Exercise 3

January 30, 2024

Problem 1.

Assume that $B \in \mathbb{R}^{d\times d}$ is a non-singular matrix (not necessarily orthogonal), $\hat{c} \in \mathbb{R}^d$, and $f\colon \mathbb{R}^d \to \mathbb{R}$ is a real-valued function. Define the function $g(x) = f(Bx + \hat{c})$.

a) Find expressions for $\nabla g(x)$ and $H_g(x)$ in the terms of f, B, and $\hat{c}$. (*You can (and maybe should) use the chain rule for the computation of* ∇g *and* H_g.)

b) Let $x_0 \in \mathbb{R}^d$ and denote by x_1 the result of one Newton step starting at x_0 for the minimisation of g with (Armijo) backtracking line search using the parameters $0 < c < 1$ (sufficient decrease parameter), $0 < \rho < 1$ (contraction factor), and $\hat{\alpha} = 1$ (initial step length).

   Moreover, let $y_0 = Bx_0 + \hat{c}$ and denote by y_1 the result of one Newton step starting at y for the minimisation of f with the (Armijo) backtracking line search and the same parameters $0 < c < 1$, $0 < \rho < 1$, and $\hat{\alpha} = 1$. Show that

$$y_1 = Bx_1 + \hat{c}. \tag{1}$$

c) Show that the relation (1) in general does not hold for the gradient descent method unless B is an orthogonal matrix.

Solution.

a) For clarity, we set $y(x) = Bx + \hat{c}$ so that $g(x) = f(y(x))$.

   The gradient $\nabla g(x)$ is the vector of the partial derivatives of g. We can compute these using the chain rule as

$$\partial_i g(x) = \partial_i f(y(x)) = \sum_{k=1}^{d} \partial_k f(y(x)) \partial_i y_k(x).$$

   Now

$$y_k(x) = \sum_{j=1}^{d} (B_{kj} x_j + \hat{c}_j)$$

and therefore

$$\partial_i y_k(x) = B_{ki}$$

Thus we obtain that

$$\partial_i g(x) = \sum_{k=1}^{d} B_{ki} \partial_k f(y(x)).$$

For the gradient, we thus obtain the expression

$$\nabla g(x) = B^T \nabla f(Bx + \hat{c}).$$

For the Hessian of g, we require the second derivatives $\partial_{ij} g(x)$. Using the above expression for the first derivatives and again the chain rule, we see that

$$\begin{aligned}\partial_{ij} g(x) = \sum_{k=1}^{d} \partial_j \big(B_{ki} \partial_k f(y(x))\big) &= \sum_{k=1}^{d} B_{ki} \partial_j f(y(x)) \\ &= \sum_{k=1}^{d} \sum_{\ell=1}^{d} B_{ki} \partial_{\ell k} f(y(x)) \partial_j y_\ell(x).\end{aligned}$$

Inserting again the expression for $\partial_j y(x)$ that we derived above, we obtain that

$$\partial_{ij} g(x) = \sum_{k=1}^{d} \sum_{\ell=1}^{d} B_{ki} \partial_{\ell k} f(Bx + \hat{c}) B_{\ell j}.$$

We can recognise this as the ij-th entry of the product $B^T H_f(Bx+\hat{c})B$, which shows that

$$H_g(x) = B^T H_f(Bx + \hat{c}) B.$$

Note that these are special cases of more general chain rules: If $G\colon \mathbb{R}^d \to \mathbb{R}^d$ and $h(x) := f(G(x))$, then

$$\nabla h(x) = J_G(x)^T \nabla f(G(x))$$

and

$$H_h(x) = J_G(x)^T H_f(G(x)) J_G(x) + \sum_{k=1}^{d} \partial_k f(G(x)) H_{G_k}(x),$$

where $J_G(x)$ is the Jacobian of G (which in our case is simply B). These formulas can be proven in the same way as the chain rules for our case.

b) Denote by p the search direction for the Newton method for g and by q the search direction for the Newton method for f. Then

$$p = -H_g(x_0)^{-1}\nabla g(x_0) \quad \text{and} \quad q = -H_f(y_0)^{-1}\nabla f(y_0).$$

Inserting the expressions for H_g and ∇g that we derived above, we obtain that

$$\begin{aligned} p &= -\big(B^T H_f(Bx_0+\hat{c})B\big)^{-1}B^T\nabla f(Bx_0+\hat{c}) \\ &= -B^{-1}H_f(y_0)^{-1}(B^T)^{-1}B^T\nabla f(y_0) \\ &= -B^{-1}H_f(y_0)^{-1}\nabla f(y_0) = B^{-1}q, \end{aligned}$$

and thus $q = Bp$.

