Norwegian University of Science and Technology
Department of Mathematical Sciences

**Solutions to exercise set 4**

**1** Let

$$A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 2 \end{bmatrix} \qquad \text{and} \qquad b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Use the CG-method with initialisation $x_0 = 0$ for solving the linear system $Ax = b$.

**Solution:** 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** Assume that $A \in \mathbb{R}^{m\times n}$ is a matrix and that $b \in \mathbb{R}^m$.

**a)** Show that $x^* \in \mathbb{R}^n$ solves the *least squares problem*

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|^2, \tag{1}$$

if and only if $x^*$ satisfies the *normal equations*

$$A^{\mathrm{T}}Ax^* = A^{\mathrm{T}}b.$$

**Solution:** 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

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

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

$$\|Ax - b\|^2 = (Ax-b)^{\mathrm{T}}(Ax-b) = x^{\mathrm{T}}A^{\mathrm{T}}Ax - 2b^{\mathrm{T}}Ax + b^{\mathrm{T}}b.$$

Matrix $A^{\mathrm{T}}A$ is symmetric, and also positive semi-definite, because

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

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,

$$A^{\mathrm{T}}Ax^* = A^{\mathrm{T}}b.$$

**b)** Show that the optimization problem (1) admits a solution $x^* \in \mathbb{R}^n$.

**Solution:** There are many ways of proving this result; in particular, this is a special case of a so-called Frank–Wolfe's theorem, which states that is a quadratic function is bounded below on a non-empty polyhedron, then it attains its infimum on this polyhedron.

The latter result can be proved by induction in the number of spatial dimentions $n$.

If $n = 1$, then $A \in \mathbb{R}^{m\times 1}$, $b \in \mathbb{R}^m$, and $f(x) = b^{\mathrm{T}}b - 2A^{\mathrm{T}}bx + x^2A^{\mathrm{T}}A$. If $A^{\mathrm{T}}A \neq 0$ then the problem admits the unique global minumum $x^* = A^{\mathrm{T}}b/(A^{\mathrm{T}}A)$; otherwise any $x \in \mathbb{R}$ is a global minimum as then $A = 0$ and therefore $f(x) = b^2$ for any $x \in \mathbb{R}$.

Suppose now any $k$-dimensinal problem admits a solution. Let us represent $x \in \mathbb{R}^{k+1}$ as $\lambda y$, where $\lambda \geq 0$ and $y$ belongs to the unit sphere $S = \{\, x \in \mathbb{R}^{k+1} \mid \|x\| = 1 \,\}$. (Indeed, for any $x \in \mathbb{R}^{k+1} \setminus \{0\}$ we can simply put $\lambda = \|x\|$ and $y = x/\|x\|$.) Therefore, (1) is equivalent to the problem

$$\min_{\lambda\geq 0, y\in S} f(sy) = \min_{\lambda\geq 0, y\in S} \|b\|^2 - 2\lambda b^{\mathrm{T}}Ay + \lambda^2\|Ay\|^2.$$

Let us put $\sigma_{\min} = \min_{y\in S} \|Ay\| \geq 0$, where the minimum is attained since we minimize a continuous function ofer a compact set.

If $\sigma_{\min} > 0$ we can estimate our objective function from below as $\|b\|^2 - 2\lambda b^{\mathrm{T}}As + \lambda^2\|As\|^2 \geq \|b\|^2 - 2\lambda\|b\|\|A\| + \lambda^2\sigma^2_{\min}$, where we have used the fact that $\|y\| = 1$. The function on the right hand side of the inequality goes to infinity when $\lambda \to \infty$, meaning that $\lim_{\|x\|\to\infty} f(x) = +\infty$. Therefore in this case the function is coercive and continuous and as such admits a global minimum.

If $\sigma_{\min} = 0$ it means that for some $y_1 \in S : Ay_1 = 0$. Let us decompose $\mathbb{R}^{k+1}$ into $L_1 = \{\, x = \alpha y_1 \mid \alpha \in \mathbb{R} \,\}$, a one-dimensional space, and its $k$-dimensional orthogonal complement $L_k = L_1^{\perp}$. Then for each $x \in \mathbb{R}^{k+1}$ we can uniquely write $x = x_1 + x_k$, where $x_1 \in L_1$, $x_k \in L_k$. Furthermore, $f(x) = f(x_1 + x_k) = \|Ax_k + Ax_1 - b\|^2 = \|Ax_k - b\|^2 = f(x_k)$, and as a result

$$\min_{x\in\mathbb{R}^{k+1}} f(x) = \min_{x_k\in L_k} f(x_k),$$

which is a $k$-dimensional optimization problem of the same type (with any choice of the basis in $L_k$) and therefore admits a solution by the induction hypothesis.

**c)** Show that the solution $x^*$ of (1) is unique, if the rank of $A$ equals $n$.

**Solution:** If $\operatorname{rank} A = n$ it means that the columns of $A$ are linearly independent, and therefore the homogeneous problem $Av = 0$ admits only a trivial solution. As a result, the Hessian of our objective function is positive definite; indeed

$$v^{\mathrm{T}}\nabla^2 f(x)v = v^{\mathrm{T}}A^{\mathrm{T}}Av = \|Av\|^2 \geq 0$$

with equality only when $v = 0$. Concequently the function is strictly convex, and the global minimum is unique.

**d)** Show that, regardless of the rank of $A$, the optimization problem

$$\min_{x\in\mathbb{R}^n} \|x\|^2 \qquad \text{s.t. } x \text{ solves (1)} \tag{2}$$

admits a unique solution $x^\dagger \in \mathbb{R}^n$.

