Exercise 5 - Solution

Date issued: 13th May 2024
Date due: 22nd May 2024

Homework Problem 5.1 (Efficiency of Wolfe-Powell Step Sizes for $C^{1,1}$ Functions) 5 Points

Let $f \in C^1$ and let $x^{(0)} \in \mathbb{R}^n$ be an initial iterate of the generic descent scheme (Algorithm 4.2). Further assume that f' is Lipschitz continuous on the sublevel set $\mathcal{M}_f(x^{(0)}) \coloneqq \{x \in \mathbb{R}^n \mid f(x) \le f(x^{(0)})\}$.

Show that step sizes $\alpha^{(k)}$ that satisfy the Wolfe-Powell-conditions at $x^{(k)}$ for the descent direction $d^{(k)}$ for all k are efficient and that there is a $c > 0$ such that

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \le -c \left(\cos \measuredangle \big(-\nabla_M f(x^{(k)}), d^{(k)} \big) \| f'(x^{(k)})^\top \|_{M^{-1}} \right)^2$$

for all $k \ge 0$.

Solution.

For efficiency of step sizes α, we need to show that there exists a $\theta > 0$ such that, as long as $d^{(k)} \neq 0$, we have that

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \le f(x^{(k)}) - \theta \left(\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \right)^2 \tag{4.11}$$

for all $k \ge 0$.

Because the Armijo conditions are satisfied, we actually have a descent scheme.

The curvature condition of the Wolfe-Powell step-lengths states that

$$f'(x^{(k)} + \alpha^{(k)}\, d^{(k)})\, d^{(k)} \ge \tau\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad \varphi'(\alpha^{(k)}) \ge \tau\, \varphi'(0) \tag{4.17}$$

for a $\tau \in (\sigma, 1)$ and all $k \geq 0$. Subtracting $f'(x^{(k)})d^{(k)}$ from both sides, applying Cauchy-Schwarz's inequality and using the Lipschitz continuity of f' (measured in the M^{-1} and the M norm, respectively), we obtain that

$$\begin{aligned}(\tau-1)\, f'(x^{(k)})\, d^{(k)} &\leq \Big(f'(x^{(k)} + \alpha^{(k)}\, d^{(k)}) - f'(x^{(k)})\Big)\; d^{(k)} \\ &= \Big(f'(x^{(k)} + \alpha^{(k)}\, d^{(k)}) - f'(x^{(k)})\Big)\; M^{-1}M\, d^{(k)} \\ &\leq \left\| M^{-1}\Big(f'(x^{(k)} + \alpha^{(k)}\, d^{(k)}) - f'(x^{(k)})\Big)^{\intercal}\right\|_{M} \left\|d^{(k)}\right\|_{M} \\ &= \left\| \Big(f'(x^{(k)} + \alpha^{(k)}\, d^{(k)}) - f'(x^{(k)})\Big)^{\intercal}\right\|_{M^{-1}} \left\|d^{(k)}\right\|_{M} \\ &\leq L_{M^{-1},M}\alpha^{(k)} \left\|d^{(k)}\right\|_{M}^{2}\end{aligned}$$

(2 Points)

Note: Note that we were able to use Lipschitz continuity because we are working with a descent scheme.

Rearranging the estimate yields the following bound on the step size

$$\alpha^{(k)} \geq \frac{(\tau-1)}{L_{M^{-1},M}} \frac{f'(x^{(k)})\, d^{(k)}}{\left\|d^{(k)}\right\|_{M}^{2}}.$$

Inserting that into the Armijo-condition

$$f(x^{(k)} + \alpha^{(k)}\, d^{(k)}) \leq f(x^{(k)}) + \sigma\, \alpha^{(k)}\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad \varphi(\alpha^{(k)}) \leq \varphi(0) + \sigma\, \alpha^{(k)}\, \varphi'(0), \tag{4.12}$$

we immediately obtain that

$$f(x^{(k)} + \alpha^{(k)}\, d^{(k)}) \leq f(x^{(k)}) + \sigma\, \alpha^{(k)} \underbrace{f'(x^{(k)})\, d^{(k)}}_{<0} \leq f(x^{(k)}) + \underbrace{\frac{\sigma(\tau-1)}{L_{M^{-1},M}}}_{=:-c,\, c>0} \left(\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_{M}}\right)^{2}$$

for all $k \geq 0$ which shows efficiency. (2 Points)

