Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4180 Optimisation I Spring 2017

Exercise set 4

1 Applying Algorithm 5.2 in Nocedal & Wright, we find that

$$\begin{aligned} &x_0 = (0,0,0), && r_0 = (-1,0,-1), && && p_0 = (1,0,1), && \alpha_0 = 1,\\ &x_1 = (1,0,1), && r_1 = (0,2,0), && \beta_1 = 2, && p_1 = (2,2,2), && \alpha_1 = 1,\\ &x_2 = (3,2,3), && r_3 = (0,0,0). \end{aligned}$$

Since $r_3 = 0$—which it should as convergence is guaranteed within 3 steps—we stop and conclude that $x = (3,2,3)$ solves the linear system.

2 **a)** The least squares problem is an unconstrained minimisation problem for the function $f(x) = \|Ax - b\|^2$ on $\mathbb{R}^n$. Observe that f is smooth, and that

$$\nabla f(x) = 2A^\top(Ax - b) \qquad \text{and} \qquad \nabla^2 f(x) = 2A^\top A.$$

Calculation of ∇f follows either from the chain rule in the multivariable setting, or by direct expansion

$$\|Ax - b\|^2 = (Ax - b)^\top(Ax - b) = x^\top A^\top Ax - 2b^\top Ax + b^\top b.$$

Matrix $A^\top A$ is symmetric, and also positive semi-definite, because

$$v^\top A^\top Av = (Av)^\top Av = \|Av\|^2 \geq 0 \quad \text{for all} \quad v \in \mathbb{R}^n.$$

Hence, f is convex and we infer that every critical point is a global minimiser (and conversely). As such, x^* minimises f if and only if $\nabla f(x^*) = 0$. In other words,

$$A^\top Ax^* = A^\top b.$$

b) If we can show that the normal equations admit a solution, then we are done. Specifically, this amounts to proving that $A^\top b \in \operatorname{ran} A^\top A$. Now,

$$\operatorname{ran} A^\top A = \big(\ker(A^\top A)^\top\big)^\perp = \big(\ker A^\top A\big)^\perp$$

from the fundamental theorem of linear algebra, where $B^\perp$ denotes the orthogonal complement of a set B. Since $\ker A^\top A = \ker A$ (why?), it follows that

$$\operatorname{ran} A^\top A = (\ker A)^\perp.$$

Observe next that if $y \in \ker A$, then $(A^\top b)^\top y = b^\top(Ay) = 0$. Consequently, we have $A^\top b \in (\ker A)^\perp = \operatorname{ran} A^\top A$, as desired.

c) If rank $A = n$, then by the rank-nullity theorem the null space (kernel) of A is trivial, that is, $\ker A = \{0\}$. Thus $Av = 0$ if and only if $v = 0$, and so $\nabla^2 f = A^\top A$ is positive definite:

$$v^\top A^\top Av = (Av)^\top Av = \|Av\|^2 > 0 \quad \text{for all} \quad v \in \mathbb{R}^n \setminus \{0\}.$$

Therefore f is strictly convex, which implies uniqueness of the global minimiser.

d) From a) and b) we know that f is a convex function whose set $\Omega \subset \mathbb{R}^n$ of minimiser(s) is nonempty. Moreover, it follows that Ω is convex (why?), so we may write the new optimisation problem as

$$\min_{x\in\Omega} g(x), \qquad \text{where} \qquad g(x) = \|x\|^2.$$

Note that $\nabla^2 g = 2I_{n\times n}$ is symmetric positive definite. In particular, g is strictly convex on Ω[1], and has at most one solution $x^\dagger \in \Omega$.

If rank $A = n$, then $\Omega = \{x^*\}$, from which we conclude that $x^\dagger = x^*$. If, however rank $A < n$, then Ω is at least a one-dimensional subspace of $\mathbb{R}^n$ (there is at least one free parameter in the normal equations). In this case, Ω is unbounded, but we are saved by coercivity of g. Indeed,

$$x \in \Omega \text{ with } \|x\| \to \infty \qquad \text{implies} \qquad g(x) = \|x\|^2 \to \infty.$$

Since g is lower semi-continuous—in fact, smooth—and coercive, it admits a global minimum $x^\dagger \in \Omega$.

e) By construction of the optimisation problem in d), $x^\dagger$ satisfies $A^\top Ax^\dagger = A^\top b$. In order to see that $x^\dagger \in \operatorname{ran} A^\top$, observe first that

$$\mathbb{R}^n = \operatorname{ran} A^\top \ominus \ker A$$

(orthogonal direct sum) by the rank–nullity theorem. Hence, $x^\dagger$ may be written uniquely as $x^\dagger = y + z$ for some $y \in \operatorname{ran} A^\top$ and $z \in \ker A$, satisfying $y^\top z = 0$. We want to show that $z = 0$. This rests upon two observations: 1) perturbing z inside $\ker A$ has no effect on the normal equations (or the value of f):

$$A^\top A(y + \widetilde{z}) = A^\top Ay + A^\top(A\widetilde{z}) = A^\top Ay$$

for any $\widetilde{z} \in \ker A$; and 2) by orthogonality between $\operatorname{ran} A^\top$ and $\ker A$ we have

$$g(x^\dagger) = \|x^\dagger\|^2 = \|y\|^2 + y^\top z + \|z\|^2 = \|y\|^2 + \|z\|^2,$$

and similarly

$$g(y + \widetilde{z}) = \|y\|^2 + \|\widetilde{z}\|^2.$$

If $z \neq 0$, we can therefore just pick any $\widetilde{z} \in \ker A$ with $\|\widetilde{z}\| < \|z\|$ and obtain that $g(y + \tilde{z}) < g(x^\dagger)$. But this is a contradiction to the fact that $x^\dagger$ minimises g, so z must indeed be 0.