Now the next iterates are

$$x_1 = x_0 + \alpha p \qquad \text{and} \qquad y_1 = y_0 + \beta q,$$

where α and β are chosen according to backtracking Armijo line search. Since

$$y_0 + \beta q = Bx_0 + \hat{c} + \beta(Bp) = B(x_0 + \beta p) + \hat{c}, \tag{2}$$

we have to show that $\alpha = \beta$. Because of the definition of the step length in backtracking Armijo line search, it is thus enough to show that the same step lengths satisfy the Armijo condition for both of the cases.

From (2) we obtain that

$$f(y_0 + \beta q) = f\big(B(x_0 + \beta p) + \hat{c}\big) = g(x_0 + \beta p)$$

for all β. Moreover

$$\langle q, \nabla f(y_0)\rangle = \langle Bp, \nabla f(Bx_0+\hat{c})\rangle = \langle p, B^T\nabla f(Bx_0+\hat{c})\rangle = \langle p, \nabla g(x_0)\rangle.$$

Thus the Armijo condition

$$f(y_0 + \beta q) \le f(y_0) + c_1\beta\langle q, \nabla f(y_0)\rangle$$

holds if and only if the Armijo condition

$$g(x_0 + \beta p) \le g(x_0) + c_1\beta\langle p, \nabla g(x_0)\rangle$$

holds. Thus the same step lengths are chosen for both of the iterations, which concludes the proof.

c) Denote now by $\hat{p}$ and $\hat{q}$ the search directions for the gradient descent methods for g and f, respectively. Following the same argumentation as in part b), for (1) to hold we will again require that $B\hat{p} = \hat{q}$. However, by using the chain rule for the gradient, we get that

$$B\hat{p} = -B\nabla g(x_0) = -BB^T\nabla f(Bx_0 + \hat{c}) = -BB^T\nabla f(y_0) = BB^T\hat{q}.$$

If B is orthogonal, then $BB^T = I$, and thus the relation $B\hat{p} = \hat{q}$ holds. If B is not orthogonal, however, this will in general not be true and thus the relation (1) will not hold.

Problem 2.

Implement both the gradient descent method and Newton's method with backtracking line search. Apply your method to the minimisation of the *Rosenbrock function* $f\colon \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = 100(y - x^2)^2 + (1 - x)^2$.

The Newton direction is not necessarily a descent direction for this function, as f is not convex, and thus it might be necessary to modify the search directions in the Newton method. Do this by switching to the negative gradient direction, whenever the inequality

$$-\left\langle \nabla f(x_k), p_k^{\text{Newton}} \right\rangle \leq \epsilon \|\nabla f(x_k)\| \|p_k^{\text{Newton}}\|$$

holds (here $\epsilon > 0$ is some fixed small parameter).

Solution.

See possible solutions on the wiki page.

Problem 3.

Consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ (see Exercise 1, Problem 3a)

$$f(x,y) = \frac{x^2}{2} + x\cos y.$$

We want to apply a line search method for the minimisation of f, starting at the point $x_0 = (1, \frac{\pi}{4})$ and with the search direction $p_0 = (-1, 0)$. Formulate the (weak) Wolfe conditions for this situation! If the parameters are chosen as $c_1 = 0.1$ and $c_2 = 0.8$, what is the range of admissible values for the step length α?

Solution.

