R. Herzog, G. Müller, E. Herberg, M. Hashemi
Heidelberg University

Exercise 5 - Solution

Date issued: 15th May 2023
Date due: 23rd May 2023

Homework Problem 5.1 (Affine Invariance of Newton's Method for Root Finding) 10 Points

Prove the statement in remark 5.29(*iii*) of the lecture notes concerning affine invariance of local Newton's method for solving the root finding problem $F(x) = 0$ with continuously differentiable $F\colon \mathbb{R}^n \to \mathbb{R}^n$ (Algorithm 5.23 of the lecture notes).

I. e., let $A \in \mathbb{R}^{n\times n}$ be regular and $b \in \mathbb{R}^n$ and consider a sequence $(x^{(k)})_{k\in\mathbb{N}_0}$ of iterates produced by Newton's method for F started from $x^{(0)} \in \mathbb{R}^n$. Prove that:

(*i*) Newton's method for the function

$$G\colon \mathbb{R}^n \mapsto \mathbb{R}^n, \quad G(y) := F(Ay + b)$$

with initial value $y^{(0)} \in \mathbb{R}^n$ such that $x^{(0)} = Ay^{(0)} + b$ is well defined and produces the sequence $(y^{(k)})_{k\in\mathbb{N}_0}$ of iterates with

$$x^{(k)} = Ay^{(k)} + b.$$

(*ii*) Newton's method for the function

$$H\colon \mathbb{R}^n \mapsto \mathbb{R}^n, \quad H(y) := AF(y)$$

with initial value $y^{(0)} \in \mathbb{R}^n$ such that $x^{(0)} = y^{(0)}$ is well defined and produces the sequence $(y^{(k)})_{k\in\mathbb{N}_0}$ of iterates with

$$x^{(k)} = y^{(k)}.$$

(*iii*) Explain why we can not expect a similar transformation result to hold for the iterates of Newton's method when we expand the transformation in Part (*ii*) by an additional constant shift, as in

$$H\colon \mathbb{R}^n \mapsto \mathbb{R}^n, \quad H(y) := AF(y) + b.$$

Solution.

(*i*) The claim is true for $k = 0$ by assumption. Now let Newton's method for G started at $y^{(0)}$ have been well defined and successfull up until the k-th iterate with

$$x^{(k)} = Ay^{(k)} + b.$$

Then we have that

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)}) \\ y^{(k+1)} &= y^{(k)} - G'(y^{(k)})^{-1}G(y^{(k)}) \end{aligned}$$

where

$$\begin{aligned} G(y) &:= F(Ay + b) \\ G'(y) &= F'(Ay + b)\, A. \end{aligned}$$

(2 Points)

Accordingly, $G'(y^{(k)})$ is singular if and only if $F'(x^{(k)})$ is singular, which it is not (by assumption), so the next iteration step is well defined as well, and we obtain that

$$\begin{aligned} Ay^{(k+1)} + b &= A\left(y^{(k)} - G'(y^{(k)})^{-1}G(y^{(k)})\right) + b \\ &= x^{(k)} - A\underbrace{G'(y^{(k)})^{-1}}_{A^{-1}F'(x^{(k)})^{-1}}\underbrace{G(y^{(k)})}_{F(x^{(k)})} \\ &= x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)}) \\ &= x^{(k+1)} \end{aligned}$$

(3 Points)

(*ii*) The claim is true for $k = 0$ by assumption. Now let Newton's method for H started at $y^{(0)}$ have been well defined and successfull up until the k-th iterate with

$$x^{(k)} = y^{(k)}.$$

Then we have that

$$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)})$$
$$y^{(k+1)} = y^{(k)} - H'(y^{(k)})^{-1}H(y^{(k)})$$

where

$$H(y) := AF(y)$$
$$H'(y) = AF'(y).$$

(2 Points)

Accordingly, $H'(y^{(k)})$ is singular if and only if $F'(x^{(k)})$ is singular, which it is not (by assumption), so the next iteration step is well defined as well, and we obtain that

$$\begin{aligned} y^{(k+1)} &= y^{(k)} - H'(y^{(k)})^{-1}H(y^{(k)}) \\ &= \underbrace{y^{(k)}}_{x^{(k)}} - F'(\underbrace{y^{(k)}}_{=x^{(k)}})^{-1}A^{-1}AF(\underbrace{y^{(k)}}_{=x^{(k)}}) \\ &= x^{(k+1)}. \end{aligned}$$

(2 Points)

(*iii*) As long as F is an affine linear function with nonsingular linear part, we can actually expect a similar transformation result to hold. When F is a fully nonlinear function though, then truely affine transformation in the image space can modify the location of the function's roots nonlinearly/non-affine. This of course influences the Newton steps because the directions will be influenced by the constant shift in the image space. Note that each update is affine-linearly dependent on the shift, but the entire sequence will depend on it nonlinearly.