The additional statement simply follows from the fact that

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_{M}} = \frac{(\nabla_M f(x^{(k)}), d^{(k)})_M}{\|d^{(k)}\|_{M}} \frac{\|\nabla_M f(x^{(k)})\|_M}{\|\nabla_M f(x^{(k)})\|_M} = -\cos\measuredangle\big(-\nabla_M f(x^{(k)}), d^{(k)}\big)\|f'(x^{(k)})^{\intercal}\|_{M^{-1}}$$

and the square. (1 Point)

Homework Problem 5.2 (Scaling Invariance of Armijo- and Curvature Conditions) 5 Points

Show the statement of remark 4.21, i. e. that when a step length α satisfies any of the Armijo- or curvature conditions (4.12), (4.17) and (4.18) for $g(x) := \gamma f(Ax+b)+\delta$ at $x \in \mathbb{R}^n$ with search direction $d \in \mathbb{R}^n$, where $A \in \mathbb{R}^{n\times n}$ is non-singular, $b \in \mathbb{R}^n$, $\gamma > 0$ and $\delta \in \mathbb{R}$, then it satisfies the respective conditions for f at $Ax+b$ with the search direction Ad.

Solution.

If α satisfies the Armijo condition for g at x and direction d with parameter σ, then

$$\begin{aligned} g(x+\alpha d) &\le g(x)+\sigma g'(x)d \\ \Rightarrow \gamma f(A(x+\alpha d)+b)+\delta &\le \gamma f(Ax+b)+\delta+\sigma\gamma f'(Ax+b)Ad \end{aligned}$$

and deviding by $\gamma > 0$ and subtracting δ from both sides yields

$$\Rightarrow f(Ax+b+\alpha Ad) \le f(Ax+b)+\sigma f'(Ax+b)Ad$$

meaning that α satisfies the Armijo condition for f at $Ax+b$ and direction Ad with parameter σ. (2 Points)

If the curvature condition is satisfied by α for g at x with direction d and parameter τ, then

$$\begin{aligned} g'(x+\alpha d)d &\ge \tau g'(x)d \\ \Rightarrow \gamma f'(A(x+\alpha d)+b)Ad &\ge \tau\gamma f'(Ax+b)Ad \end{aligned}$$

and deviding by $\gamma > 0$ yields

$$\Rightarrow f'(Ax+b+\alpha Ad)Ad \ge \tau f'(Ax+b)Ad,$$

which means that the curvature condition is satisfied by α for f at $Ax+b$ with direction Ad and parameter τ. (2 Points)

If the strong curvature condition is satisfied by α for g at x with direction d and parameter τ, then

$$\begin{aligned} |g'(x+\alpha d)d| &\le -\tau g'(x)d \\ \Rightarrow |\gamma f'(A(x+\alpha d)+b)Ad| &\le -\tau\gamma f'(Ax+b)Ad \end{aligned}$$

and deviding by $\gamma > 0$ yields

$$\Rightarrow |f'(Ax+b+\alpha Ad)Ad| \le -\tau f'(Ax+b)Ad$$

which means that the strong curvature condition is satisfied by α for f at $Ax + b$ with direction Ad and parameter τ. (1 Point)

Note: We did not require A to be nonsingular anywhere in the proof. This requirement is merely needed to show the inverse by applying the result we just proved to the inverted form that generates f from g.

Homework Problem 5.3 (Implementation of Nonlinear Steepest Descent and Armijo Backtracking) 8 Points

Implement the M-steepest descent method as outlined in Algorithm 4.22 with the original Armijo backtracking as outlined in Algorithm 4.11. You can also try to use the modified (interpolating) Armijo backtracking as outlined in Algorithm 4.15.

Visualize and examine the effect of the parameters of the step size strategy on the behavior of the algorithm when applied to quadratic, strongly convex functions, Rosenbrock's and/or Himmelblau's functions.

Solution.

Figures 0.1 and 0.3 show the behavior of the iterates and Figures 0.2 and 0.4 show the (approximate) $f''(x^*)$-error in a semilogarithmic plot for the steepest descent method with armijo backtracking applied to quadratic optimization problem and the minimization of the Rosenbrock function (with minimizer $x^* = (1, 1)$), respectively, for backtracking parameters $\beta \in \{0.01, 0.5, 0.99\}$ and $\sigma \in \{10^{-2}, 0.3, 0.7\}$.

We can observe that when β is chosen very small (top rows), the choice of σ does not influence the behavior greatly. This is due to one backtracking step reducing the trial step size so much that we essentially end up with gradient flow like behavior. Only when σ is really large (very relaxed acceptance of trial steps) we can get lucky with the first trials in the rosenbrock case.