The Wolfe conditions (weak) are given as

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1\alpha_k\langle \nabla f(x_k), p_k\rangle \text{ (Armijo condition)},$$

$$\langle \nabla f(x_k + \alpha_k p_k), p_k\rangle \geq c_2\langle \nabla f(x_k), p_k\rangle \text{ (Curvature condition)}.$$

Choosing the step lengths such that the Wolfe conditions are satisfied ensures sufficiently decrease of the objective function from one iteration to the next one, and to ensure convergence of the line search method. For $x_0 = \left(1, \frac{\pi}{4}\right)$ and $p_0 = (-1, 0)$, the Armijo condition $f(x_0+\alpha p_0) \leq f(x_0)+c_1\alpha\langle \nabla f(x_0), p_0\rangle$ gives

$$\frac{(1-\alpha)^2}{2} + \frac{(1-\alpha)}{\sqrt{2}} \leq \frac{1}{2} + \frac{1}{\sqrt{2}} - c_1\alpha\left(1 + \frac{1}{\sqrt{2}}\right)$$

$$\begin{aligned}
&\Rightarrow \frac{(1-\alpha)(1-\alpha+\sqrt{2})}{2} \leq \frac{1+\sqrt{2}}{2} - c_1\alpha\frac{(2+\sqrt{2})}{2} \\
&\Rightarrow \alpha(\alpha - 2 - \sqrt{2} + c_1(2+\sqrt{2})) \leq 0 \text{ (since } \alpha > 0) \\
&\Rightarrow \alpha - (2+\sqrt{2}) + c_1(2+\sqrt{2}) \leq 0 \\
&\Rightarrow \alpha \leq (1-c_1)(2+\sqrt{2}).
\end{aligned}$$

Similarly, curvature condition $\langle \nabla f(x_0 + \alpha p_0), p_0\rangle \geq c_2\langle \nabla f(x_0), p_0\rangle$ gives

$$\left(1 + \frac{1}{\sqrt{2}} - \alpha, \frac{\alpha - 1}{\sqrt{2}}\right)\begin{pmatrix}-1\\0\end{pmatrix} \geq c_2\left(1 + \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)\begin{pmatrix}-1\\0\end{pmatrix}$$

$$\Rightarrow -1 - \frac{1}{\sqrt{2}} + \alpha \geq -c_2\left(1 + \frac{1}{\sqrt{2}}\right)$$

$$\Rightarrow \alpha \geq (1-c_2)\left(1 + \frac{1}{\sqrt{2}}\right).$$

By putting the values $c_1 = 0.1$ and $c_2 = 0.8$, we obtain the following admissible value range of α

$$0.3414 \leq \alpha \leq 3.0727.$$

Problem 4.

Consider the optimisation problem

$$\min_{x \in \mathbb{R}^2} f(x), \tag{3}$$

where the objective function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is defined as

$$f(x, y) = 2x^2 - 2xy^2 - 12x + y^4 + 2y^2 + 36.$$

a) Compute all stationary (critical) points of the optimisation problem (3) and also find all local and global minima.

b) Starting at the point $x_0 = (1, 2)$, compute one step of Newton's method with backtracking (Armijo) line search. Start with an initial step length $\alpha_0 = 1$, and use the parameters $c = 1/8$ (sufficient decrease parameter) and $\rho = 1/2$ (contraction factor).

Solution.

a) The gradient and the Hessian of f can be computed as

$$\nabla f(x, y) = \begin{pmatrix} 4x - 2y^2 - 12 \\ -4xy + 4y^3 + 4y \end{pmatrix}$$

and

$$H_f(x, y) = \begin{pmatrix} 4 & -4y \\ -4y & -4x + 12y^2 + 4 \end{pmatrix}.$$

The first-order necessary condition of optimality reads $\nabla f(x, y) = 0$, that is,

$$4x - 2y^2 - 12 = 0 \qquad \text{and} \qquad -4xy + 4y^3 + 4y = 0.$$

The second equation holds if either $y = 0$ or $-4x + 4y^2 + 4 = 0$.

In the case $y = 0$, the first equation simplifies to $4x - 12 = 0$, which gives us the stationary point $(x, y) = (3, 0)$.

In the case $4x + 4y^2 + 4 = 0$, the first equation simplifies to

$$4y^2 + 4 - 2y^2 - 12 = 0$$

which has the two solutions $y = \pm 2$. In both cases, the corresponding value of x is 5. Thus we have the additional stationary points $(x, y) = (5, \pm 2)$.

In total, we have the stationary points $(3, 0)$, $(5, 2)$, and $(5, -2)$.

