Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4180 Optimization Theory Spring 2016

Exercise set 4 solution

1 The file `BFGS.m` on the course webpage contains an implementation of the BFGS method with a line search procedure satisfying the Wolfe conditions.

2 In the one-dimensional case, we no longer deal with matrices and vectors in the BFGS and DFP methods; H_k, p_k, y_k and s_k are scalars, and all gradients become standard derivatives. In addition, we are considering methods without line search, e.g. with step lengths $\alpha_k = 1$. For the BFGS method, we thereby get

$$x_{k+1} = x_k - H_k^{BFGS} f'(x_k), \tag{1}$$

where (equation 6.17 in N&W)

$$H_{k+1}^{BFGS} = \left(1 - \frac{s_k y_k}{s_k y_k}\right) H_k^{BFGS} \left(1 - \frac{y_k s_k}{s_k y_k}\right) + \frac{s_k^2}{s_k y_k} = \frac{s_k}{y_k}.$$

With $s_k = x_{k+1} - x_k$ and $y_k = f'(x_{k+1}) - f'(x_k)$, we find that

$$H_k^{BFGS} = \frac{s_{k-1}}{y_{k-1}} = \frac{x_k - x_{k-1}}{f'(x_k) - f'(x_{k-1})},$$

and by inserting into (1):

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f'(x_k) - f'(x_{k-1})} f'(x_k).$$

The same reasoning goes for the DFP method; we have

$$x_{k+1} = x_k - H_k^{DFP} f'(x_k),$$

where (equation 6.15 in N&W)

$$H_{k+1}^{DFP} = H_k^{DFP} - \frac{H_k^{DFP} y_k y_k H_k^{DFP}}{y_k H_k^{DFP} y_k} + \frac{s_k^2}{s_k y_k} = \frac{s_k}{y_k} = \frac{x_{k+1} - x_k}{f'(x_{k+1}) - f'(x_k)},$$

and hence

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f'(x_k) - f'(x_{k-1})} f'(x_k).$$

3 We wish to show that the BFGS method where H_k is reset to the identity matrix at each iteration is equivalent to a certain nonlinear CG method, i.e. that the iterates x_k produced by the methods coincide. In the BFGS method with resetting, we have

$$x_{k+1} = x_k + \alpha_k p_k^{BFGS}, \tag{2}$$

where α_k results from an exact line search and

$$p_k^{BFGS} = -H_k \nabla f_k,$$
$$H_k = \left(I - \frac{s_{k-1}y_{k-1}^T}{y_{k-1}^T s_{k-1}}\right)\left(I - \frac{y_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}\right) + \frac{s_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}},$$

where I denotes the identity matrix, $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f_k - \nabla f_{k-1}$.

In the nonlinear CG method under consideration, we have (following Algorithm 5.4 in N&W)

$$x_{k+1} = x_k + \alpha_k p_k^{CG}, \tag{3}$$

where α_k results from an exact line search and

$$p_k^{CG} = -\nabla f_k + \beta_k p_{k-1},$$
$$\beta_k = \frac{\nabla f_k^T(\nabla f_k - \nabla f_{k-1})}{(\nabla f_k - \nabla f_{k-1})^T p_{k-1}},$$

with $p_0 = -\nabla f_0$. Comparing (2) and (3), we see that if $p_k^{BFGS} = p_k^{CG}$, the iterates x_k will coincide since the α_k will then be equal as they are given by an exact line search method. We therefore only need to show that the p_k coincide for both algorithms. But first, we may observe the useful result that for both p_k^{CG} and p_k^{BFGS},

$$\nabla f_{k+1}^T p_k = 0 = \nabla f_{k+1}^T s_k.$$

This is because $\phi(\alpha) = f(x_k + \alpha_k p_k)$ has a minimum at α_k since we are using an exact line search. Hence,

$$\phi'(\alpha_k) = \nabla f(x_k + \alpha_k p_k)^T p_k = \nabla f_{k+1}^T p_k = 0.$$

