Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4180
Optimisation I
Spring 2018

Solutions to exercise set 3

$\boxed{1}$ Consider the quadratic function

$$f(x) = \tfrac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x,$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric and positive definite matrix and $b \in \mathbb{R}^n$.

a) Let $p \in \mathbb{R}^n$ be a direction satisfying the inequality $\nabla f(x)^{\mathrm{T}}p < 0$. Compute analytically the steplength $\alpha_{x,p}$, which solves the linesearch problem $\min_{\alpha>0} f(x+\alpha p)$

Solution: First of all, to avoid trivial cases let us note that $p \neq 0$ and $\nabla f(x) = Ax - b \neq 0$ owing to the inequality $\nabla f(x)\mathrm{T}p < 0$.
Now, let us look at the first order necessary conditions for $\alpha_{x,p}$ to be a minimizer:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f(x+\alpha_{x,p}p) = p^{\mathrm{T}}\nabla f(x+\alpha_{x,p}p) = p^{\mathrm{T}}[A(x+\alpha_{x,p}p) - b] = 0,$$

or

$$\alpha_{x,p} = -\frac{p^{\mathrm{T}}[Ax-b]}{p^{\mathrm{T}}Ap} > 0,$$

since $p^{\mathrm{T}}Ap > 0$ owing to A being positive definite, and $p^{\mathrm{T}}[Ax-b] = p^{\mathrm{T}}\nabla f(x) < 0$ by our assumption.
Since $\mathrm{d}^2/\mathrm{d}\alpha^2 f(x+\alpha p) = p^{\mathrm{T}}Ap > 0$ the linesearch problem is strictly convex, and therefore $\alpha_{x,p}$ is the unique global minimum.

b) Let $x, p \in \mathbb{R}^n$ and $\alpha_{x,p} > 0$ be as in the previous question. Show that the steplength $\alpha_{x,p}$ satisfies the strong Wolfe conditions if and only if $c_1 \leq 1/2$.

Solution: Clearly the strong curvature condition is satisfied because $\nabla f(x+\alpha_{x,p}p)^{\mathrm{T}}p = \mathrm{d}/\mathrm{d}\alpha f(x+\alpha_{x,p}p) = 0$, thus the "new" slope is 0 and must be smaller than or equal in magnitude than the slope we have started with.
We check the sufficient decrease condition now:

$$f(x+\alpha_{x,p}p) - f(x) = \frac{1}{2}\alpha_{x,p}^2 p^{\mathrm{T}}Ap + \alpha p^{\mathrm{T}}(Ax-b) = -\frac{1}{2}\frac{[p^{\mathrm{T}}(Ax-b)]^2}{p^{\mathrm{T}}Ap} < 0,$$

while

$$c_1\alpha_{x,p}\nabla f(x)^{\mathrm{T}}p = -c_1\frac{[p^{\mathrm{T}}(Ax-b)]^2}{p^{\mathrm{T}}Ap}$$

. Thus the sufficient decrease condition implies the inequality $c_1 \leq 1/2$.

c) Let $A = Q\Lambda Q^{\mathrm{T}}$ be the eigenvalue decomposition of A, where Λ is a diagonal matrix with eigenvalues on the diagonal, and columns of Q are the orthonormal eigenvectors of A. In particular, $Q^{\mathrm{T}}Q = I$, where $I \in \mathbb{R}^{n\times n}$ is the identity matrix.

Show that applying the steepest descent method with exact linesearch to the problem $\min_{x\in\mathbb{R}^n} 0.5x^{\mathrm{T}}Ax - b^{\mathrm{T}}x$ is equivalent to applying the steepest descent method with exact linesearch to $\min_{y\in\mathbb{R}^n} 0.5y^{\mathrm{T}}\Lambda y$, in the following sense: if $x_0 = Qy_0 + A^{-1}b$ then the iterates generated by the two methods satisfy the same relation, $x_k = Qy_k + A^{-1}b$, $k \geq 1$.

In this sense, the behaviour of the steepest descent method is insensitive with respect to translation or orthogonal transformation of coordinates.

Solution: Assume that $x_k = Qy_k + A^{-1}b$, $k \geq 0$, and let us establish the same relation after one step of steepest descent.