For rather strict acceptance of trial step sizes (large σ, right columns) we also observe gradient flow type behavior (as this forces the step sizes to be small).

When β is chosen large (bottom rows) we can expect fine adjustments of the step sizes. When σ is small an we are lenient with accepting step sizes, we end up with pretty extreme zig-zagging. We observe the best behavior for moderate parameter choices. Small β and large σ dominate the behavior in any case.

Near minimizers with s. p. d. $f''(x^*)$ (this is the sufficient second order condition) we can expect nonlinear problems to behave alsmost quadratically, so the difference in function values $f(x^{(k)}) - f(x^*) \approx \|x^{(k)} - x^*\|_{f''(x^*)}$, which conincides with the A-error in the quadratic case. Looking at those plots over the iterations in semilogarithmic axes shows the expected behavior once close to the minimizer. Zig-zagging and sudden drops are experienced well outside close neighbourhoods of the minimizer.

Note that the steepest descent method is struggling strongly with the rosenbrock function due to zig zagging along the barely-sloping banana-shaped valley leading to the minimizer, as no local curvature information is used in the model hessians (identity preconditioner for all plots). Depending on the safe-guarding conditions on might actually end up with the step size computation failing because step sizes become to small and the function almost appears locally constant.

Additionally, using interpolation over simple backtracking can improve the number of iterations needed until the termination criteria is used, but it actually may increase them as well.

(8 Points)

