E. Herberg, M. Marić, V. Stein
Heidelberg University

Exercise 6 - Solution

Date issued: 21st May 2024
Date due: 28th May 2024

Homework Problem 6.1 (Example for convergence of the local Newton's method) 6 Points

Let $p > 2$ and $f : \mathbb{R} \to \mathbb{R}, f(x) \coloneqq |x|^p$ be given. Consider the local Newton's method (Algorithm 4.23) for minimization of f, i.e. $F(x) = \nabla f(x)$, with some initial guess $x^{(0)} > 0$.

(i) Show that the method converges to the global minimizer $x^* = 0$ of f.

(ii) Which rate of convergence do you observe?

(iii) Why is this result not in contradiction with Theorem 4.27?

Solution.

We have $f \in C^2(\mathbb{R};\mathbb{R})$ with

$$\begin{aligned} f(x) &= |x|^p = (\operatorname{sgn}(x) \cdot x)^p, \\ f'(x) &= p(\operatorname{sgn}(x) \cdot x)^{p-1} \operatorname{sgn}(x) = p|x|^{p-1} \operatorname{sgn}(x), \\ f''(x) &= p(p-1)(\operatorname{sgn}(x) \cdot x)^{p-2} (\operatorname{sgn}(x))^2 = p(p-1)|x|^{p-2}. \end{aligned}$$

Note that in one dimension $f'(x) = \nabla f(x) \in \mathbb{R}$. (1 Point)

(i) For $x^{(k)} \neq 0$ we have by definition of the Newton's direction

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - (f''(x^{(k)}))^{-1} f'(x^{(k)}) \\ &= x^{(k)} - \frac{p|x^{(k)}|^{p-1}\operatorname{sgn}(x^{(k)})}{p(p-1)|x^{(k)}|^{p-2}} \\ &= x^{(k)} - \frac{|x^{(k)}|\operatorname{sgn}(x^{(k)})}{(p-1)} \\ &= \left(1 - \frac{1}{p-1}\right) x^{(k)}. \end{aligned}$$

(1 Point)

Next, we prove by induction that indeed $x^{(k)} > 0 \ \forall k$:
For $k = 0$ the claim holds by assumption. Let $x^{(k)} > 0$ for some $k \in \mathbb{N}_0$, then

$$x^{(k+1)} = \underbrace{\left(1 - \frac{1}{p-1}\right)}_{>0, \text{ since } p>2} \underbrace{x^{(k)}}_{>0} > 0.$$

(1 Point)

Together we see that $(x^{(k)})$ converges to $x^* = 0$ (from above):

$$x^{(k)} = \underbrace{\left(1 - \frac{1}{p-1}\right)}_{<1} x^{(k-1)} = \ldots = \left(1 - \frac{1}{p-1}\right)^k x^{(0)} \xrightarrow{k\to\infty} 0.$$

(1 Point)

(ii) To determine the convergence rate, we observe

$$|x^{(k+1)} - x^*| = \left|\left(1 - \frac{1}{p-1}\right) x^{(k)} - \underbrace{x^*}_{=0}\right| = \left|\left(1 - \frac{1}{p-1}\right)(x^{(k)} - x^*)\right| = \left(1 - \frac{1}{p-1}\right)|x^{(k)} - x^*|.$$

This shows Q-linear convergence with factor $\left(1 - \frac{1}{p-1}\right) \in (0, 1)$. (1 Point)

(iii) Theorem 4.27 (iii) states the Q-superlinear convergence of the local Newton's method. Here, we only observe Q-linear convergence. However, this is not a contradiction, because the assumptions of Theorem 4.27 (iii) are not fulfilled. In detail, we have

$$f''(x^*) = f''(0) = p(p-1)|0|^{p-2} = 0,$$

which is not invertible. (1 Point)

Homework Problem 6.2 (On the Restriction $\sigma \in (0, \frac{1}{2})$ in Globalized Newton) 7 Points

In the globalized Newton's method for optimization (Algorithm 4.30 of the lecture notes), the Armijo-parameter, which is typically chosen as $\sigma \in (0, 1)$, is restricted to the interval $(0, \frac{1}{2})$ so that the full Newton step size $\alpha^{(k)} = 1$ can in fact be accepted by the Armijo condition for $k \geq k_0$ and some $k_0 > 0$, in order to facilitate quadratic convergence in the final stages of the algorithm. We will investigate why that is:

(i) Show that the step length $\alpha^{(k)} = 1$ satisfies the Armijo condition for the Newton direction $d^{(k)} \neq 0$ for the quadratic function

$$f(x) = \frac{1}{2}x^\intercal Ax + b^\intercal x + c$$

with s. p. d. $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, und $c \in \mathbb{R}$ if and only if $\sigma \leq \frac{1}{2}$.

(ii) Explain why we need to restrict ourselves to $\sigma < \frac{1}{2}$ for general nonquadratic problems.

Solution.

