Oregon State
University

Homework 4. Gradient Descent Method

Hyuntaek Oh

ohhyun@oregonstate.edu

Due: May 7, 2025

ECE599/ AI539 Nonlinear Optimization (Spring 2025)
Homework 4. Gradient Descent Method (Due: 11:59pm on May 7, Wednesday.)

Instruction: Students should provide enough detail of the logical procedure of deriving answers. Answers without sufficient justification will receive partial or no credit.

Reading: Section 8.1-8.6 of the textbook (Luenberger and Ye).

1. In Section 7.7 of [Luenberger and Ye], you can find the proof of Global Convergence Theorem. Study Section 7.7 and provide the proof of Global Convergence Theorem in your solution.

According to "Linear and Non-Linear Programming", Global Convergence Theorem (GCT) is:

Let $\mathbf{A}$ *be an algorithm on* X*, and suppose that, given* $\mathbf{x}_0$ *the sequence* $\{\mathbf{x}_k\}_{k=0}^{\infty}$ *is generated satisfying*

$$\mathbf{x}_{k+1} \in \mathbf{A}(\mathbf{x}_k).$$

Let a solution set $\Gamma \subset X$ *be given, and suppose*

(i) all points $\mathbf{x}_k$ *are contained in a compact set* $S \subset X$

(ii) there is a continuous function Z *on* X *such that*

(a) if $\mathbf{x} \notin \Gamma$*, then* $Z(\mathbf{y}) < Z(\mathbf{x})$ *for all* $\mathbf{y} \in \mathbf{A}(\mathbf{x})$

(b) if $\mathbf{x} \in \Gamma$*, then* $Z(\mathbf{y}) \leq Z(\mathbf{x})$ *for all* $\mathbf{y} \in \mathbf{A}(\mathbf{x})$

(iii) the mapping $\mathbf{A}$ *is closed at points outside* Γ*.*

Then the limit of any convergent subsequence of $\{\mathbf{x}_k\}$ *is a solution.*

Since Z *is a descent function, and* $x_k \notin \Gamma$*, the function strictly decreases:*

$$f(x_1) < f(x_0), \quad f(x_2) < f(x_1), \quad \cdots \Rightarrow f(x_{k+1}) < f(x_k)$$

This means:

$$f(x_k) \leq f(x_0), \forall k$$

So, all points x_k *are contained in the level set:*

$$\mathcal{L} := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$$

According to the lecture 8 slides, the level set $\mathcal{L}$ is compact since f is continuous, and $\mathcal{L}$ is closed and bounded. Based on the Bolzano-Weierstrass Theorem, if $\mathcal{S}$ is compact, and $\{x_k\}$ is such that $x_k \in \mathcal{S}, \forall k$, then $\{\mathbf{x}_k\}$ has a limit point in $\mathcal{S}$, meaning that it has a convergent subsequence that converges to some point in $\mathcal{S}$.

Therefore, by compactness, $\{x_k\}$ has at least one convergent subsequence:

$$x_{k_j} \to x^* \quad \text{for some } x^* \in \mathbb{R}^n$$

By applying it to the algorithm A, next iterates also converge:

$$x_{k_j+1} \in A(x_{k_j})$$

$$x_{k_j+1} \to y^*$$

By the definition of closedness of a point-to-set mapping, if

$$x_{k_j} \to x^*$$

$$x_{k_j+1} \in A(x_k) \to y^*$$

then,

$$y^* \in A(x^*)$$

Thus, $x^* \in A(x^*)$ holds, meaning that it's a fixed point of the mapping.

Now, the limit point x^* above is needed to check whether it is a stationary point or not. This means:

$$x^* \in \Gamma, \ \text{where } \Gamma = \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$$

For contradiction, assume that the limit point x^* is not a stationary point:

$$x^* \notin \Gamma \to \nabla f(x^*) \neq 0$$

By the descent function condition, the condition below should hold:

$$\text{For } \textbf{any } y \in A(x^*) \to f(y) < f(x^*)$$

However, from the previous proof y, this cannot hold since it is contradicted:

$$x^* \in A(x^*) \Rightarrow f(x^*) < f(x^*)$$

In conclusion, because of the contradiction above, $x^* \in \Gamma$, meaning that x^* is a stationary point, and there are no further possible descent points.

2. Show that if **A** is a continuous point-to-point mapping, the Global Convergence Theorem is valid even without assumption (i). Compare with Example 2, Section 7.7.