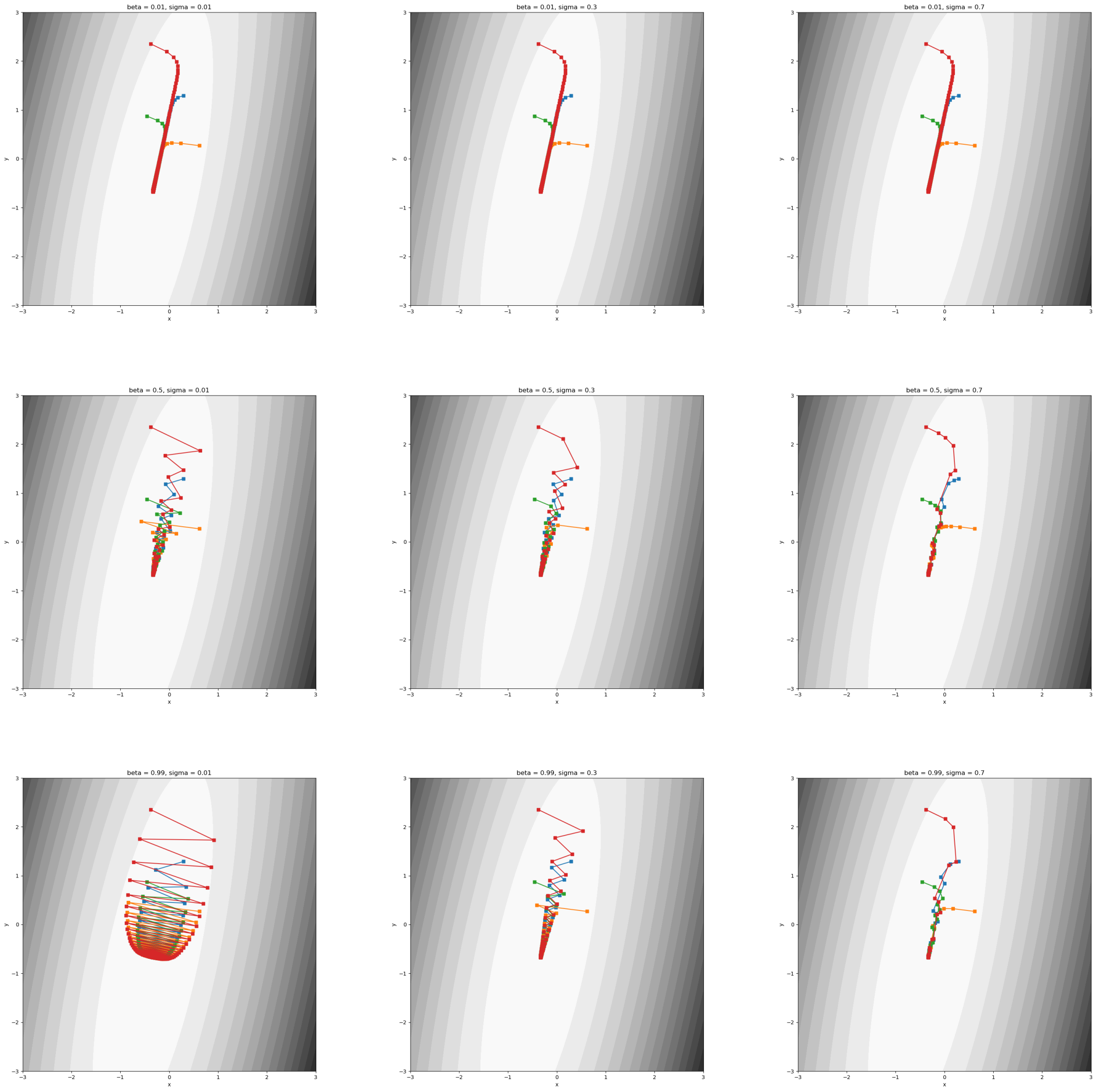

Figure 0.1: Iterates for steepest descent with armijo step length rule applied to quadratic optimization.

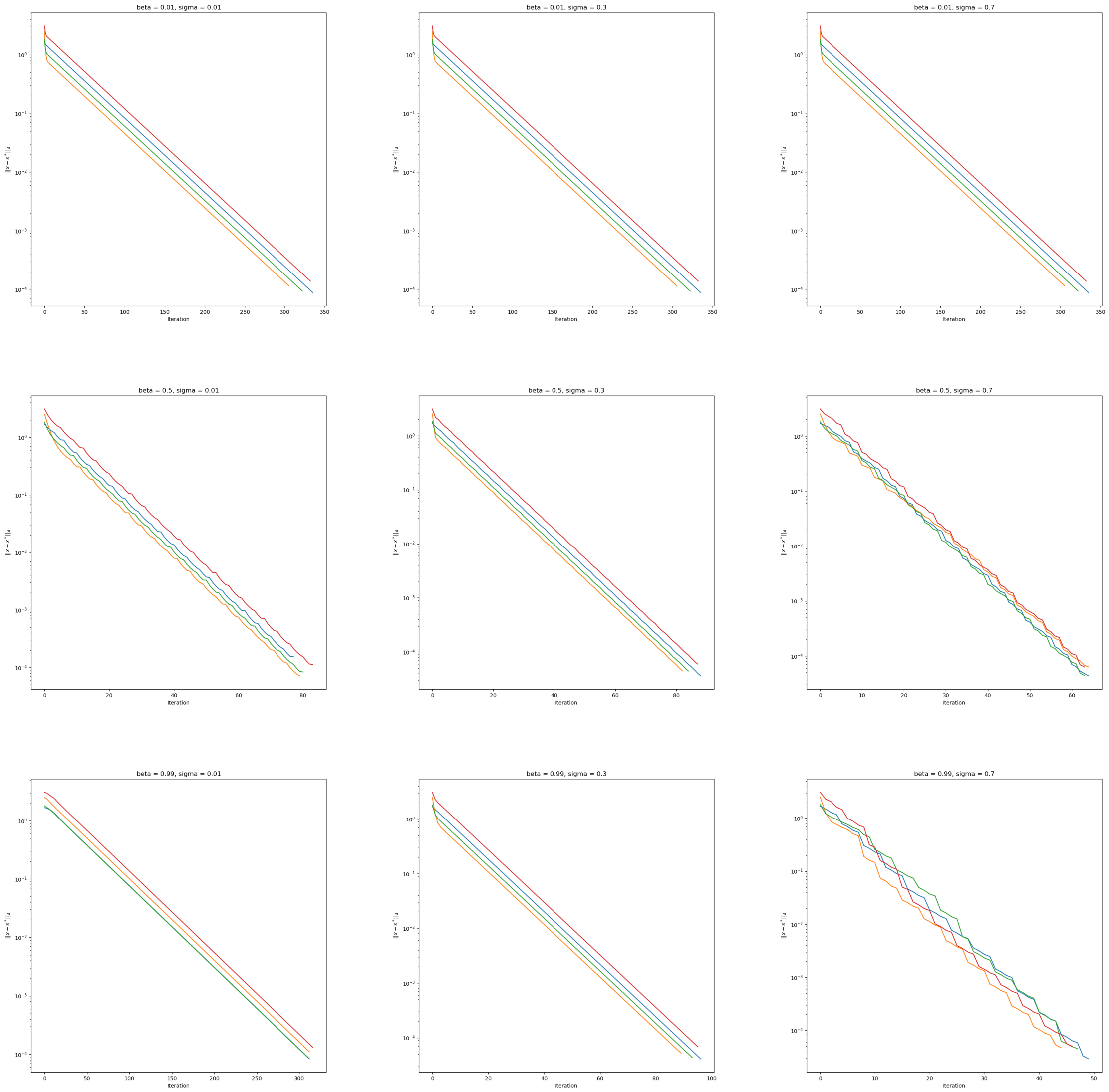

Figure 0.2: A- norm of errors for steepest descent with armijo step length rule applied to quadratic optimization.

