Exercise Sheet #6

Solution

December 3, 2018

1 Newton Optimization

We consider the real-valued function $L : \mathbb{R}^d \to \mathbb{R}$ with real-vector-valued inputs $f \in \mathbb{R}^d$. We denote the gradient as $\nabla L(f) = \left[\frac{\partial L(f)}{\partial f_i}\right]_{i=1,\dots,d}$ and the Hessian as $B(f) = \left[\frac{\partial^2 L(f)}{\partial f_i \partial f_j}\right]_{i,j}$.

(a)

Taylor's expansion of $L(f)$ around f_0 gives us:

$$L(f_0 + \delta) = \underbrace{L(f_0) + \nabla^T L(f_0)\delta + \frac{1}{2}\delta^T B(f_0)\delta}_{\tilde{L}(f_0+\delta)} + \mathcal{O}(\delta^3) \; . \tag{1}$$

The minimum of $\tilde{L}(f_0 + \delta)$ can be found by computing the gradient with respect to the step δ, and finding the step where it's zero:

$$\nabla_\delta \tilde{L}(f_0 + \delta) = \nabla L(f_0) + \frac{1}{2}\left(B(f_0) + B(f_0)^T\right)\delta \overset{B=B^T}{=} \nabla L(f_0) + B(f_0)\delta \overset{!}{=} 0 \tag{2}$$

$$\iff \delta = -B^{-1}(f_0)\nabla L(f_0) \tag{3}$$

Note, that we can verify that the gradient of $\delta^T B \delta$ is $B\delta$, if $B^T = B$, using the sum notation:

$$\frac{\partial}{\partial \delta_k} \sum_{i,j} \delta_i B_{ij} \delta_j = \sum_j B_{kj}\delta_j + \sum_i \delta_i B_{ik} \overset{B^T=B}{=} \sum_i B_{ki}\delta_i + \sum_i B_{ki}\delta_i = 2\sum_i B_{ki}\delta_i$$

Therefore the minimum of the approximation $\tilde{L}(f_0 + \delta)$ is at

$$f_1 := f_0 + \delta = f_0 - B^{-1}(f_0)\nabla L(f_0) \tag{4}$$

(b)

Taylor's expansion of $L(\phi w)$ around w_0 gives us:

$$L(\phi^T(w_0+\epsilon)) = \underbrace{L(\phi^T w_0) + \nabla^T L(\phi^T w_0)\phi^T\epsilon + \frac{1}{2}(\phi^T\epsilon)^T B(\phi^T w_0)\phi^T\epsilon}_{\tilde{L}(w_0+\epsilon)} + \mathcal{O}(\epsilon^3)\ . \tag{5}$$

Again, we find the minimum, by finding the point where the gradient is zero:

$$\nabla_\epsilon \tilde{L}(w_0+\epsilon) \overset{B=B^T}{=} \phi\nabla L(\phi^T w_0) + \phi B(\phi^T w_0)\phi^T\epsilon \overset{!}{=} 0 \tag{6}$$

$$\iff \epsilon = -(\phi B(\phi^T w_0)\phi^T)^{-1}\phi\nabla L(\phi^T w_0) \tag{7}$$

therefore, the Newton updates in w is

$$w_1 = w_0 + \epsilon = w_0 - (\phi B(\phi^T w_0)\phi^T)^{-1}\phi\nabla L(\phi^T w_0) \tag{8}$$

2 Logistic Link Functions

We want to show two identities for the sigmoid function $\sigma(f) = \frac{1}{1+\exp(-f)}$. Note that in the following, y denotes a binary variable that can take on exactly two values, such that $y \in \{-1, 1\}$.

Throughout, we will use two important properties of the sigmoid function:

$$\sigma(f) = 1 - \sigma(-f) \tag{9}$$

$$\sigma'(f) = \sigma(f)(1-\sigma(f)) \tag{10}$$

which we quickly prove here:

$$1 - \sigma(-f) = 1 - \frac{1}{1+e^f} = \frac{1+e^f-1}{1+e^f} = \frac{e^f}{1+e^f} = \frac{1}{1+e^{-f}}$$

and

$$\sigma'(f) = \frac{-e^{-f}}{(1-e^{-f})^2} = \sigma(f)\left(\frac{1}{1-e^f}\right) = \sigma(f)\sigma(-f) = \sigma(f)(1-\sigma(f))$$

using (9) in the last equation.

First Identity

$$\frac{\partial \log\sigma(y\cdot f)}{\partial f} \overset{(10)}{=} \frac{1}{\sigma(y\cdot f)} \cdot y \cdot \sigma(y\cdot f)\,(1-\sigma(y\cdot f)) \tag{11}$$

Now we do a simple case-by-case analysis.

y = 1

$$\frac{\partial \log \sigma(1 \cdot f)}{\partial f} \overset{(11)}{=} \frac{1}{\sigma(1 \cdot f)} \cdot 1 \cdot \sigma(1 \cdot f)\,(1 - \sigma(1 \cdot f)) \tag{12}$$

$$= (1 - \sigma(f)) = \frac{y+1}{2} - \sigma(f) \tag{13}$$

y = -1

$$\frac{\partial \log \sigma(-1 \cdot f)}{\partial f} \overset{(11)}{=} \frac{1}{\sigma(-1 \cdot f)} \cdot -1 \cdot \sigma(-1 \cdot f)\,(1 - \sigma(-1 \cdot f)) \tag{14}$$

$$= -1 + \sigma(-f) \tag{15}$$

$$\overset{(9)}{=} -1 + (1 - \sigma(f)) \tag{16}$$

$$= -\sigma(f) = \frac{y+1}{2} - \sigma(f) \tag{17}$$

Second Identity

From the first identity we now that

$$\frac{\partial \log \sigma(y \cdot f)}{\partial f} = \frac{y+1}{2} - \sigma(f) \tag{18}$$

computing the derivative of this with respect to f, give us

$$\frac{\partial^2 \log \sigma(y \cdot f)}{(\partial f)^2} \overset{(10)}{=} -\sigma(f)(1 - \sigma(f)) \tag{19}$$