Also, since $s_k = x_{k+1} - x_k = \alpha_k p_k^{BFGS}$, we see that $\nabla f_{k+1}^T s_k = 0$. We are now ready to prove, by induction, that $p_k^{BFGS} = p_k^{CG}$. In the base case $k = 0$, we see that

$$p_0^{BFGS} = -H_0 \nabla f_0 = -I \nabla f_0 = -\nabla f_0 = p_0^{CG},$$

so the hypothesis holds. In the general case, we have

$$p_k^{BFGS} = -H_k \nabla f_k = -\left[\left(I - \frac{s_{k-1}y_{k-1}^T}{y_{k-1}^T s_{k-1}}\right)\left(I - \frac{y_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}\right) + \frac{s_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}\right]\nabla f_k.$$

But, since $s_{k-1}^T \nabla f_k = 0$, most of the terms drop out, and we are left with

$$\begin{aligned} p_k^{BFGS} &= -\left(I - \frac{s_{k-1}y_{k-1}^T}{y_{k-1}^T s_{k-1}}\right)\nabla f_k \\ &= -\nabla f_k + \frac{\alpha_{k-1} p_{k-1}^{BFGS}(\nabla f_k - \nabla f_{k-1})^T}{(\nabla f_k - \nabla f_{k-1})^T \alpha_{k-1} p_{k-1}^{BFGS}}\nabla f_k \\ &= -\nabla f_k + \frac{\nabla f_k^T(\nabla f_k - \nabla f_{k-1})}{(\nabla f_k - \nabla f_{k-1})^T p_{k-1}^{CG}} p_{k-1}^{CG} = p_k^{CG}, \end{aligned}$$

where we have used that $(\nabla f_k - \nabla f_{k-1})^T \nabla f_k = \nabla f_k^T(\nabla f_k - \nabla f_{k-1})$ in the last equality, in addition to $p_{k-1}^{BFGS} = p_{k-1}^{CG}$ from the induction hypothesis.

4 **a)** The Gauss-Newton method uses updates of the form

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k p_k, \\ p_k &= -(J_k^T J_k)^{-1} J_k^T r_k. \end{aligned}$$

The method is well-defined in a neighborhood of x^* if $(J_k^T J_k)^{-1}$ exists there. Since $J(x^*)$ has full rank, $J(x^*)^T J(x^*)$ is invertible, and by the inverse function theorem, an inverse exists in a neighborhood of x^*. Next, we see that the directional derivative of $f(x) = \frac{1}{2}||r(x)||^2$ in the search direction p_k is:

$$p_k^T \nabla f_k = p_k^T J_k^T r_k = p_k^T J_k^T J_k (J_k^T J_k)^{-1} J_k^T r_k = -p_k^T J_k^T J_k p_k = -||J_k p_k|| \leq 0.$$

Since J_k has full rank, equality is obtained if and only if $p_k = 0$. Thus, p_k is a descent direction.

b) We take the hint and use Theorem 3.7 in N&W. The theorem states that if we consider iterations of the form $x_{k+1} = x_k + p_k$, where $p_k = -B_k^{-1} \nabla f_k$, B_k is SPD, and $\{x_k\}$ converge to a point x^* such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, then the iterations converge superlinearily if and only if

$$\lim_{k \to \infty} \frac{||(B_k - \nabla^2 f(x^*))p_k||}{||p_k||} = 0. \tag{4}$$

In the Gauss-Newton algorithm, we have iterations of the form

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k p_k, \\ p_k &= -(J_k^T J_k)^{-1} \nabla f_k, \end{aligned}$$

so we have $B_k = J_k^T J_k$. As $x^T J_k^T J_k x = ||J_k x||^2 \geq 0$, with equality iff $x = 0$ (due to J_k having full rank), we see that the B_k are SPD. Furthermore, we know that the Gauss-Newton iterations converge to a minimum of $f(x) = \frac{1}{2}||r(x)||^2$, i.e. to the point x^*, where $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. What remains to show is that (4) holds.

First, we observe that

$$\nabla^2 f(x) = J(x)^T J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x),$$

meaning, since $r(x^*) = 0$, that

$$\nabla^2 f(x^*) = J(x^*)^T J(x^*).$$

Since each r_j is continuously differentiable, $J(x)$ is a continuous function, and so is $J(x)^T J(x)$. Therefore, since $x_k \to x^*$, we also have $J(x_k)^T J(x_k) \to J(x^*)^T J(x^*) = \nabla^2 f(x^*)$, meaning condition (4) is satisfied, so the convergence is indeed superlinear.