdot -1 + \frac{(1-w)}{(1-w)} \cdot -1 \\
&= \log w - \log(1-w)
\end{aligned}$$

Stationary points occur when $dg/dw = 0$:

$$\begin{aligned}
0 &= \log \frac{w}{1-w} \\
1 &= \frac{w}{1-w}
\end{aligned}$$

$$\boxed{w = 1/2}$$

The plot (at end of the homework) shows that this point is a minimum.

2. $g(w) = \log(1 + e^w)$

$$\begin{aligned}
\frac{dg}{dw} &= \frac{\log(1+e^w)}{d(1+e^w)}\frac{d(1+e^w)}{dw} \\
&= \frac{e^w}{(1+e^w)} \\
0 &= e^w
\end{aligned}$$

$$\boxed{w = -\infty}$$

The plot shows that the derivative function tending toward $w = -\infty$ is a minimum.

3. $g(w) = w \tanh w$

$$\begin{aligned}
\frac{dg}{dw} &= \tanh w \, \frac{dw}{dw} + w \, \frac{d \tanh w}{dw} \\
&= \tanh w + w \, \operatorname{sech}^2 w \\
0 &= \sinh w \cosh w + w \\
&= \frac{1}{2} \sinh 2w + w
\end{aligned}$$

$$\boxed{w = 0}$$

The plot shows that this stationary point is a minimum.

4. $g(w) = \frac{1}{2}\mathbf{w}^T\mathbf{C}\mathbf{w} + \mathbf{b}^T\mathbf{w}$

$$\begin{aligned}
\nabla_w g(w) &= \frac{1}{2}\nabla(\mathbf{w}^T\mathbf{C}\mathbf{w}) + \nabla(\mathbf{b}^T\mathbf{w}) \\
&= \frac{1}{2}(\mathbf{C} + \mathbf{C}^T)\mathbf{w} + \mathbf{b}, \qquad \mathbf{C}^T = \mathbf{C} \\
\mathbf{0} &= \mathbf{C}\mathbf{w} + \mathbf{b} \\
-\mathbf{b} &= \mathbf{C}\mathbf{w}
\end{aligned}$$

The problem states that:

$$\mathbf{C} = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

$$\boxed{(w_1 \quad w_2)^T = (-2/5 \quad -1/5)^T}$$

The plot shows this stationary point is a minimum.

Question 3.3

Start with Rayleigh's quotient:

$$g(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{C}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}.$$

Taking the gradient of the numerator and denominators separately:

$$\begin{aligned}
\nabla(\mathbf{w}^T\mathbf{C}\mathbf{w}) &= (\mathbf{C} + \mathbf{C}^T)\mathbf{w}, \\
\nabla(\mathbf{w}^T\mathbf{w}) &= 2\mathbf{w}.
\end{aligned}$$

Now taking the gradient of g:

$$\begin{aligned}
\nabla g(\mathbf{w}) &= \frac{\nabla(\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{w}^T\mathbf{w} - \nabla(\mathbf{w}^T\mathbf{w})\mathbf{w}^T\mathbf{C}\mathbf{w}}{(\mathbf{w}^T\mathbf{w})^2} \\
&= \frac{(\mathbf{C}+\mathbf{C}^T)\mathbf{w}\mathbf{w}^T\mathbf{w} - 2\mathbf{w}\mathbf{w}^T\mathbf{C}\mathbf{w}}{||\mathbf{w}||_2^4}.
\end{aligned}$$

The stationary points occur when $\nabla g = \mathbf{0}$:

$$\mathbf{0} = (\mathbf{C}+\mathbf{C}^T)\mathbf{w}(\mathbf{w}^T\mathbf{w}) - 2\mathbf{w}(\mathbf{w}^T\mathbf{C}\mathbf{w}).$$

Since $\mathbf{w}^T\mathbf{w} \in \mathbb{R}$, it can freely be divided:

$$\begin{aligned}
\mathbf{0} &= (\mathbf{C}+\mathbf{C}^T)\mathbf{w} - 2\mathbf{w}\frac{\mathbf{w}^T\mathbf{C}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \\
&= (\mathbf{C}+\mathbf{C}^T)\mathbf{w} - 2g(\mathbf{w})\mathbf{w} \\
&= (\mathbf{C}+\mathbf{C}^T - 2g(\mathbf{w})\mathbf{I})\mathbf{w} \qquad (1)
\end{aligned}$$

There, the last equation is multiplied by the identity matrix $\mathbf{I}$ to make the sum inside of the parenthesis a matrix. This is the classic eigenvalue problem which can more easily been shown if $\mathbf{C} = \mathbf{C}^T$:

$$\mathbf{0} = (2\mathbf{C} - 2g(\mathbf{w})\mathbf{I})\mathbf{w} \implies \det(\mathbf{C} - 2g(\mathbf{w})\mathbf{I}) = 0.$$

Consider $\mathbf{C} \in \mathbb{R}^{N\times N}$. This implies that $|\mathbf{C}+\mathbf{C}^T - 2g(\mathbf{w})\mathbf{I}| = 0$ can be solved to determine N eigenvalues, each of which correspond to one eigenvectors (N total eigenvectors).
Therefore, the stationary points of Rayleigh's quotient correspond to the N eigenvectors of the matrix $\mathbf{C}$.

Question 3.5

Here we start with the function:

$$g(w) = \frac{w^4 + w^2 + 10w}{50}$$

The derivative of this is trivial:

$$\frac{dg}{dw} = \frac{4w^3 + 2w + 10}{50}$$

The cost function plots for this problem are at the end of this document. With this combination of step length and initial position, $\alpha = 1$ converges the quickest to the minimum $w \sim -1.234, g(w) \sim -0.17$.

Question 3.6

See end of document for plot. Please note that the fixed-step line stops at $k = 4$ where it encounters the minimum which is a non-differentiable point.

Question 3.8

The cost function we're trying to minimize is:

$$g(\mathbf{w}) = \mathbf{w}^T\mathbf{w}$$

and has a gradient of the form:

$$\nabla_{\mathbf{w}} g(\mathbf{w}) = 2\mathbf{w}.$$

The cost function history plots are below and a step length of $\alpha = 0.1$ converges quickest to the minimum located at $\mathbf{w} = \mathbf{0}$.