On the "x"-side we have

$$x_{k+1} = x_k - \alpha_{x_k,-\nabla f(x_k)}\nabla f(x_k) = x_k - \frac{(Ax_k - b)^{\mathrm{T}}(Ax_k - b)}{(Ax_k - b)^{\mathrm{T}}A(Ax_k - b)}(Ax_k - b).$$

We substitute now $x_k = Qy_k + A^{-1}b$, which in particular means that $Ax_k - b = AQy_k = Q\Lambda y_k$, to get the equality

$$\begin{aligned} x_{k+1} &= Qy_k + A^{-1}b - \frac{y_k^{\mathrm{T}}\Lambda^{\mathrm{T}}Q^{\mathrm{T}}Q\Lambda y_k}{y_k^{\mathrm{T}}\Lambda^{\mathrm{T}}Q^{\mathrm{T}}Q\Lambda Q^{\mathrm{T}}Q\Lambda y_k}Q\Lambda y_k \\ &= Q\left[y_k - \frac{y_k^{\mathrm{T}}\Lambda^2 y_k}{y_k^{\mathrm{T}}\Lambda^3 y_k}\Lambda y_k\right] + A^{-1}b. \end{aligned}$$

Similarly, on the "y" side we can write:

$$y_{k+1} = y_k - \alpha_{y_k,-\Lambda y_k}\Lambda y_k = y_k - \frac{[\Lambda y_k]^{\mathrm{T}}\Lambda y_k}{[\Lambda y_k]^{\mathrm{T}}\Lambda\Lambda y_k}\Lambda y_k = y_k - \frac{y_k^{\mathrm{T}}\Lambda^2 y_k}{y_k^{\mathrm{T}}\Lambda^3 y_k}\Lambda y_k,$$

where we have used the fact that $\nabla_y[0.5y^{\mathrm{T}}\Lambda y] = \Lambda y$.

In view of the two equalities above, the proof is complete.

$\boxed{2}$ Let f be twice continuously differentiable in a vicinity of $x_0 \in \mathbb{R}^n$. Assume that $\nabla^2 f(x_0)$ is positive definite and consider the Newton's direction $p_x = -[\nabla^2 f(x_0)]^{-1}\nabla f(x_0)$ together with the unit Newton's step $x_1 = x_0 + p_x$.

Let us now perform an affine transformation (translation, rotation, and scaling) of coordinates $x = By + c$, where $B \in \mathbb{R}^{n\times n}$ is a non-singular matrix (not necessarily orthogonal), and $c \in \mathbb{R}^n$ is some vector. Demonstrate that Newton's method is insensitive with respect to such transformations: that is, if $g(y) = f(By + c) = f(x)$, $x_0 = By_0 + c$, and finally $y_1 = y_0 - [\nabla^2 g(y_0)]^{-1}\nabla g(y_0)$ then $x_1 = By_1 + c$.

Solution:

$$\begin{aligned} \frac{\partial g}{\partial y_i}(y) &= \sum_{k=1}^{n}\frac{\partial f}{\partial x_k}(By + c)\frac{\partial x_k}{\partial y_i} = \sum_{k=1}^{n}\frac{\partial f}{\partial x_k}(By + c)B_{ki}, \\ \frac{\partial^2 g}{\partial y_i \partial y_j}(y) &= \frac{\partial}{\partial y_j}\sum_{k=1}^{n}\frac{\partial f}{\partial x_k}(By + c)B_{ki} = \sum_{k=1}^{n}\sum_{\ell=1}^{n}\frac{\partial^2 f}{\partial x_k \partial x_\ell}(By + c)B_{ki}B_{\ell j}, \end{aligned}$$

and therefore

$$\begin{aligned}\nabla_y g(y) &= B^{\mathrm{T}}\nabla_x f(By+c)\\ \nabla_y^2 g(y) &= B^{\mathrm{T}}\nabla_x^2 f(By+c)B.\end{aligned}$$

As a result

$$\begin{aligned}y_1 &= y_0 - [B^{\mathrm{T}}\nabla_x^2 f(By+c)B]^{-1}B^{\mathrm{T}}\nabla_x f(By+c)\\ &= y_0 - B^{-1}[\nabla_x^2 f(By+c)]^{-1}B^{-\mathrm{T}}B^{\mathrm{T}}\nabla_x f(By+c)\\ &= y_0 - B^{-1}[\nabla_x^2 f(By+c)]^{-1}\nabla_x f(By+c),\\ x_1 &= x_0 - [\nabla_x^2 f(x_0)]^{-1}\nabla_x f(x_0)\\ &= B\{y_0 - B^{-1}[\nabla_x^2 f(By_0+c)]^{-1}\nabla_x f(By_0+c)\} + c = By_1 + c.\end{aligned}$$

