MATH-329 Nonlinear optimization
Exercise session 5: Trust-region with Cauchy step

Instructor: Nicolas Boumal
TAs: Andrew McRae, Andreea Musat

Document compiled on October 2, 2024

Exercises marked with $(*)$ will be used in later exercises or in the homeworks: you might want to prioritize those.

1. The Cauchy step. Consider the quadratic model for $f\colon \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ given by

$$m(v) = f(x) + \langle \nabla f(x), v\rangle + \frac{1}{2}\langle v, Hv\rangle$$

for some symmetric matrix $H \in \mathbb{R}^{n\times n}$. We let $\Delta > 0$ be a radius and we let $g = \nabla f(x)$ denote the gradient.

1. Remember that the Cauchy step is defined by

$$u^C = -t^C \cdot g \qquad \text{with} \qquad t^C \in \operatorname*{arg\,min}_{0\leq t\leq \frac{\Delta}{\|g\|}} m(-t\cdot g).$$

   Show that the step size is given by

$$t^C = \begin{cases} \min\left(\frac{\|g\|^2}{\langle g, Hg\rangle}, \frac{\Delta}{\|g\|}\right) & \text{if } \langle g, Hg\rangle > 0, \\ \frac{\Delta}{\|g\|} & \text{otherwise.} \end{cases}$$

2. Show that the Cauchy step leads to the following decrease in model value

$$m(0) - m(u^C) \geq \frac{1}{2}\min\left(\Delta, \frac{\|g\|}{\|H\|}\right)\|g\|.$$

Answer.

1. Finding the Cauchy point amounts to minimizing the 1-dimensional quadratic function

$$\begin{aligned} r(t) &= m(-t\cdot g) \\ &= f(x) - t\|g\|^2 + \frac{t^2}{2}\langle g, Hg\rangle. \end{aligned}$$

   on the closed interval $[0, \frac{\Delta}{\|g\|}]$. If $\langle g, Hg\rangle \leq 0$, then r is non-increasing on $\mathbb{R}_+$. So the minimum is attained at the upper boundary of the interval, that is, $t = \frac{\Delta}{\|g\|}$.

   Suppose now that $\langle g, Hg\rangle > 0$. Then r is convex and has a unique critical point:

$$r'(t^\star) = 0 \qquad \Leftrightarrow \qquad t^\star = \frac{\|g\|^2}{\langle g, Hg\rangle} > 0.$$

This is the global minimum of r. If this critical point is in $[0, \frac{\Delta}{\|g\|}]$ then $t^C = t^\star$. Otherwise $t^\star > \frac{\Delta}{\|g\|}$ and the convexity of r implies that r is non-increasing on $[0, \frac{\Delta}{\|g\|}]$. So we deduce that $t^C = \frac{\Delta}{\|g\|}$. We conclude that

$$t^C = \min\left(\frac{\|g\|^2}{\langle g, Hg\rangle}, \frac{\Delta}{\|g\|}\right).$$

2. Assume first that $\langle g, Hg\rangle \leq 0$. Then we have

$$\begin{aligned} m(u^C) - m(0) &= f(x) - \frac{\Delta}{\|g\|}\|g\|^2 + \frac{\Delta^2}{2\|g\|^2}\langle g, Hg\rangle - f(x) \\ &= -\Delta\|g\| + \frac{\Delta^2}{2\|g\|^2}\langle g, Hg\rangle \\ &\leq -\Delta\|g\|, \end{aligned}$$

where the inequality follows from the assumption $\langle g, Hg\rangle \leq 0$. Therefore:

$$m(0) - m(u^C) \geq \Delta\|g\| \geq \frac{1}{2}\Delta\|g\| \geq \frac{1}{2}\min\left(\Delta, \frac{\|g\|}{\|H\|}\right)\|g\|$$