(i) The Armijo condition

$$f(x + t\, d) \leq f(x) + \sigma\, t\, f'(x)d.$$

holds at $t = 1$ if and only if

$$\begin{aligned}
& f(x + d) - f(x) \leq \sigma\, f'(x)d \\
\overset{\text{Taylor}}{\Leftrightarrow}\quad & \frac{1}{2}d^\intercal f''(\xi)d + f'(x)d \leq \sigma\, f'(x)d \\
\overset{\text{Rearrange}}{\Leftrightarrow}\quad & \frac{1}{2}d^\intercal f''(\xi)d \leq (\sigma - 1)\, f'(x)d \\
\overset{\text{form of } d}{\Leftrightarrow}\quad & \frac{1}{2}\nabla f(x)^\intercal f''(x)^{-1} f''(\xi) f''(x)^{-1}\nabla f(x) \leq (1 - \sigma)\, \nabla f(x)^\intercal f''(x)^{-1}\nabla f(x) \qquad (*)
\end{aligned}$$

where ξ is on the line connecting x and $x+d$ (Lagrangian form of error in Taylor's theorem). In the quadratic case, all second derivatives coincide with A, so that we can continue equivalently reformulating to

$$\begin{aligned} &\overset{f''\equiv A}{\Leftrightarrow} && \frac{1}{2}\nabla f(x)^\intercal A^{-1}AA^{-1}\nabla f(x) \le (1-\sigma)\,\nabla f(x)^\intercal A^{-1}\nabla f(x) \\ &\Leftrightarrow && 0 \le \left(\frac{1}{2}-\sigma\right)\underbrace{\nabla f(x)^\intercal A^{-1}\nabla f(x)}_{>0}. \end{aligned}$$

The direction $d \neq 0$ if and only if $\nabla f(x) \neq 0$, i.e., the Armijo condition holds for $t=1$ if and only if

$$\sigma \le \frac{1}{2}.$$

(4 Points)

(ii) Intuitively: The previous part showed that minimizing a quadratic functional and using the newton direction with $t=1$, we can only expect half the linearly predicted descent. For general nonlinear problems, we can argue as above until reaching the estimate $(*)$. Instead of $f''\equiv A$, we then have nonconstant terms. If the sequence of iterates converges and the hessians are "sufficiently well behaved", then we get almost quadratic behavior around the limit point, but with an additional error (for the higher order terms in Taylor's approximation). This error could potentially lead to nonacceptance of $t=1$, so the criterion needs to be a bit more lenient than in the quadratic case. (3 Points)

A bit more technical: Estimate $(*)$ can be obtained verbatim for any C^2 function as long as we are sufficiently close to a minimizer with nonsingular hessian. From that estimate, we can show that $\sigma < \frac{1}{2}$ is sufficient for the armijo condition holding for $t=1$.

When $x^{(k)}$ converges to x^* then $f'(x^{(k)})$ converges to 0 and with sufficiently uniformly regular hessians along the iterates, the directions $d^{(k)}$ will also converge to 0, meaning that the $\xi^{(k)}$ converge to $x^{(k)}$. The error

$$e^{(k)} := \frac{1}{2}\nabla f(x^{(k)})^\intercal f''(x^{(k)})^{-1}f''(\xi^{(k)})f''(x^{(k)})^{-1}\nabla f(x^{(k)}) - \frac{1}{2}\nabla f(x^{(k)})^\intercal f''(x^{(k)})^{-1}\nabla f(x^{(k)})$$

therefore converges to 0. For $\sigma < \frac{1}{2}$ we hence have a k_0, such that for $k \ge k_0$

$$\begin{aligned} \frac{1}{2}\nabla f(x^{(k)})^\intercal f''(x^{(k)})^{-1}f''(\xi^{(k)})f''(x^{(k)})^{-1}\nabla f(x^{(k)}) &= \frac{1}{2}\nabla f(x^{(k)})^\intercal f''(x^{(k)})^{-1}\nabla f(x^{(k)}) + e^{(k)} \\ &\le \underbrace{(1-\sigma)}_{>\frac{1}{2}}\ \underbrace{\nabla f(x^{(k)})^\intercal f''(x^{(k)})^{-1}\nabla f(x^{(k)})}_{>0 \text{ and converging to } 0} \end{aligned}$$

so the Armijo condition holds for $t = 1$ eventually. We have already seen in the preious part, that $\sigma < \frac{1}{2}$ is generally not necessary for the Armijo condition to hold for $t = 1$.

Homework Problem 6.3 (Characterization of fast local convergence) 6 Points

The proof of Lemma 4.36 is given in the lecture notes. Your task is to carefully read and understand the proof. Then write it down in your own words.

Solution.

See lecture notes. (6 Points)

Homework Problem 6.4 (Globalized Newton's Method in Optimization) 8 Points

Implement the globalized Newton's method for optimization (Algorithm 4.30 of the lecture notes), run it for the Rosenbrock's and/or Himmelblau's functions and compare its performance to that of your gradient descent implementation.

Solution.

For the implementation, see driver_ex_020_compare_newton_gradient_rosenbrock_himmelblau.py.