Next, to check whether these are local or global minima, we have to find the Hessian matrix at these points. At $(3, 0)$ we have

$$H_f(3, 0) = \begin{pmatrix} 4 & 0 \\ 0 & -8 \end{pmatrix},$$

which is obviously indefinite. Thus this point is a saddle point.

At $(5, \pm 2)$ we obtain the Hessians

$$H_f(5, 2) = \begin{pmatrix} 4 & -8 \\ -8 & 32 \end{pmatrix} \quad \text{and} \quad H_f(5, -2) = \begin{pmatrix} 4 & 8 \\ 8 & 32 \end{pmatrix}.$$

Since the diagonal entries of both these matrices are positive and the determinants are 64 in both cases, both matrices are positive definite. Thus $(5, -2)$ and $(5, 2)$ are strict local minima.

Finally, we need to decide if one of the local minima is actually a global minimum. For this, we first show that the function f actually has a global minimum. Being a polynomial, f is obviously continuous. Moreover, we can rewrite f as

$$f(x, y) = x^2 - 12x + 36 + x^2 - 2xy^2 + y^4 + 2y^2 = (x-6)^2 + (x-y^2)^2 + 2y^2.$$

Thus

$$f(x, y) \geq (x - 6)^2 + 2y^2$$

for all $(x, y) \in \mathbb{R}^2$, which shows that f is coercive. We have therefore shown that f is coercive and continuous, which implies that it has a global minimum.

As seen above, the only candidates for a global minimum are the local minima $(5, \pm 2)$. Since $f(5, 2) = f(5, -2) = 10$, it follows that both points are actually global minimisers of f.

b) First, we have to find the search direction for the starting point $x_0 = (1, 2)$, which is

$$p_0 = -(H_f(x_0))^{-1} \nabla f(x_0). \tag{4}$$

We compute

$$\nabla f(x_0) = \begin{pmatrix} -16 \\ 32 \end{pmatrix}, \quad H_f(x_0) = \begin{pmatrix} 4 & -8 \\ -8 & 48 \end{pmatrix}$$

and

$$(H_f(x_0))^{-1} = \frac{1}{128}\begin{pmatrix} 48 & 8 \\ 8 & 4 \end{pmatrix}.$$

By using these values in (4), we obtain $p_0 = (4, 0)^T$. We now need to check whether the Armijo condition is satisfied for the step length $\alpha_0 = 1$. The Armijo condition is

$$f(x_0 + \alpha_0 p_0) \leq f(x_0) + c\alpha_0 \nabla f(x_0)^T p_0,$$

Inserting the values for x_0, p_0, α, and c, this becomes the inequality

$$10 \leq 34,$$

which holds true. Therefore, the Armijo condition is satisfied at $\alpha_0 = 1$. Consequently, we can choose the initial step length $\alpha_0 = 1$, which gives the first step of Newton's method as

$$x_1 = x_0 + \alpha_0 p_0 = (5, 2).$$

Note: Since $(5, 2)$ is a global minimiser too, the iteration will stop here.

Problem 5.

Consider the optimization problem

$$\min_{x \in \mathbb{R}^2} f(x), \tag{5}$$

where the objective function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(x, y) = x^2 - 2x + x^2y^2 - 2xy.$$

a) Compute all stationary (critical) points of the optimisation problem (5) and also find all its local and global minima.

b) Starting at the point $x_0 = (0, 0)$ compute one step of the gradient (steepest) descent method. Ensure that the step length satisfies the (weak) Wolfe conditions with $c_1 = \frac{1}{4}$ and $c_2 = \frac{3}{4}$.

c) In part b), can you also use the Newton search direction?

Solution.

a) The gradient and Hessian of f are

$$\nabla f(x, y) = \begin{pmatrix} 2x - 2 + 2xy^2 - 2y \\ 2x^2y - 2x \end{pmatrix}$$

and

$$H_f(x,y) = \begin{pmatrix} 2+2y^2 & 4xy-2 \\ 4xy-2 & 2x^2 \end{pmatrix}.$$

The first-order necessary optimality condition requires that $\nabla f(x,y) = 0$. Therefore, we have to solve the equations

$$2x - 2 + 2xy^2 - 2y = 0 \qquad \text{and} \qquad 2x^2y - 2x = 0.$$

The second equation can be rewritten as

$$2x(xy-1) = 0,$$

which implies that either $x = 0$ or $xy = 1$.

