AMS230 – Homework #4

Zayd Hammoudeh

May 20, 2018

Name: Zayd Hammoudeh
Course Name: AMS230
Assignment Name: Homework #4
Due Date: May 30, 2018
Student Discussions: I discussed the problems with the following students. All write-ups were prepared separately and independently.

- Ben Sherman
- Bernardo Torres

Exercise #1
Exercise 6.1 in Nocedal and Wright.

Hint: If a function $f(x)$ is strongly convex, there exists $\sigma > 0$ such that $v^{\mathrm{T}}\nabla^2 f(x)v \geq \sigma\|v\|^2$.

(a) Show that if f is strongly convex, then

$$s_k^{\mathrm{T}} y_k > 0$$

holds for any vectors x_k and x_{k+1}.

Proof. A function, f, is *strongly convex* if it holds that:

$$\Big(\nabla f(x) - \nabla f(y)\Big)^{\mathrm{T}} (x - y) \geq m\|x - y\|^2$$

for some $m > 0$. Define $s_k = x_{k+1} - x_k$ as the change in position between consecutive iterates. Define $y_k = \nabla f_{k+1} - \nabla f_k$. Given a strongly convex $f(x)$, we get:

$$y_k^{\mathrm{T}} s_k \geq m\|s_k\|^2.$$

If $x_k \neq x_{k+1}$, then by definition $\|s_k\| > 0$. Also, $y_k^{\mathrm{T}} s_k = s_k^{\mathrm{T}} y_k$. Therefore,

$$s_k^{\mathrm{T}} y_k > 0.$$

□

(b) Give an example of a function of one variable satisfying $g(0) = -1$ and $g(1) = -\frac{1}{4}$ and show that (6.7) does not hold in this case.

Consider the univariate function $f(x) = \frac{1}{x+1}$ whose gradient $\nabla f(x) = g(x) = -\frac{1}{(x+1)^2}$. This function is obviously not convex, but its gradient satisfies the property that $g(0) = -1$ and $g(1) = -0.25$. Consider $x_1 = -2$ and $x_2 = 0$. $g(x_1) = g(x_2) = -1$ making

$$s^{\mathrm{T}} y = 0.$$

Clearly, this does not satisfy the above condition that $s^{\mathrm{T}} y > 0$.

Exercise #2
Exercise 6.4 in Nocedal and Wright.

Show that

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^{\mathrm{T}}}{(y_k - B_k s_k)^{\mathrm{T}} s_k}$$

is the inverse of

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^{\mathrm{T}}}{(s_k - H_k y_k)^{\mathrm{T}} y_k}.$$

Note: By the **Sherman-Morrison-Woodbury formula**, if the square nonsingular matrix A undergoes a rank-one update to become

$$\bar{A} = A + ab^{\mathrm{T}}$$

where $a, b \in \mathbb{R}^n$, then if $\bar{A}$ is nonsingular, we have:

$$\bar{A}^{-1} = A^{-1} - \frac{A^{-1}ab^{\mathrm{T}}A^{-1}}{1 + b^{\mathrm{T}}A^{-1}a}.$$

By definition $H_k = B_k^{-1}$. Define $a = b = y_k - B_k s_k$ and $\rho = \frac{1}{(y_k - B_k s_k)^{\mathrm{T}} s_k}$. Substituting

$$B_{k+1} = B_k + \rho ab^{\mathrm{T}}.$$

This can be substituted into the Sherman-Morrison-Woodbury formula as:

$$B_{k+1}^{-1} = B_k^{-1} - \frac{B_k^{-1}\rho(y_k - B_k s_k)(y_k - B_k s_k)^{\mathrm{T}} B_k^{-1}}{1 + (y_k - B_k s_k)^{\mathrm{T}} B_k^{-1}\rho(y_k - B_k s_k)}$$

B_k is a symmetric, nonsingular matrix. As such, B_k^{-1} is also a symmetric, nonsingular matrix. Therefore,

$$\begin{aligned}
B_{k+1}^{-1} &= B_k^{-1} - \frac{(B_k^{-1}y_k - s_k)(B_k^{-1}y_k - s_k)^{\mathrm{T}}}{(y_k - B_k s_k)^{\mathrm{T}} s_k + (y_k - B_k s_k)^{\mathrm{T}}(B_k^{-1}y_k - s_k)} \\
&= B_k^{-1} - \frac{(s_k - B_k^{-1}y_k)(s_k - B_k^{-1}y_k)^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k - (B_k s_k)^{\mathrm{T}} s_k + y_k^{\mathrm{T}} B_k^{-1} y_k - y_k^{\mathrm{T}} s_k - s_k^{\mathrm{T}} y_k + (B_k s_k)^{\mathrm{T}} s_k} \\
&= B_k^{-1} - \frac{(s_k - B_k^{-1}y_k)(s_k - B_k^{-1}y_k)^{\mathrm{T}}}{y_k^{\mathrm{T}} B_k^{-1} y_k - s_k^{\mathrm{T}} y_k} \\
&= B_k^{-1} - \frac{(s_k - B_k^{-1}y_k)(s_k - B_k^{-1}y_k)^{\mathrm{T}}}{(B_k^{-1}y_k - s_k)^{\mathrm{T}} y_k} \\
&= B_k^{-1} + \frac{(s_k - B_k^{-1}y_k)(s_k - B_k^{-1}y_k)^{\mathrm{T}}}{(s_k - B_k^{-1}y_k)^{\mathrm{T}} y_k}
\end{aligned}$$