We obtain the behavior in Figures 0.1 and 0.2. Our convergence measure (difference of the function values and the optimal value, approximately corresponds to energy norm of error) shows the typical linear (SD) and quadratic (Newton) convergence modes. With absurdly large η and ρ, we can even force Newton's method to take a few gradient steps in the beginning but soon the Newton directions will be accepted and the step length $t = 1$ is in fact accepted towards the end of the method while the gradient steps are slowed down terribly by the step size control. In Himmelblau's function, steepest descent is not as significantly worse than globalized newton is compared to the Rosenbrock example, where the steepest descent method is struggling notably in the low angle curved valley towards the minimizer.

(8 Points)

Please submit your solutions as a single pdf and an archive of programs via moodle.

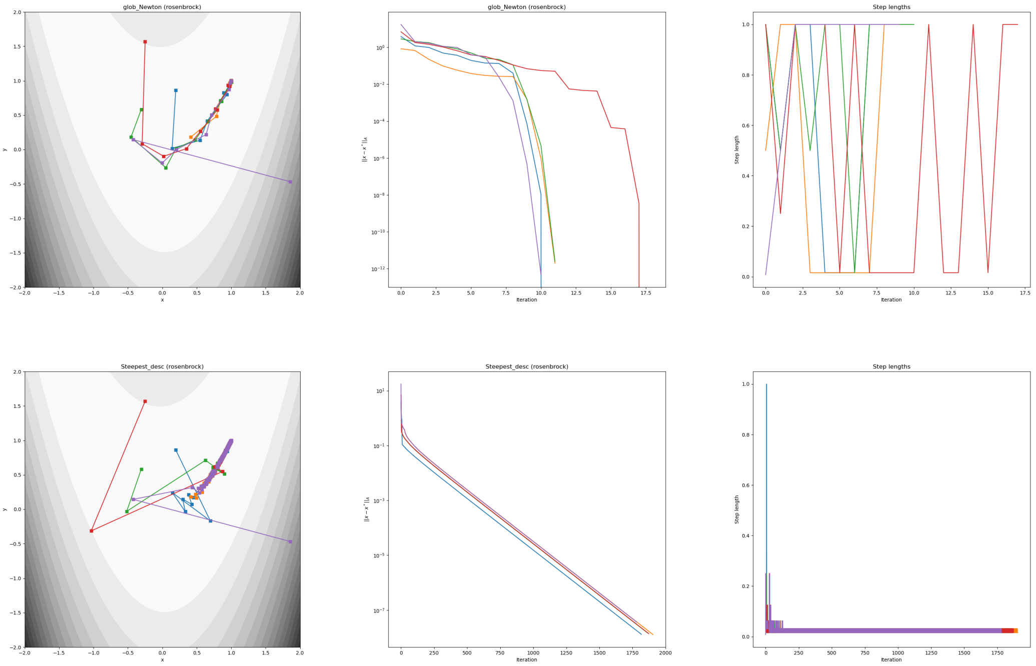

Figure 0.1: Newton (top) vs steepest descent (bottom) for Rosenbrock function.

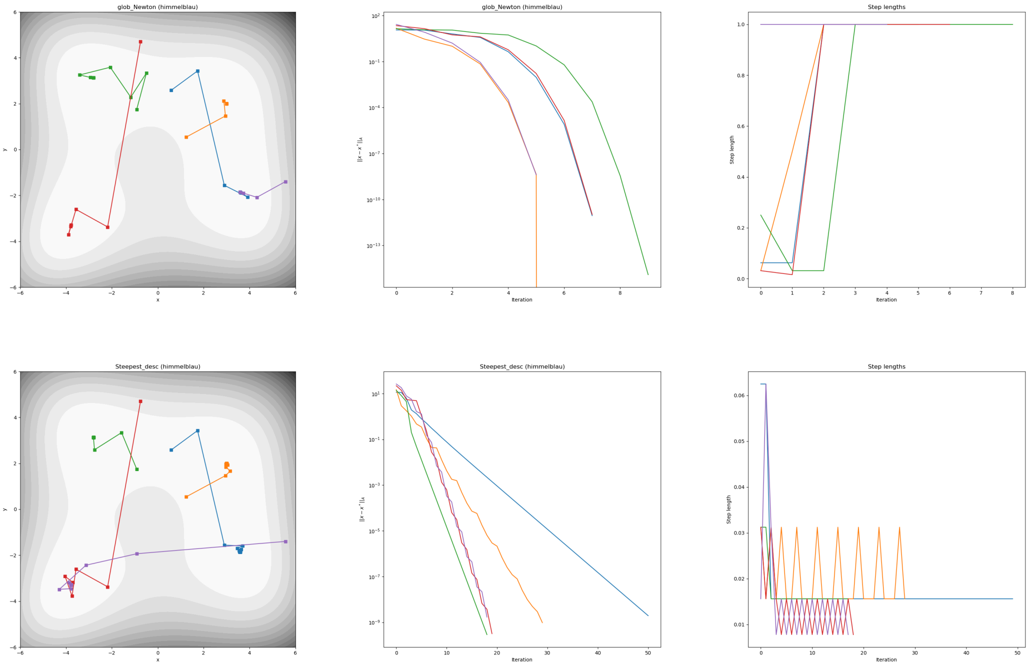

Figure 0.2: Newton (top) vs steepest descent (bottom) for Himmelblau function.