Now, for $x = 0$ we obtain from the first equation that $y = -1$. Thus we have the stationary point $(x,y) = (0,-1)$.

For the case $xy = 1$, the first equation simplifies to $2x - 2 = 0$, and thus we have the stationary point $(x,y) = (1,1)$.

In total, we have two stationary points of the optimisation problem (5), namely $(0,-1)$ and $(1,1)$.

Next, to check whether these are local or global minima, we have to find the Hessian matrix. At the stationary point $(0,-1)$, the Hessian is

$$H_f(0,-1) = \begin{pmatrix} 4 & -2 \\ -2 & 0 \end{pmatrix},$$

which is indefinite, as the trace is positive but the determinant is negative. Therefore, the stationary point $(0,-1)$ is saddle point, that means, at $(0,-1)$ the optimisation problem (5) has neither a local minimum nor a local maximum.

Further, the Hessian matrix at $(1,1)$ is

$$\nabla^2 f(1,1) = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}.$$

This matrix is positive definite, as its trace and determinant are both positive. Thus $(1,1)$ is a strict local minimum.

Now, we have to check that whether $(1,1)$ is a global minimum. The function f is obviously continuous (being a polynomial). Thus, if f were coercive, we could conclude that f has a global minimum with $(1,1)$ being the only candidate, and we were done. Unfortunately, this is not the case: For $x = 0$, the function value $f(0,y)$ is 0 independent of y, and thus the 0-level set of f is unbounded. We therefore devise another method for showing that $(1,1)$ is a global minimum.

We note that we can write f as

$$f(x,y) = (x-1)^2 + (xy-1)^2 - 2.$$

Thus

$$f(x,y) \geq -2 \text{ for all } (x,y) \in \mathbb{R}^2$$

with equality if and only if $x = 1$ and $xy = 1$, that is, if and only if $(x,y) = (1,1)$. This shows that $(1,1)$ is indeed the global minimiser of f.

b) First, we have to find the search direction for the starting point $x_0 = (0,0)$, which is

$$p_0 = -\nabla f(x_0). \tag{6}$$
$$\nabla f(x_0) = (-2,0)^T.$$

By using above values in (6), we obtain $p_0 = (2,0)^T$. The Wolfe conditions (weak) are given as

$$f(x_0 + \alpha_0 p_0) \leq f(x_0) + c_1 \alpha_0 \langle \nabla f(x_0), p_0 \rangle \qquad \text{(Armijo condition)} \tag{7}$$

and

$$\langle \nabla f(x_0 + \alpha_0 p_0), p_0 \rangle \geq c_2 \langle \nabla f(x_0), p_0 \rangle \qquad \text{(Curvature condition)}. \tag{8}$$

Condition (5) with $c_1 = \frac{1}{4}$ gives $\alpha \leq \frac{3}{4}$, and (8) with $c_2 = \frac{3}{4}$ gives $\alpha \geq \frac{1}{8}$. Therefore, we have $\frac{1}{8} \leq \alpha \leq \frac{3}{4}$. Now, we can choose the initial step length $\alpha_0 = \frac{1}{2}$ for which one step of gradient method is $x_1 = x_0 + \alpha_0 p_0 = (1,0)$.

c) We compute the search direction for Newton's method at the starting point $x_0 = (0,0)$ as

$$p_0 = -H_f(x_0)^{-1} \nabla f(x_0). \tag{9}$$

Here

$$\nabla f(x_0) = \begin{pmatrix} -2, 0 \end{pmatrix},$$
$$H_f(x_0) = \begin{pmatrix} 2 & -2 \\ -2 & 0 \end{pmatrix},$$
$$H_f(x_0))^{-1} = -\frac{1}{4} \begin{pmatrix} 0 & 2 \\ 2 & 2 \end{pmatrix}.$$

By using the above values in (9), we obtain $p_0 = (0,-1)^T$. Now $\langle p_0, \nabla f(x_0) \rangle = 0 \not< 0$. Therefore, the search direction $p_0 = (0,-1)^T$ is not a descent direction, and thus we cannot use this search direction for the line search.