Accordingly, we can expect to find examples of functions F and affine-linear transformations in the image space where there is no affine-linear connection between the iterates whatsoever. (1 Point)

The function $F(x) = e^x$ on $\mathbb{R}$ with vertical shift comes to mind. We can set $H(x) = e^x + b$ and examine the first three iterates for two different initial values and shifts b. We need three iterates each because for two, we can always find an affine linear transformation between the iterates. We omit further details here.

Note:

- The scaling (in-)variance property of Newton's and the steepest descent method in optimization can be nicely discussed when the cost functional is a quadratic function, where Newton always

converges in a single step while the steepest descent scheme's convergence depends on the scaling matrix A.

- Keep in mind that some of the analytical results, especially the size of the basin of attraction of a root, may be transformed by such transformations

Homework Problem 5.2 (Newton Fractals in Root Finding) 10 Points

The convergence of Newton's method for varying initial values can be quite chaotic, depending on the initial value. A nice visualization of its behavior are so called fractal plots for root finding problems, which color each starting point according to the root that the method converged to when started at that point. Fractal plots are typically created for Newton's method in the complex numbers, but this is equivalent to working in $\mathbb{R}^2$, as we will see.

(*i*) Let $\phi\colon \mathbb{R}^2 \to \mathbb{C}$ denote the canonical isomorphism $\phi(x, y) = x + yi$ as well as $F_{\mathbb{C}}\colon \mathbb{C} \to \mathbb{C}$ be continuously differentiable and $F := \phi^{-1} \circ F_{\mathbb{C}} \circ \phi\colon \mathbb{R}^2 \to \mathbb{R}^2$.

Further let $z^{(0)} = x^{(0)} + y^{(0)}i = \phi(x^{(0)}, y^{(0)}) \in \mathbb{C}$ be an initial value and let $z^{(k)}$ be a sequence of iterates of the local Newton's method in the complex numbers, defined as

$$z^{(k+1)} = z^{(k)} - F'_{\mathbb{C}}(z^{(k)})^{-1} F_{\mathbb{C}}(z^{(k)}),$$

started from $z^{(0)}$.

Show that Newton's method for F started at $(x^{(0)}, y^{(0)})$ is well defined and yields the sequence $(x^{(k)}, y^{(k)})^\mathsf{T}$ with $\phi(x^{(k)}, y^{(k)}) = z^{(k)}$ for all $k \in \mathbb{N}$.

(*ii*) Implement the local Newton's method for solving problems of the type

$$F(x) = 0$$

for $F\colon \mathbb{R}^n \to \mathbb{R}^n$ (Algorithm 5.23 of the lecture notes).

(*iii*) Reformulate the root finding problem $F_{\mathbb{C}}(z) = 0$, $z \in \mathbb{C}$ for

$$F_{\mathbb{C}}\colon \mathbb{C} \to \mathbb{C}, \quad F_{\mathbb{C}}(z) = z^3 - 1$$

into an equivalent problem in $\mathbb{R}^2$, solve the problem numerically for a grid of initial values and create a fractal plot by coloring each starting value corresponding to the root that the local Newton's method converged to starting from that point.

Solution.

(i) The claim is true for $k = 0$ by assumption. Now let the two dimensional real local Newton's method started at $(x^{(0)}, y^{(0)})$ have been well defined and have successfully reached the iterate $(x^{(k)}, y^{(k)})$.