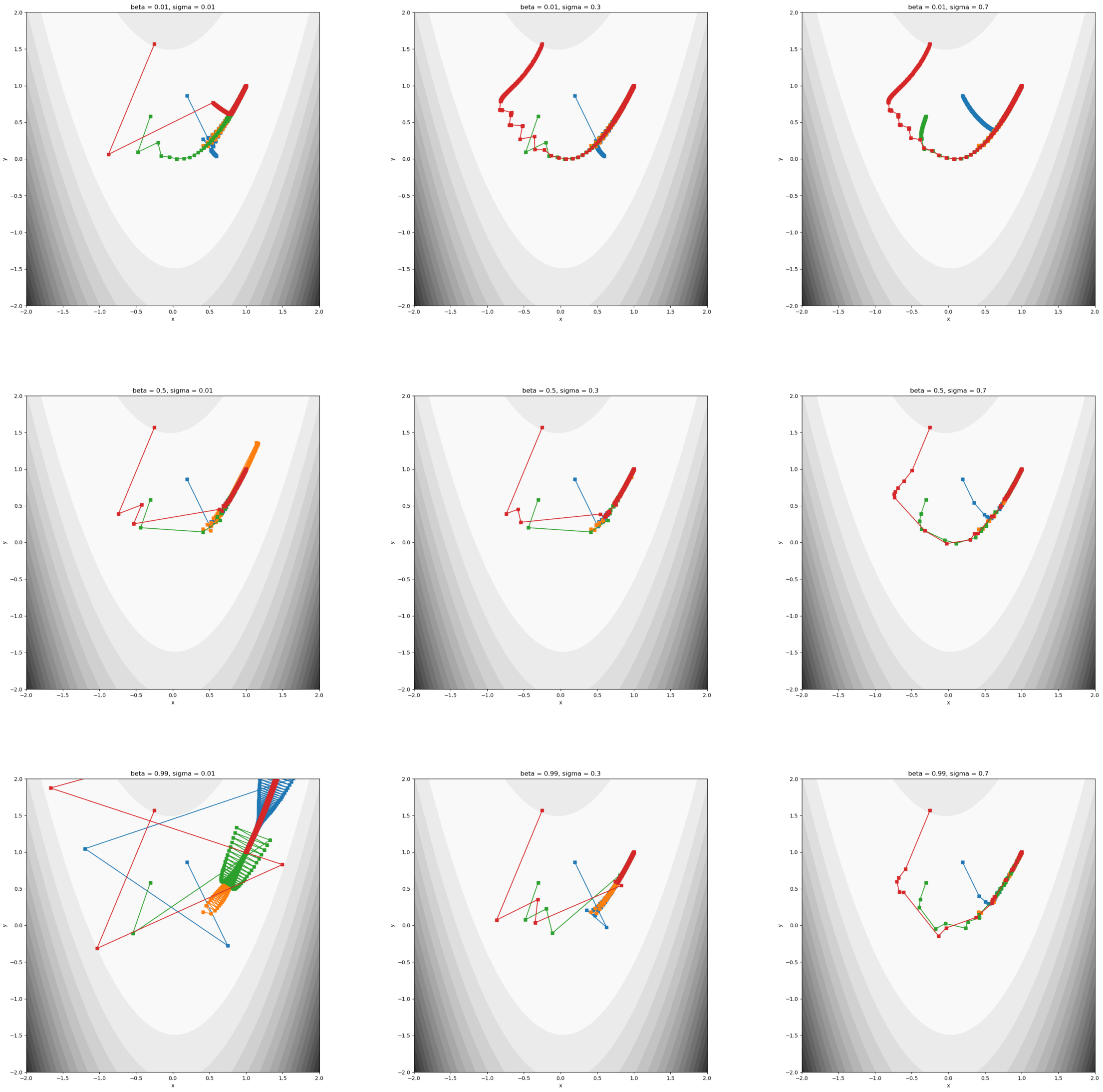

Figure 0.3: Iterates for steepest descent with armijo step length rule applied to rosenbrock optimization.

