MS-E2122 - Nonlinear optimization **Exercise sheet 6**
Prof. Fabricio Oliveira **Thursday 21.10.2021**

Project Work 1 is online from Monday 11.10.2021 https://mycourses.aalto.fi/mod/assign/view.php?id=765881 and is due no later than **Friday 5.11.2021 23:55**. No homework this week.

Problem 6.1: Convergence of Gradient Methods

Consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(x) = \frac{1}{2}x^\top Qx \tag{1}$$

where $Q \in \mathbb{R}^{n\times n}$ is a positive definite symmetric matrix. Suppose that $f(x)$ is minimized with a Gradient method using the update rule

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \tag{2}$$

and exact line search where the stepsize α_k at iteration k is computed as the minimum α of

$$\theta(\alpha) = f(x_k - \alpha\nabla f(x_k)) \tag{3}$$

Let $\underline{\lambda}$ and $\overline{\lambda}$ be the minimum and maximum eigenvalues of the (Hessian) matrix Q, respectively. Show that for all iterations k, we have

$$f(x_{k+1}) \leq \left(\frac{\overline{\lambda}-\underline{\lambda}}{\overline{\lambda}+\underline{\lambda}}\right)^2 f(x_k) \quad \text{or} \quad \frac{f(x_{k+1})}{f(x_k)} \leq \left(\frac{\overline{\lambda}-\underline{\lambda}}{\overline{\lambda}+\underline{\lambda}}\right)^2 \tag{4}$$

Solution.

To simplify notation, let us denote the gradient of f at x_k as

$$g_k = \nabla f(x_k) = \nabla(\frac{1}{2}x_k^\top Qx_k) = Qx_k \tag{5}$$

We can see that (4) holds if $g_k = 0$, so assume that $g_k \neq 0$.

Let us first compute the optimal stepsize α_k from (3). By taking the derivate of (3) and setting it to zero, we get

$$\begin{aligned}\theta'(\alpha) &= -g_k^\top \nabla f(x_k - \alpha g_k)\\ &= -g_k^\top Q(x_k - \alpha g_k)\\ &= -g_k^\top Qx_k + \alpha g_k^\top Qg_k\\ &= -g_k^\top g_k + \alpha g_k^\top Qg_k = 0\end{aligned}$$

and solving for α, we get the optimal stepsize at iteration k as

$$\alpha_k = \frac{g_k^\top g_k}{g_k^\top Qg_k} \tag{6}$$

Now, from (1) and (2) we have

$$\begin{aligned}f(x_{k+1}) &= \frac{1}{2}(x_k - \alpha_k g_k)^\top Q(x_k - \alpha_k g_k)\\ &= \frac{1}{2}(x_k^\top - \alpha_k g_k^\top)Q(x_k - \alpha_k g_k)\\ &= \frac{1}{2}(x_k^\top Qx_k - \alpha_k x_k^\top Qg_k - \alpha_k g_k^\top Qx_k + \alpha_k^2 g_k^\top Qg_k)\\ &= \frac{1}{2}(x_k^\top Qx_k - 2\alpha_k g_k^\top Qx_k + \alpha_k^2 g_k^\top Qg_k)\\ &= \frac{1}{2}(x_k^\top Qx_k - 2\alpha_k g_k^\top g_k + \alpha_k^2 g_k^\top Qg_k) \qquad (7)\end{aligned}$$

and substituting the optimal stepsize (6) to (7) we get

$$\begin{aligned} f(x_{k+1}) &= \frac{1}{2}\left(x_k^\top Qx_k - 2\frac{(g_k^\top g_k)^2}{g_k^\top Qg_k} + \frac{(g_k^\top g_k)^2}{(g_k^\top Qg_k)^2}g_k^\top Qg_k\right) \\ &= \frac{1}{2}\left(x_k^\top Qx_k - \frac{(g_k^\top g_k)^2}{g_k^\top Qg_k}\right) \end{aligned} \tag{8}$$

Now, by writing

$$\begin{aligned} f(x_k) &= \frac{1}{2}x_k^\top Qx_k \\ &= \frac{1}{2}x_k^\top QQ^{-1}Qx_k \\ &= \frac{1}{2}(Qx_k)^\top Q^{-1}(Qx_k) \\ &= \frac{1}{2}g_k^\top Q^{-1}g_k \end{aligned} \tag{9}$$

and substituting (9) to (8), we get

$$\begin{aligned} f(x_{k+1}) &= f(x_k) - \frac{1}{2}\frac{(g_k^\top g_k)^2}{g_k^\top Qg_k} \\ &= f(x_k) - \frac{(g_k^\top g_k)^2\frac{1}{2}g_k^\top Q^{-1}g_k}{(g_k^\top Qg_k)(g_k^\top Q^{-1}g_k)} \\ &= f(x_k) - \frac{(g_k^\top g_k)^2 f(x)}{(g_k^\top Qg_k)(g_k^\top Q^{-1}g_k)} \\ &= \left(1 - \frac{(g_k^\top g_k)^2}{(g_k^\top Qg_k)(g_k^\top Q^{-1}g_k)}\right)f(x_k) \end{aligned} \tag{10}$$

