Continuous Optimization: Assignment 6

Due on June 4, 2024

Honglu Ma **Hiroyasu Akada** **Mathivathana Ayyappan**

Exercise 1

The strong Wolfe condition states that for some $\eta \in (\gamma, 1), \gamma \in (0, 1)$, the following holds:

$$\left|\langle \nabla f(x^{(k)} + \tau_k d^{(k)}), d^{(k)} \rangle\right| \leq \eta \left|\langle \nabla f(x^{(k)}), d^{(k)} \rangle\right|$$

We know the iterative update step for $x^{(k+1)}$ is defined as: $x^{(k+1)} = x^{(k)} + \tau_k d^{(k)}$. The strong curvature condition can be rewritten as such:

$$\left|\langle \nabla f(x^{(k+1)}), d^{(k)} \rangle\right| \leq \eta \left|\langle \nabla f(x^{(k)}), d^{(k)} \rangle\right|$$

By the definition of descent direction, $\langle \nabla f(x^{(k)}), d^{(k)} \rangle < 0$ and $\eta > 0$, we get

$$\begin{aligned}
\langle \nabla f(x^{(k+1)}), d^{(k)} \rangle &\geq \eta \langle \nabla f(x^{(k)}), d^{(k)} \rangle \\
\langle \nabla f(x^{(k+1)}), d^{(k)} \rangle - \langle \nabla f(x^{(k)}), d^{(k)} \rangle &\geq \eta \langle \nabla f(x^{(k)}), d^{(k)} \rangle - \langle \nabla f(x^{(k)}), d^{(k)} \rangle \\
\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), d^{(k)} \rangle &\geq (\eta - 1) \langle \nabla f(x^{(k)}), d^{(k)} \rangle > 0
\end{aligned}$$

We know $\tau_k > 0$

$$\begin{aligned}
\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), \tau_k d^{(k)} \rangle > 0 \\
\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle > 0 \\
\langle y^{(k)}, s^{(k)} \rangle > 0
\end{aligned}$$

Exercise 2

The secant equation is given by $B_{k+1} s^{(k)} = y^{(k)}$ which is a system of n linear equations (assume the dimension is n). The choice of B_{k+1} is constrained by these n equations which results in a degree of freedom of n. On the other hand, the curvature condition:

$$\begin{aligned}
\langle s^{(k)}, B_{k+1} s^{(k)} \rangle &= \langle s^{(k)}, y^{(k)} \rangle \\
\langle s^{(k)}, B_{k+1} s^{(k)} \rangle - \langle s^{(k)}, y^{(k)} \rangle &= 0 \\
\langle s^{(k)}, B_{k+1} s^{(k)} - y^{(k)} \rangle &= 0
\end{aligned}$$

It can be satisfied not only by setting $B_{k+1} s^{(k)} - y^{(k)} = 0$ which is the same as the secant equation, but also by setting $B_{k+1} s^{(k)} - y^{(k)}$ to be orthogonal to $s^{(k)}$. This gives more degree of freedom of choosing B_{k+1}.

Exercise 3

Theory 5.10 states that if f is strictly convex and differentiable, then the following holds:

$$\langle x_1 - x_0, \nabla f(x_1) - \nabla f(x_0) \rangle > 0$$

which can be interprerated as the curvature condition.

Exercise 4

(a)

An useful identity:

$$1 - h_w(x) = \frac{e^{-\langle w,x\rangle}}{1 + e^{-\langle w,x\rangle}} = h_w(x) \cdot e^{-\langle w,x\rangle}$$

We first calculate $\frac{\partial h_w(x)}{\partial w}$:

$$\begin{aligned}
\frac{\partial h_w(x)}{\partial w} &= (1 + e^{-\langle w,x\rangle})^{-2} \cdot e^{-\langle w,x\rangle} \cdot x \\
&= h_w(x)^2 \cdot e^{-\langle w,x\rangle} \cdot x \\
&= h_w(x) \cdot (1 - h_w(x)) \cdot x
\end{aligned}$$

Using chain rule, we can calculate $\frac{\partial f(w)}{\partial w}$:

$$\begin{aligned}
\frac{\partial f(w)}{\partial w} &= -\frac{1}{m} \sum_{i=1}^{m} \left(y_i \cdot \frac{\partial \log(h_w(x_i))}{\partial w} + (1 - y_i) \cdot \frac{\partial \log(1 - h_w(x_i))}{\partial w} \right) \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left(y_i \cdot \frac{1}{h_w(x_i)} \cdot \frac{\partial h_w(x_i)}{\partial w} - (1 - y_i) \cdot \frac{1}{1 - h_w(x_i)} \cdot \frac{\partial h_w(x_i)}{\partial w} \right) \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left(y_i \cdot \frac{1}{h_w(x_i)} \cdot h_w(x_i) \cdot (1 - h_w(x_i)) \cdot x_i - (1 - y_i) \cdot \frac{1}{1 - h_w(x_i)} \cdot h_w(x_i) \cdot (1 - h_w(x_i)) \cdot x_i \right) \\
&= -\frac{1}{m} \sum_{i=1}^{m} (y_i \cdot (1 - h_w(x_i)) \cdot x_i - (1 - y_i) \cdot h_w(x_i) \cdot x_i) \\
&= -\frac{1}{m} \sum_{i=1}^{m} ((y_i \cdot (1 - h_w(x_i)) - (1 - y_i) \cdot h_w(x_i)) \cdot x_i) \\
&= -\frac{1}{m} \sum_{i=1}^{m} ((y_i - y_i \cdot h_w(x_i) - h_w(x_i) + y_i \cdot h_w(x_i)) \cdot x_i) \\
&= -\frac{1}{m} \sum_{i=1}^{m} ((y_i - h_w(x_i)) \cdot x_i)
\end{aligned}$$

Now we calculate the Hessian $\frac{\partial \frac{\partial f(w)}{\partial w}}{\partial w}$

$$\frac{\partial \frac{\partial f(w)}{\partial w}}{\partial w} = \frac{1}{m} \sum_{i=1}^{m} \left(h_w(x_i) \cdot (1 - h_w(x_i)) \cdot x_i \cdot x_i^T \right)$$

(b)

Observe that $0 < h_w(x_i) < 1$ for all $i = 1 \cdots m$ which means $h_w(x_i) > 0$ and $1 - h_w(x_i) > 0$. Furthermore, the term $x_i \cdot x_i^\top$ results in the a matrix which each entry is the square of the entries in x_i. This shows that the Hessian is positive semi-definite thus function is convex.