$\boxed{3}$ Let $A \in \mathbb{R}^{n\times n}$ be an SPD matrix with the eigenvalue decomposition $A = Q\Lambda Q^T$, and let $b \in \mathbb{R}^n$ be an arbitrary vector. We put $x^* = A^{-1}b$ to be the optimal solution of the quadratic unconstrained minimization problem $\min_{x\in\mathbb{R}^n} 0.5x^{\mathrm{T}}Ax - b^{\mathrm{T}}x$. Suppose that the eigenvaluse of A are sorted as $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. During the lecture we have discussed that for starting point of the type $x_0 = x^* + \lambda_1^{-1}q_1 + \lambda_n^{-1}q_n$, where q_i are orthonormal eigenvectors of A (columns of Q) corresponding to eigenvalues λ_i, the steepest descent method with exact linesearch for this problem generates iterates satisfying

$$\|x_k - x^*\| = \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^k \|x_0 - x^*\|,$$

which converges to zero linearly, and arbitrarily slowly for large condition numbers $\mathrm{cond}(A) = \lambda_n/\lambda_1$. Approximately, the number of iterations needed to achieve some prescribed tolerance scales proportionally to the condition number of A.

a) Implement the steepest descent method with exact linesearch for this problem and verify the estimate above numerically.

Hint: one can generate random positive definite matrices for example as follows:

```
import numpy as np
N = 10
# generate NxN random matrix
X = np.random.randn(N,N)
# generate NxN orthogonal matrix from it
Q = np.linalg.qr(X)[0]
# generate some random eigenvalues between lam_min and lam_max
lam_min   = 1.0
lam_max   = 100.0
lmbda     = lam_min + (lam_max-lam_min)*np.sort(np.random.rand(N))
lmbda[0]  = lam_min
lmbda[-1]=lam_max
Lambda    = np.diag(lmbda)
A = np.matmul(Q,np.matmul(Lambda,Q.T))
# random vector
b = np.random.randn(N)
# A^{-1}b
xstar = np.linalg.solve(A,b)
```

```
# starting point
x0      = xstar + 1.0/lmbda[0]*Q[:,0] + 1.0/lmbda[-1]*Q[:,-1]
```

Solution: See a possible implementation on the Wiki

b) Not everyone has given up on the steepest descent method. Consider for example the following accelerated version of the method due to Nesterov:

$$\begin{aligned} p_k &= -\nabla f(x_k), \\ y_{k+1} &= x_k + \lambda_n^{-1} p_k, \\ x_{k+1} &= s_1 y_{k+1} + s_0 y_k, \end{aligned}$$

where we put $y_0 = x_0$, $s_0 = -(\lambda_n^{1/2} - \lambda_1^{1/2})/(\lambda_n^{1/2} + \lambda_1^{1/2})$, and $s_1 = 1.0 - s_0$. Implement this method and verify numerically, that the number of iterations needed to achieve some prescribed tolerance scales proportionally to the square root of the condition number of A, $\lambda_n^{1/2}/\lambda_1^{1/2}$.

Solution: See a possible implementation on the Wiki

$\boxed{4}$ Implement both the steepest descent method and the Newton's method with linesearch satisfying Wolfe conditions (use a bisection algorithm for this).

Apply the method to minimizing the Rosenbrock function:

$$f(x, y) := 100(y - x^2)^2 + (1 - x)^2.$$

As Newtons direction is not necessarily a descent direction, we can simply use the steepest descent direction when the following inequality holds:

$$-\nabla f(x_k)^{\mathrm{T}} p_k^{\mathrm{Newton}} \leq \varepsilon \|\nabla f(x_k)\| \|p_k^{\mathrm{Newton}}\|,$$

that is, when the angle between the Newton's direction and the steepest descent direction gets dangerously close to $\pi/2$ or exceeds this value (ε is a small positive parameter in the inequality above).

Verify numerically that the unit Newton's steps are accepted by the linesearch algorithm provided that the sufficient decrease parameter satisfies the inequality $0 < c_1 < 1/2$.

Solution: See a possible implementation on the Wiki