Since $F_{\mathbb{C}}$ is complex differentiable and Φ is a linear, bounded map, we immediately obtain that

$$F'(x, y) = \phi^{-1} \circ F_{\mathbb{C}}'(\phi(x, y)) \circ \phi.$$

This matrix is nonsingular at $(x^{(k)}, y^{(k)})$ because the mapping $F_{\mathbb{C}}{}'\phi(x^{(k)}, y^{(k)})$ is invertible by assumption, where (using the Frechet derivative) we can find

$$F'(x^{(k)}, y^{(k)})^{-1} = \phi^{-1} \circ F_{\mathbb{C}}{}'(\phi(x^{(k)}, y^{(k)}))^{-1} \circ \phi,$$

so the next iteration step is well defined and from linearity of ϕ we obtain that

$$\begin{aligned}
\phi(x^{(k+1)}, y^{(k+1)}) &= \phi\left((x^{(k)}, y^{(k)}) - F'(x^{(k)}, y^{(k)})^{-1} F(x^{(k)}, y^{(k)})\right) \\
&= z^{(k)} - \phi \underbrace{F'(x^{(k)}, y^{(k)})^{-1}}_{\phi^{-1} \circ F_{\mathbb{C}}'(\phi(x^{(k)}, y^{(k)}))^{-1} \circ \phi} \underbrace{F(x^{(k)}, y^{(k)})}_{\phi^{-1} \circ F_{\mathbb{C}}(\phi(x^{(k)}, y^{(k)}))} \\
&= z^{(k)} - F_{\mathbb{C}}'(z^{(k)})^{-1} F_{\mathbb{C}}(z^{(k)}) \\
&= z^{(k+1)}.
\end{aligned}$$

(ii) See the implementation in local_newton_root.py.

(iii) We can of course simply compute the $\mathbb{R}^2$ representation of the complex cubic function and its derivative as

$$\begin{aligned}
F_3(x, y) &= \phi^{-1}\left((x + yi)^3 - 1\right) = \begin{pmatrix} x^3 - 3xy^2 - 1 \\ -y^3 + 3x^2 y \end{pmatrix} \\
F_3'(x, y) &= \phi^{-1}\left(3(x + yi)^2\right) = \begin{pmatrix} 3x^2 - 3y^2 & -6xy \\ 6xy & 3x^2 - 3y^2 \end{pmatrix},
\end{aligned}$$

where the derivative can either be found by hand or using the Cauchy-Riemann equations for $F_{\mathbb{C}}(x + yi) = u(x, y) + v(x, y)i$ leading to

$$F'(x, y) = \phi^{-1} \circ F_{\mathbb{C}}'(\phi(x, y)) \circ \phi = \begin{pmatrix} \frac{\partial}{\partial x} u(x, y) & -\frac{\partial}{\partial x} v(x, y) \\ \frac{\partial}{\partial x} v(x, y) & \frac{\partial}{\partial x} u(x, y) \end{pmatrix}.$$

Note: Showing that this matrix is nonsingular is an alternative to using the Frechet derivative before. Its determinant is $\frac{\partial}{\partial x}u(x,y)^2 + \frac{\partial}{\partial x}v(x,y)^2 = |F'_{\mathbb{C}}(\phi(x,y))|$ which can not be zero by assumption, because that would mean that the complex Newton step would not have been well defined either.

This formulation is quite efficient. However, for more general exponents, the polynomial

$$z^n - 1$$

and its derivative can be computed easily using the polar coordinates r, φ of the points $z = \phi(x,y)$, where we obtain

$$F_n(x,y) = \phi^{-1}\left((x+yi)^n - 1\right) = r^n \begin{pmatrix}\cos(n\varphi)\\ \sin(n\varphi)\end{pmatrix}$$
$$F_n'(x,y) = \phi^{-1}\left(n(x+yi)^{n-1}\right) = nr^{n-1}\begin{pmatrix}\cos((n-1)\varphi) & -\sin((n-1)\varphi)\\ \sin((n-1)\varphi) & \cos((n-1)\varphi)\end{pmatrix},$$

so if one plans to also do computations for general polynomial orders, the decrease in performance may be worth it.

For the implementation see driver_ex_018_newton_fractals.py.

We obtain the fractal plots in Figure 0.1, which clearly show somewhat structured (with respect to n) chaotic convergence behavior.

Note that it would be interesting to make such a plot for the globalized Newton's method in optimization applied to the functional $x \mapsto \frac{1}{2}\|F(x)\|_2$. All roots of F are stationary points of the optimization functional, but we obtain additional stationary points.