**Solution:** We have already shown that the function $f$ is convex, and that its set of global minimizers is non-empty regardless of $A$. Owing to the convexity of $f$, its set of global minimizers is also a convex set; let us call it $\Omega$ — these are precisely the points satisfying (1). Clearly $\Omega$ is closed (this is true for any l.s.c. function $f$). Therefore, a continuous and coercive function $g(x) = \|x\|^2$ admits at least one minimizer on $\Omega$. Further, since $g$ is strictly convex ($\nabla^2 g = 2I$), there cannot be more than one minimizer in $\Omega$ (otherwise their convex combination would be an even better solution in $\Omega$).

3 Assume that $A \in \mathbb{R}^{n\times n}$ is symmetric and positive *semi*-definite, $b \in \operatorname{ran} A$ (equivalently, $b \perp \ker A$, or equivalently there exists a solution to the system $Ax = b$). Show that, in exact arithmetics, the CG algorithm converges in at most $m = \dim \operatorname{ran} A$ iterations to a solution to the system $Ax = b$ from any starting point $x_0 \in \mathbb{R}^n$.

Thus the requirement for $A$ to be positive definite can be somewhat relaxed, and the algorithm still works.

**Solution:** The main difficulty is in showing that the algorithm does not break down with divisions by zero when the steplength $\alpha_k$ is computed, as there are could be directions $p \neq 0$ such that $p^\mathrm{T} A p = 0$. For this to be the case, however, the direction $p$ needs to be in $\ker A$: indeed, if we expand $p = \sum_i c_i v_i$ in terms of orthonormal eigenvectors $v_i$ of $A$, which correspond to eigenvalues $\lambda_i \geq 0$, then $p^\mathrm{T} A p = \sum_i \lambda_i c_i^2$. For the latter sum to be zero $p$ must be a linear combination of eigenvectors, corresponding to the zero eigenvalue.

We will first show that throughout the usual CG algorithm we maintain $p_k \in \operatorname{ran} A$ so that divisions by zero are avoided. We will then show the estimate on the number of iterations.

At iteration 0 we have $p_0 = r_0 = b - Ax_0 \in \operatorname{ran}(A) - \operatorname{ran}(A) \in \operatorname{ran} A$. Assuming that $p_k \in \operatorname{ran}(A)$, we compute $p_{k+1} = r_{k+1} + \beta_{k+1} p_k \in \operatorname{ran}(A) - \beta_{k+1} \operatorname{ran}(A) \in \operatorname{ran}(A)$, because $r_{k+1} = b - Ax_{k+1} \in \operatorname{ran}(A) - \operatorname{ran}(A) \in \operatorname{ran} A$.

The usual inductive proof of convergence of CG implies that the algorithm constructs orthogonal residuals $\{r_0, r_1, \dots\}$ and conjugate directions $\{p_0, p_1, \dots\}$. Normally we rely on the fact that the number of conjugate or orthogonal directions in the $n$-dimensional space is $n$, therefore the algorithm must converge in at most $n$ steps. Hovever, all residuals are by construction in $\operatorname{ran} A$, which in the present case has dimension $m \leq n$. Thus the algorithm will generate a zero residual (in exact arithmetics) after at most $m$ steps.

4 Assume that $m > n$, that $A \in \mathbb{R}^{m\times n}$, and that $b \in \mathbb{R}^m$. Consider the following algorithm:

- Choose $x_0 \in \mathbb{R}^n$ arbitrary, set $r_0 \leftarrow Ax_0 - b$, $s_0 \leftarrow A^\mathrm{T} r_0$, $p_0 \leftarrow -s_0$, and $k \leftarrow 0$.

- While $s_k \neq 0$:

$$\begin{aligned} \alpha_k &\leftarrow \frac{\|s_k\|^2}{\|Ap_k\|^2}, \\ x_{k+1} &\leftarrow x_k + \alpha_k p_k, \\ r_{k+1} &\leftarrow r_k + \alpha_k A p_k, \\ s_{k+1} &\leftarrow A^{\mathrm{T}} r_{k+1}, \\ \beta_{k+1} &\leftarrow \frac{\|s_{k+1}\|^2}{\|s_k\|^2}, \\ p_{k+1} &\leftarrow -s_{k+1} + \beta_{k+1} p_k, \\ k &\leftarrow k+1. \end{aligned}$$

Assume that the matrix $A$ has full rank. Show that the algorithm above is actually identical with the CG-algorithm for the solution of $A^{\mathrm{T}}Ax = A^{\mathrm{T}}b$ (in the sense that the iterates $x_k$ of both methods coincide).

**Solution:** We provide an inductive argument, showing that

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

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 $A^{\mathrm{T}}A$ is symmetric positive definite (rank $A = n$).

Base case $k = 1$ follows from

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

so that

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

and

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

Therefore

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

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

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

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

and

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

so, most importantly,

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

**5** Exercise 5.1 in Nocedal & Wright.

(Note that in MATLAB the Hilbert matrix can be produced with the command `hilb`, and in Python using `scipy.linalg.hilbert`.)

**Solution:** See possible solutions on the wiki.

**6** Exercise 5.12 in Nocedal & Wright: show that Lemma 5.6 holds for any choice of $\beta_k$ in the non-linear CG algorithm with $|\beta_k| \leq \|\beta_k^{\mathrm{FR}}$. In particular, this explains the strategy (5.48) in the book (FR–PR CG algorithm).

**Solution:** Induction/direct computation as in the proof of Lemma 5.6, utilizing the strong curvature condition.