To proceed with the proof, we need the following *Kantorovich inequality.* Let $Q \in \mathbb{R}^{n\times n}$ be a positive definite symmetric matrix. Then, for any vector $y \in \mathbb{R}^n$ with $y \neq 0$, we have

$$\frac{(y^\top y)^2}{(y^\top Qy)(y^\top Q^{-1}y)} \geq \frac{4\overline{\lambda}\,\underline{\lambda}}{(\overline{\lambda}+\underline{\lambda})^2} \tag{11}$$

where $\underline{\lambda}$ and $\overline{\lambda}$ are the minimum and maximum eigenvalues of Q, respectively. Now, by applying the Kantorovich inequality (11) to (10), we get

$$\begin{aligned} f(x_{k+1}) &\leq \left(1 - \frac{4\overline{\lambda}\,\underline{\lambda}}{(\overline{\lambda}+\underline{\lambda})^2}\right)f(x_k) \\ &= \left(\frac{(\overline{\lambda}+\underline{\lambda})^2 - 4\overline{\lambda}\,\underline{\lambda}}{(\overline{\lambda}+\underline{\lambda})^2}\right)f(x_k) \\ &= \left(\frac{\overline{\lambda}^2 + 2\overline{\lambda}\,\underline{\lambda} + \underline{\lambda}^2 - 4\overline{\lambda}\,\underline{\lambda}}{(\overline{\lambda}+\underline{\lambda})^2}\right)f(x_k) \\ &= \left(\frac{\overline{\lambda}^2 - 2\overline{\lambda}\,\underline{\lambda} + \underline{\lambda}^2}{(\overline{\lambda}+\underline{\lambda})^2}\right)f(x_k) \\ &= \left(\frac{(\overline{\lambda}-\underline{\lambda})^2}{(\overline{\lambda}+\underline{\lambda})^2}\right)f(x_k) \end{aligned} \tag{12}$$

Dividing both sides by $f(x_k)$ ($f(x_k) > 0$ since Q is PD), we finally get

$$\frac{f(x_{k+1})}{f(x_k)} \leq \left(\frac{\overline{\lambda}-\underline{\lambda}}{\overline{\lambda}+\underline{\lambda}}\right)^2 \tag{13}$$

Notice that if we denote the condition number of the (Hessian) matrix Q as

$$\kappa = \frac{\overline{\lambda}}{\underline{\lambda}}$$

we can rewrite (13) as

$$\frac{f(x_{k+1})}{f(x_k)} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \tag{14}$$

from which we can clearly see how the convergence rate of the Gradient method depends on the condition number κ of the corresponding (Hessian) matrix Q. For large condition numbers κ, the right side of (14) evaluates close to 1, which implies poor convergence rate.

Problem 6.2: Effect of Scaling on Gradient Method Convergence

Consider the following unconstrained optimization problem

$$\min_{x}. \; f(x) = (x_1 - 2)^2 + 5(x_2 + 6)^2 \tag{15}$$

where we denote the (quadratic) objective function $f : \mathbb{R}^2 \to \mathbb{R}$ of (15) as

$$f(x) = (x_1 - 2)^2 + 5(x_2 + 6)^2 \tag{16}$$

Suppose that we want to solve the problem (15) with a Gradient method using the update rule

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \tag{17}$$

and exact line search where the stepsize α_k at iteration k is computed as the minimum α of

$$\theta(\alpha) = f(x_k - \alpha \nabla f(x_k)) \tag{18}$$

(a) Evaluate the convergence rate of the Gradient method applied to problem (15) with an arbitrary starting point. *Hint:* Use the the results of Exercise 6.1.

(b) Can you solve the problem faster by first modifying the objective function (16) and then applying the Gradient method to the modified problem? *Hint:* Try to find a variable substitution that gives the best convergence rate according to the results of Exercise 6.1.

Solution.

(a) We will derive an upper bound for the convergence rate of the Gradient method applied to (15) based on the result of Exercise 6.1. To this end, notice that we have similar conditions: the objective function is quadratic, although we have an additional affine term. Nevertheless, we can write $f(x)$ in the form

$$\begin{aligned} f(x) &= \frac{1}{2} x^\top Q x + c^\top x + b \\ &= \frac{1}{2}[x_1 \; x_2] \begin{bmatrix} 2 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + [-4 \; 60] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 184 \end{aligned}$$

and we use the exact line search just like in 6.1. Notice that Q corresponds to the Hessian of $f(x)$ as in Exercise 6.1. To verify this, computing the gradient and the Hessian of $f(x)$ yields