(10 Points)

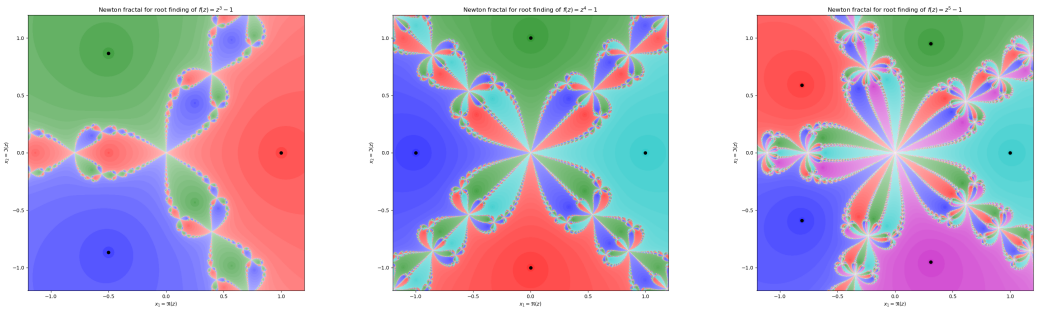

Figure 0.1: Newton fractals for $z^n - 1$ with $n \in \{3, 4, 5\}$, left to right. Each point in $\mathbb{R}^2$/the complex plane is colored according to the root Newton's method converged to when started at said point. Opacity according to number of iterations needed to converge log-relative to maximum number of iterations needed. White color means nonconvergence. Roots are marked as black dots.

Homework Problem 5.3 (On the Restriction $\sigma \in (0, \frac{1}{2})$ in Globalized Newton) 7 Points

In the globalized Newton's method for optimization (Algorithm 5.30 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

$$f(x) = \frac{1}{2} x^\mathsf{T} A x + b^\mathsf{T} x + c$$

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

$$
\begin{aligned}
& f(x+d) - f(x) \leq \sigma f'(x)d \\
\overset{\text{Taylor}}{\Leftrightarrow} \quad & \frac{1}{2} d^\mathsf{T} f''(\xi) d + f'(x)d \leq \sigma f'(x)d \\
\overset{\text{Rearrange}}{\Leftrightarrow} \quad & \frac{1}{2} d^\mathsf{T} f''(\xi) d \leq (\sigma - 1) f'(x)d \\
\overset{\text{form of } d}{\Leftrightarrow} \quad & \frac{1}{2} \nabla f(x)^\mathsf{T} f''(x)^{-1} f''(\xi) f''(x)^{-1} \nabla f(x) \leq (1-\sigma) \nabla f(x)^\mathsf{T} f''(x)^{-1} \nabla f(x) \qquad (*)
\end{aligned}
$$

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

$$
\begin{aligned}
\overset{f'' \equiv A}{\Leftrightarrow} \quad & \frac{1}{2} \nabla f(x)^\mathsf{T} A^{-1} A A^{-1} \nabla f(x) \leq (1-\sigma) \nabla f(x) A^{-1} \nabla f(x) \\
\Leftrightarrow \quad & 0 \leq \left(\frac{1}{2} - \sigma \right) \underbrace{\nabla f(x)^\mathsf{T} A^{-1} \nabla f(x)}_{>0} .
\end{aligned}
$$

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

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

(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

$$e^{(k)} := \frac{1}{2}\nabla f(x^{(k)})^\mathsf{T} f''(x^{(k)})^{-1} f''(\xi^{(k)}) f''(x^{(k)})^{-1} \nabla f(x^{(k)}) - \frac{1}{2}\nabla f(x^{(k)})^\mathsf{T} f''(x^{(k)})^{-1}\nabla f(x^{(k)})$$

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

$$\begin{aligned}\frac{1}{2}\nabla f(x^{(k)})^\mathsf{T} f''(x^{(k)})^{-1} f''(\xi^{(k)}) f''(x^{(k)})^{-1} \nabla f(x^{(k)}) &= \frac{1}{2}\nabla f(x^{(k)})^\mathsf{T} f''(x^{(k)})^{-1}\nabla f(x^{(k)}) + e^{(k)} \\ &\leq \underbrace{(1-\sigma)}_{>\frac{1}{2}} \underbrace{\nabla f(x^{(k)})^\mathsf{T} f''(x^{(k)})^{-1}\nabla f(x^{(k)})}_{>0 \text{ and converging to } 0}\end{aligned}$$

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 5.4 (Globalized Newton's Method in Optimization) 8 Points

Implement the globalized Newton's method for optimization (Algorithm 5.30 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 0.2 and 0.3. 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.