Now assume that $\langle g, Hg\rangle \geq 0$ and that $\frac{\|g\|^2}{\langle g, Hg\rangle} < \frac{\Delta}{\|g\|}$, implying that $t^C = \frac{\|g\|^2}{\langle g, Hg\rangle}$. Then:

$$\begin{aligned} m(u^C) - m(0) &= -\frac{\|g\|^2}{\langle g, Hg\rangle}\|g\|^2 + \frac{1}{2}\frac{\|g\|^4}{\langle g, Hg\rangle^2}\langle g, Hg\rangle \\ &= -\frac{1}{2}\frac{\|g\|^4}{\langle g, Hg\rangle} \\ &\leq -\frac{1}{2}\frac{\|g\|^2}{\|H\|}, \end{aligned}$$

where the inequality follows from the fact that $\langle g, Hg\rangle \leq \|H\|\|g\|^2$ implying that $-\frac{1}{\langle g, Hg\rangle} \leq -\frac{1}{\|H\|\|g\|^2}$. Therefore:

$$m(0) - m(u^C) \geq \frac{1}{2}\frac{\|g\|^2}{\|H\|} \geq \frac{1}{2}\min\left(\Delta, \frac{\|g\|}{\|H\|}\right)\|g\|.$$

Finally, assume that $\langle g, Hg\rangle \geq 0$ and $\frac{\|g\|^2}{\langle g, Hg\rangle} \geq \frac{\Delta}{\|g\|}$, implying that $t^C = \frac{\Delta}{\|g\|}$ and $\langle g, Hg\rangle \leq \frac{\|g\|^3}{\Delta}$. Then:

$$\begin{aligned} m(u^C) - m(0) &= -\Delta\|g\| + \frac{\Delta^2}{2\|g\|^2}\langle g, Hg\rangle \\ &\leq -\Delta\|g\| + \frac{\Delta^2}{2\|g\|^2}\frac{\|g\|^3}{\Delta} \\ &= -\frac{1}{2}\Delta\|g\|. \end{aligned}$$

So finally, as before, we can say:

$$m(0) - m(u^C) \geq \Delta\|g\| \geq \frac{1}{2}\Delta\|g\| \geq \frac{1}{2}\min\left(\Delta, \frac{\|g\|}{\|H\|}\right)\|g\|.$$

■

2. $(*)$ **Implementing the TR algorithm.**

1. Implement the trust-region algorithm (see lecture notes). Use $H_k = \nabla^2 f(x_k)$ for the quadratic model and the Cauchy step to approximately solve the trust-region subproblem.
2. Consider the n-dimensional Rosenbrock function (see the definition at the end of the exercise sheet). We provide files on Moodle to compute the function value, the gradient and the Hessian. Run your implementation of the TR algorithm with $n = 10$ and $x_0 =$ `randn(n,1)`. You may choose parameters $\bar{\Delta} = \sqrt{n}$, $\Delta_0 = \bar{\Delta}/8$ and $\rho' = 0.1$.
3. Compare the performance with the line-search gradient descent algorithm. Both algorithms should have approximately the same convergence speed; can you guess why? Can you see pros and cons for TR with Cauchy steps rather than line-search gradient descent? Soon we will see how to exploit second-order information better to greatly improve the convergence speed.

Answer.

1. See lecture notes for pseudocode.
2. The TR algorithm with Cauchy point needs around 2.5×10^4 iterations to reach a point where the gradient norm is below 10^{-6} (see Figure 1). Due to random initialization, sometimes this critical point is the local minimum $(-1, 1, \ldots, 1)^\top$ and sometimes the global minimum $(1, \ldots, 1)^\top$.
3. With great disappointment, we find that gradient descent with line-search performs just as good, both in terms of iterations and computation time. The convergence is linear in both cases (see Figures 1 and 2). This is because the Cauchy point is only a rough estimate of subproblem solution. With a single Hessian call we are using very little second order information, especially when n is large.

   There is some good news: trust-region appears to be a globally convergent method, just as gradient descent. Both gradient descent with line-search and trust-region with Cauchy point try to find an adaptive step size. In the case of Cauchy point we use second order information to choose the step. The trust-region mechanism also has a memory of the previous steps in the sense that the radius is large if the model is a good local approximation of the function. In contrast, the line-search does not need the Hessian.

   What is left is to figure out a way to exploit more efficiently the second order information to make TR method also locally quadratically convergent. To do so we will discuss the truncated conjugate gradient method in class.