By definition $B_k^{-1} = H_k$. Therefore

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^{\mathrm{T}}}{(s_k - H_k y_k)^{\mathrm{T}} y_k}$$

Exercise #3

Code Algorithm 7.5, and test it on the extended Rosenbrock function

$$f(x) = \sum_{i=1}^{n/2} \left[\alpha(x_{2i} - x_{2i-1}^2)^2 + (1 - x_{2i-1})^2\right].$$

where α is a parameter that you can vary (for example, 1 or 100). The solution is $x^* = (1, 1, ..., 1)^{\mathrm{T}}$, $f^* = 0$. Choose the starting point as $(-1, -1, ..., -1)^{\mathrm{T}}$. Observe the behavior of your program for various values of the memory parameter m.

For even valued n, the gradient of f is:

$$\nabla f(x) = \begin{cases} -4\alpha x_j(x_{j+1} - x_j^2) - 2x_j(1 - x_j) & j \text{ is odd} \\ 2\alpha(x_j - x_{j-1}^2) & j \text{ is even} \end{cases}$$

where $j \in \{1, \dots, n\}$. Since LBFGS relies on line search, $\phi(\alpha) = f(x_k + \alpha_k p_k)$. In addition, by the chain rule,

$$\begin{aligned} \phi'(\alpha) &= \frac{\partial f(x_k + \alpha p_k)}{\partial \alpha} \\ &= \frac{\partial f(x_k + \alpha p_k)}{\partial x_k} \cdot \frac{\partial x_k}{\partial \alpha} \\ &= \nabla f(x_k + \alpha p_k) \cdot p_k. \end{aligned}$$

Table 1 lists the experiment parameters for this problem. The implementation of line search remains essentially unchanged from that used in homeworks #1 and #2. When α in the extended Rosenbrock function was set to 1, the model converged too quickly. As such, α was set to 100.

Table 1: Experiment parameters for problem #3

Name	**Value**
n	1,000
α	100
x_0	$[-1]^n$
α_0	0
c_1	0.1
c_2	0.45

When $m = 0$, i.e., the algorithm behaves in a memoryless fashion and converges as traditional line search, it took 2,340 iterations to converge (not shown). Figure 1 shows the results for different values of m. As expected, when $m = 1$, it took the longest to converge. For $m \geq 2$, the convergence rate was essentially the same. While for $m = 2$ and $m = 5$, the convergence was marginally, the difference was marginal and not unexpected as excluding some gradients may cause the algorithm to perform better in some cases. For $m = 10$, $m = 20$, and maximum m (i.e., save all previous results), the algorithm performed essentially the same. This means that only the most recent gradients affected the results and memory of distant iterations does not improve the results for this function.

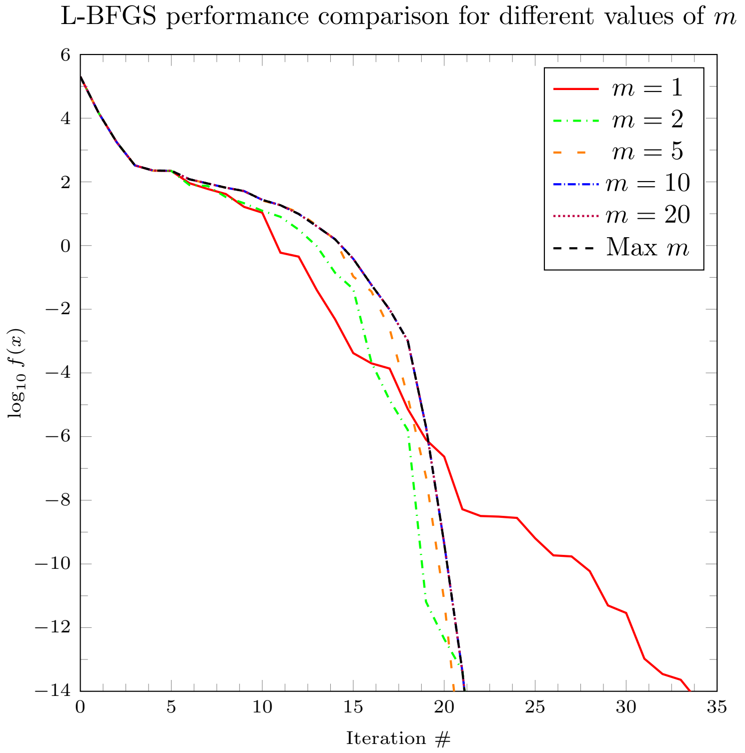

Figure 1: Convergence of L-BFGS for different values of m on the extended Rosenbrock function with $\alpha = 100$