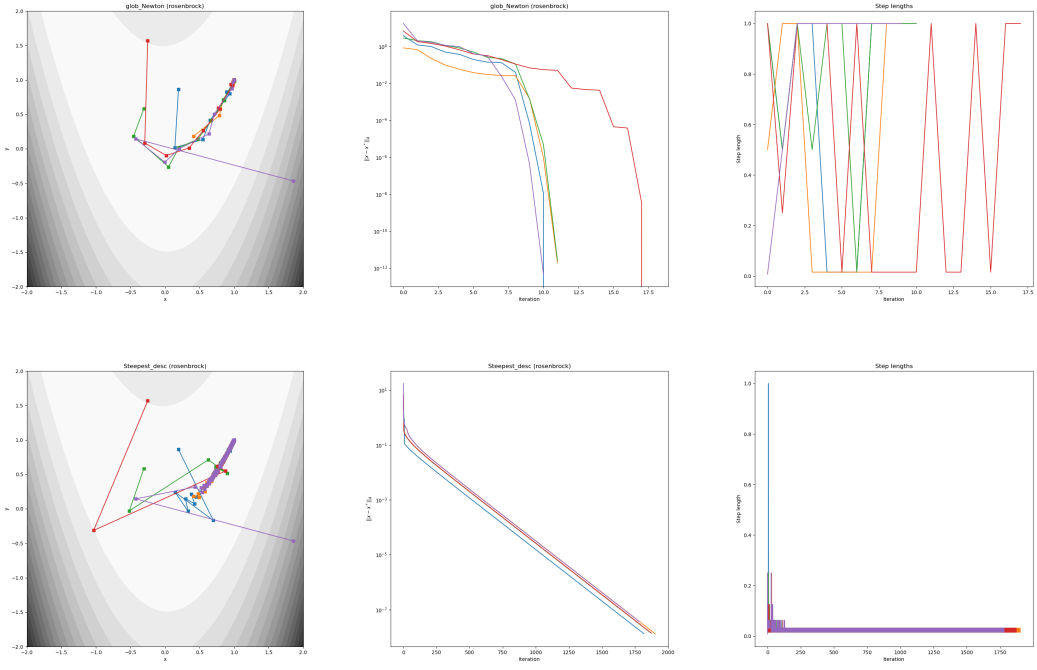

Figure 0.2: Newton (top) vs steepest descent (bottom) for Rosenbrock function.

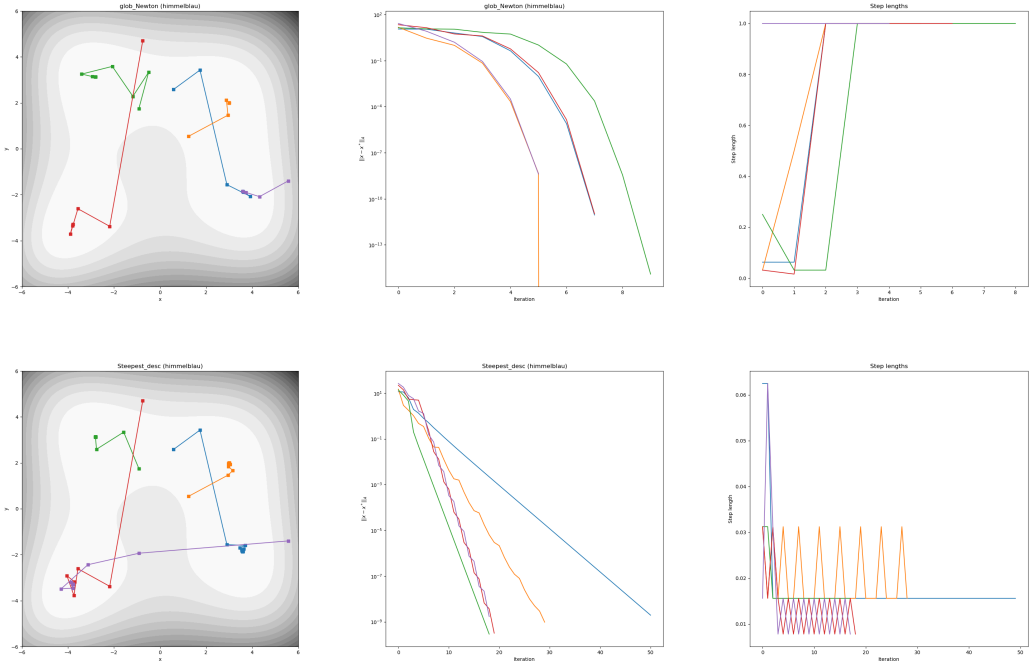

Figure 0.3: Newton (top) vs steepest descent (bottom) for Himmelblau function.