■

Multidimensional Rosenbrock function. We generalize the Rosenbrock function in n dimensions as

$$f(x) = \sum_{i=1}^{n-1} \left[100 \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2 \right].$$

The vector of ones is the unique global minimum (because the function is non-negative and is zero if and only if all entries are ones). The gradient at $x \in \mathbb{R}^n$ is given by

$$\nabla f(x)_i = \begin{cases} -2(1 - x_1) - 400 x_1 (x_2 - x_1^2) & \text{if } i = 1 \\ 200(x_i - x_{i-1}^2) - 2(1 - x_i) - 400 x_i (x_{i+1} - x_i^2) & \text{if } 1 < i < n \\ 200(x_n - x_{n-1}^2) & \text{if } i = n. \end{cases}$$

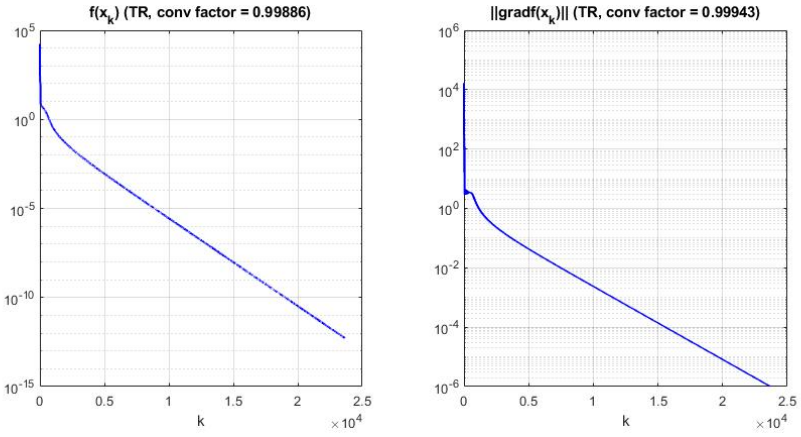

Figure 1: Objective value and gradient norm throughout the iterations of TR with Cauchy point on the multidimensional Rosenbrock function ($n = 10$, $x_0 =$ `randn(n,1)`).

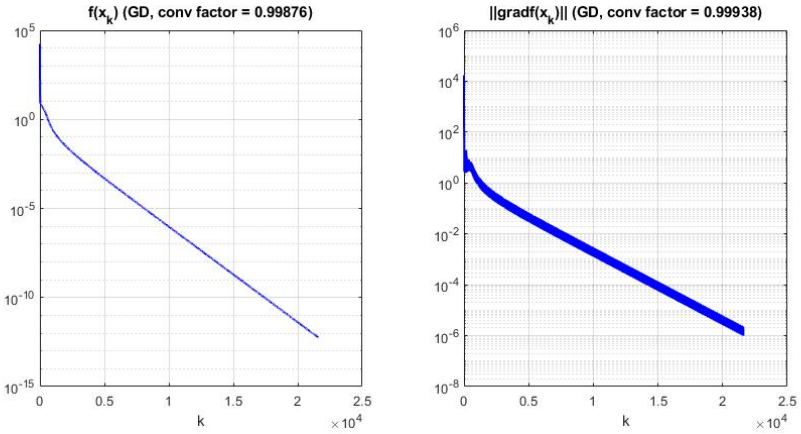

Figure 2: Objective value and gradient norm throughout the iterations of GD with backtracking line-search on the multidimensional Rosenbrock function ($n = 10$, $x_0 =$ `randn(n,1)`).

The Hessian at x is a symmetric tridiagonal $n \times n$ matrix. The main diagonal and the first diagonal above are given by

$$\begin{bmatrix} 2 + 1200x_1^2 - 400x_2 \\ 202 + 1200x_2^2 - 400x_3 \\ \vdots \\ 202 + 1200x_{n-1}^2 - 400x_n \\ 200 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -400x_1 \\ \vdots \\ -400x_{n-1} \end{bmatrix}$$

respectively. In practice we never build the full matrix but solely compute matrix/vector products. This can be done efficiently because the matrix is sparse.