[1] Remark: never forget that we cannot talk about convexity of a function unless its underlying domain is convex (why not?).

3 **a)** We provide an inductive argument, showing that

$$r_{k-1}^{\text{CG}} = s_{k-1}, \qquad p_{k-1}^{\text{CG}} = p_{k-1}, \qquad \alpha_{k-1}^{\text{CG}} = \alpha_{k-1}, \qquad \text{and} \qquad x_k^{\text{CG}} = x_k$$

for any k, assuming x_0 arbitrary but equal for both methods, with superscript "CG" for the CG-parameters. Remark: CG-algorithm is well-defined because $A^\top A$ is symmetric positive definite ($\operatorname{rank} A = n$).
Base case $k = 1$ follows from

$$r_0^{\text{CG}} = (A^\top A)x_0 - A^\top b, \qquad r_0 = Ax_0 - b, \qquad \text{and} \qquad s_0 = A^\top r_0 = r_0^{\text{CG}},$$

so that

$$p_0^{\text{CG}} = -r_0^{\text{CG}} = -s_0 = p_0,$$

and

$$\alpha_0^{\text{CG}} = \frac{\left\|r_0^{\text{CG}}\right\|^2}{\left(p_0^{\text{CG}}\right)^\top (A^\top A) p_0^{\text{CG}}} = \frac{\left\|r_0^{\text{CG}}\right\|^2}{\left\|Ap_0^{\text{CG}}\right\|^2} = \frac{\|s_0\|^2}{\|Ap_0\|^2} = \alpha_0.$$

Therefore

$$x_1^{\text{CG}} = x_0 + \alpha_0^{\text{CG}} p_0 = x_0 + \alpha_0 p_0 = x_1.$$

Suppose next that the induction hypothesis is true for some $k \in \mathbb{Z}_+$. Then

$$\begin{aligned} r_k^{\text{CG}} &= r_{k-1}^{\text{CG}} + \alpha_{k-1}^{\text{CG}} A^\top A p_{k-1}^{\text{CG}} \\ &= s_{k-1} + \alpha_{k-1} A^\top A p_{k-1} \\ &= A^\top \left(r_k + \alpha_{k-1} A p_{k-1}\right) \\ &= A^\top r_k \\ &= s_k, \end{aligned}$$

$$p_k^{\text{CG}} = -r_k^{\text{CG}} + \frac{\left\|r_k^{\text{CG}}\right\|^2}{\left\|r_{k-1}^{\text{CG}}\right\|^2} p_{k-1}^{\text{CG}} = -s_k + \frac{\|s_k\|^2}{\|s_{k-1}\|^2} p_k = p_k,$$

and

$$\alpha_k^{\text{CG}} = \frac{\left\|r_k^{\text{CG}}\right\|^2}{\left\|Ap_k^{\text{CG}}\right\|^2} = \frac{\|s_k\|^2}{\|Ap_k\|^2} = \alpha_k,$$

so, most importantly,

$$x_k^{\text{CG}} = x_{k-1}^{\text{CG}} + \alpha_{k-1}^{\text{CG}} \, p_{k-1}^{\text{CG}} = x_{k-1} + \alpha_{k-1} \, p_{k-1} = x_k.$$

b) There are two key arguments in this exercise: 1) the equivalence between the given algorithm and the CG-algorithm for the solution of $A^\top Ax = A^\top b$, and 2) the characterisation in 2 e) of optimisation problem (2).
Hence, by algorithmic equivalence, if the new algorithm converges to some $x^\dagger$, then necessarily $A^\top Ax^\dagger = A^\top b$. Since $x_0 = 0$, it follows that $p_0 = A^\top b \in \operatorname{ran} A^\top$, and $x_1 = \alpha_0 p_0 \in \operatorname{ran} A^\top$. All subsequent search directions are of the form

$$p_k = -A^\top r_k + \beta_k p_{k-1},$$

from which we infer both that $p_k \in \operatorname{ran} A^\top$ and $x_k = x_{k-1} + \alpha_{k-1} p_{k-1} \in \operatorname{ran} A^\top$. Therefore, if the given algorithm converges to some $x^\dagger$, then $x^\dagger \in \operatorname{ran} A^\top$, yielding a solution of optimisation problem (2).

But does the algorithm converge, and if so, in at most r steps? We first need to examine if all algorithmic operations are legal; especially, whether we risk division by zero anywhere. This could occur if

$$\|Ap_k\| = 0, \qquad \text{or} \qquad \|s_k\| = 0.$$

The latter is not a problem, because the algorithm has converged if $s_k = 0$ (in general, s_k is the current residual of $A^\top Ax = A^\top b$). Moreover, $Ap_k = 0$ if and only if $p_k \in \ker A = \big(\operatorname{ran} A^\top\big)^\perp$. In the previous paragraph we showed that $p_k \in \operatorname{ran} A^\top$ for all k, and so $Ap_k = 0$ if and only if $p_k = 0$. But $p_k = 0$ implies $s_k = 0$ also, and consequently, convergence.

Since all operations in the algorithm are legal, we can use the algorithmic equivalence with the CG-algorithm for the solution of $A^\top Ax = A^\top b$, and follow the proof of Theorem 5.3 in Nocedal & Wright, which shows that the generated p_k's are conjugate with respect to $A^\top A$. In the end of N&W's proof, Theorem 5.1 is invoked, stating that the algorithm will converge in at most n iterations. In our case, however, this reduces to at most r iterations, because solution $x^\dagger$ lies in the r-dimensional subspace $\operatorname{ran} A^\top \subset \mathbb{R}^n$, spanned by the r linearly independent vectors $p_0, \ldots, p_{r-1}$.

4 See file tma4180s17_ex04_4.m on the website.