$$\nabla f(x) = \begin{pmatrix} 2x_1 - 4 \\ 10x_2 + 60 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(x) = \begin{pmatrix} 2 & 0 \\ 0 & 10 \end{pmatrix} = Q \tag{19}$$

and we can compute the eigenvalues of the Hessian from the eigenvalue equation

$$(\nabla^2 f(x) - \lambda I)v = 0$$

which has a solution if and only if

$$\mathbf{det}(\nabla^2 f(x) - \lambda I) = 0 \quad \Leftrightarrow \quad (2-\lambda)(10-\lambda) = 0$$

and we get $\overline{\lambda} = 10$ and $\underline{\lambda} = 2$. Thus, by substituting these values to (13), we get

$$\frac{f(x_{k+1})}{f(x_k)} \leq \left(\frac{\overline{\lambda} - \underline{\lambda}}{\overline{\lambda} + \underline{\lambda}}\right)^2 = \left(\frac{10-2}{10+2}\right)^2 = \frac{4}{9}$$

For example, it takes about 22 iterations to converge to optimum $(x_1, x_2) = (2, -6)$ starting from the point $x_0 = (4, -5)$. The progress is shown in Figure 1

(b) If we perform a change of variables $y_1 = (x_1 - 2)$ and $y_2 = \sqrt{5}(x_2 + 6)$, we can write the objective function as

$$f(y) = y_1^2 + y_2^2$$

By computing the gradient and the Hessian of $f(y)$, we get

$$\nabla f(y) = \begin{pmatrix} 2y_1 \\ 2y_2 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \tag{20}$$

and we can compute the eigenvalues of the Hessian from the eigenvalue equation

$$(\nabla^2 f(y) - \lambda I)v = 0$$

which has a solution if and only if

$$\mathbf{det}(\nabla^2 f(y) - \lambda I) = 0 \quad \Leftrightarrow \quad (2-\lambda)(2-\lambda) = 0$$

and we get $\overline{\lambda} = 2$ and $\underline{\lambda} = 2$. Thus, by substituting these values to (13), we get

$$\frac{f(y_{k+1})}{f(y_k)} \leq \left(\frac{\overline{\lambda} - \underline{\lambda}}{\overline{\lambda} + \underline{\lambda}}\right)^2 = \left(\frac{2-2}{2+2}\right)^2 = 0$$

Thus, the Gradient method will converge to the optimum in one iteration, regardless of the starting point as long as we use the exact line search. Thus, we can solve the modified problem using the Gradient method in one iteration to get the optimal solution $(y_1, y_2) = (0, 0)$, from which we can compute the optimal solution to the original problem by simple substitution: $x_1 = y_1 + 2 = 2$ and $x_2 = y_2/\sqrt{5} - 6 = -6$. The convergence plot is shown in Figure 2

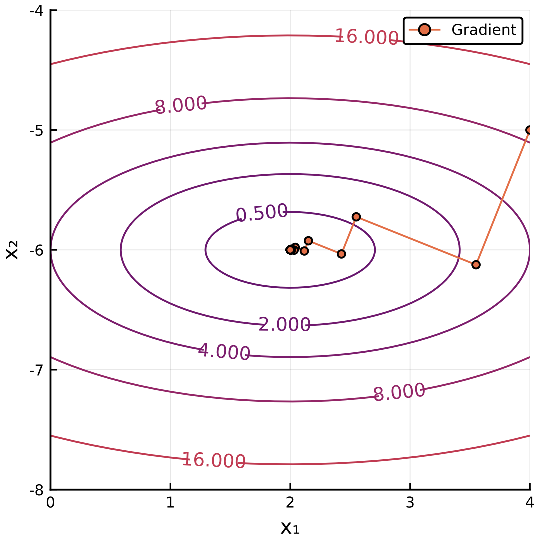

Figure 1: Convergence of Gradient method in part (a)

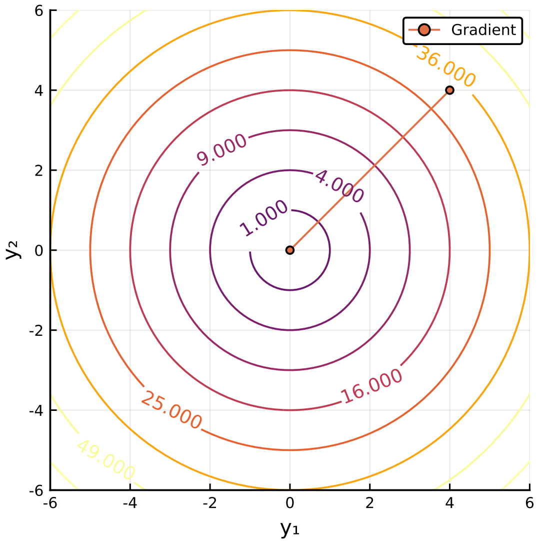

Figure 2: Convergence of Gradient method in part (b)