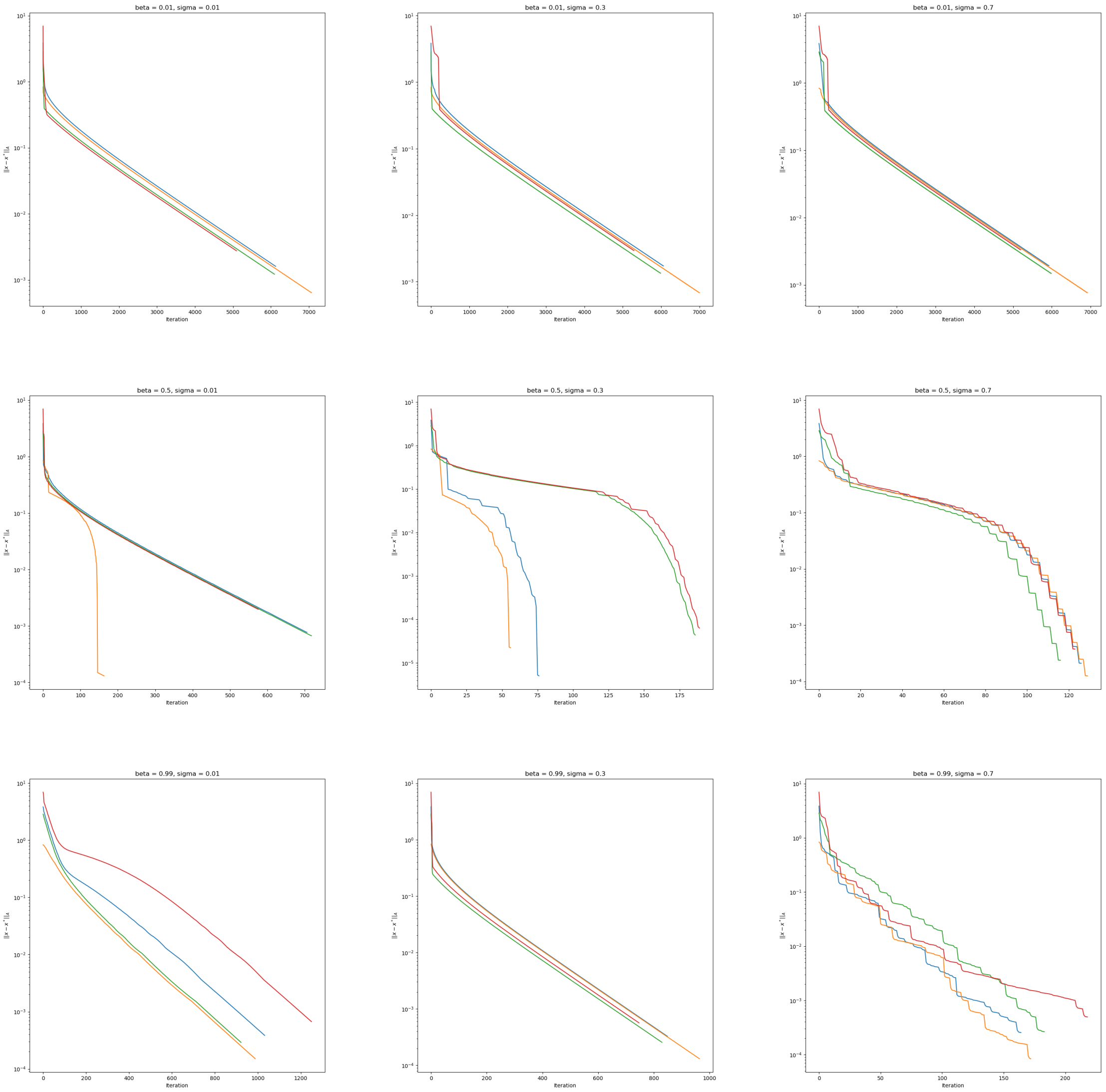

Figure 0.4: Approximate $f''(x^*)$-error for steepest descent with armijo step length rule applied to rosenbrock optimization.

Homework Problem 5.4 (Affine Invariance of Newton's Method for Root Finding) 10 Points

Prove the statement in remark 4.29(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 4.23 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

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

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

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

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

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

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

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

where

$$\begin{aligned} H(y) &:= AF(y) \\ H'(y) &= AF'(y). \end{aligned}$$

(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

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