An algorithm $A : \mathbb{R}^n \to \mathbb{R}^n$ is continuous point-to-point mapping, then it is closed at every point in its domain. According to Lecture 8 slides, a continuous point-to-point mapping is closed. Due to this, continuity and closed mapping are respectively:

$$x_k \to x \Rightarrow A(x_k) \to A(x)$$

$$x_k \to x, \quad A(x_k) \to y \Rightarrow y \in A(x)$$

When $\{x_k\}$ is the sequence generated by the algorithm A:

$$x_{k+1} = A(x_k), \quad x_0 \in \mathbb{R}^n$$

$Z : \mathbb{R}^n \to \mathbb{R}$ is a descent function for this algorithm:

$$x \notin \Gamma \Rightarrow f(A(x)) < f(x)$$

where $\Gamma := \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$ is the set of stationary points. So, the sequence $\{x_k\}$ is contained in a compact level set:

$$x_k \in \mathcal{L} := \{x \in \mathbb{R}^n | f(x) \leq f(x_0)\}$$

There exists a convergent subsequence:

$$x_{k_j} \to x^*$$

Now by continuity of A:

$$x_{k_j+1} = A(x_{k_j}) \to A(x^*)$$

So, This means:

$$x^* = \lim x_{k_j}, \text{ and } x^{*+1} := \lim x_{k_j+1} = A(x^*)$$

Because of the function f's continuity:

$$x_{k_j} \to x^* \text{ and } x_{k_j+1} \to x^{*+1}$$

Since $f(x^{*+1}) < f(x^*)$ and both are real numbers, there is a positive gap between them:

$$\epsilon := f(x^*) - f(x^{*+1}) > 0$$

The gap between them is strictly positive because $x^* \notin \Gamma$, and f is strictly decreasing unless at a stationary point.

Furthermore, for sufficiently large j, it would be:

$$f(x^{*+1}) < f(x^*) - \epsilon, \text{ for some } \epsilon > 0$$

However, the function values decrease by at least ϵ infinitely often, contradicting the fact that $f(x_k)$ is monotonically decreasing and is bounded below, meaning that it must converge. So, this infinite descent cannot happen. Due to it, the assumption $x^* \notin \Gamma$ must be false, and any limit point x^* must lie in Γ. Thus, the Global Convergence Theorem is valid even without assumption (i).

3. For $\delta > 0$ define the map $\mathbf{S}^\delta$ by

$$\mathbf{S}^\delta(\mathbf{x}, \mathbf{d}) = \{\mathbf{y} : \mathbf{y} = \mathbf{x} + \alpha\mathbf{d}, \quad 0 \le \alpha \le \delta; \quad f(\mathbf{y}) = \min_{0 \le \beta \le \delta} f(\mathbf{x} + \beta\mathbf{d})\}.$$

Thus $\mathbf{S}^\delta$ searches the interval [0, δ] for a minimum of $f(\mathbf{x} + \alpha\mathbf{d})$, representing a "limited range" line search. Show that if f is continuous, $\mathbf{S}^\delta$ is closed at all $(\mathbf{x}, \mathbf{d})$.

Suppose that f is continuous, $(x_k, d_k) \to (x, d)$ as $k \to \infty$, and $y_k \in S^\delta(x_k, d_k)$. There exists $\alpha_k \in [0, \delta]$ such that $y_k = x_k + \alpha_k d_k$ and $f(y_k) = \min_{\beta \in [0,\delta]} f(x_k + \beta d_k)$ to assume $y_k \to y$.

Since $y_k = x_k + \alpha_k d_k$, $x_k \to x$, $d_k \to d$, and $\alpha_k \in [0, \delta]$, by Bolzano-Weierstrass, the sequence $\{\alpha_k\}$ has a convergent subsequence, and because of $y_k \to y$:

$$\alpha_k \to \alpha \in [0, \delta] \Rightarrow y_k = x_k + \alpha_k d_k \to y = x + \alpha d$$

This means $y = x + \alpha d$ for some $\alpha \in [0, \delta]$. Since f is continuous and $y_k \to y$, it would be:

$$f(y_k) \to f(y)$$

Also, from definition of $y_k \in S^\delta(x_k, d_k)$:

$$f(y_k) = \min_{\beta \in [0,\delta]} f(x_k + \beta d_k)$$

Since f is continuous, for each k, the functions $\phi_k(\beta)$ and $\phi(\beta)$ can be defined to show uniform convergence:

$$\phi_k(\beta) := f(x_k + \beta d_k), \text{ for } \beta \in [0, \delta]$$

$$\phi(\beta) := f(x + \beta d), \text{ for } \beta \in [0, \delta]$$

Based on the uniform convergence, the inequality below is uniformly in β:

$$\beta \in [0, \delta] \Rightarrow ||x_k + \beta d_k - (x + \beta d)|| \leq ||x_k - x|| + \delta ||d_k - d|| \rightarrow 0$$

Because f is uniformly continuous on compact sets, uniform convergence is preserved:

$$\phi_k(\beta) \rightarrow \phi(\beta)$$

This can be applied to the original inequality:

$$\lim_{k \to \infty} \min_{0 \leq \beta \leq \delta} f(x_k + \beta d_k) = \min_{0 \leq \beta \leq \delta} f(x + \beta d)$$

However, $f(y_k) = \min_\beta f(x_k + \beta d_k)$, so it would be:

$$f(y_k) \rightarrow \min_\beta f(x + \beta d)$$

By using the conditions that are $y_k \rightarrow y$ and f is continuous:

$$f(y_k) \rightarrow f(y) = lim f(y_k) = \min_{0 \leq \beta \leq \delta} f(x + \beta d)$$

Thus, $y \in S^\delta(x, d) \Rightarrow S^\delta$ is closed at (x, d).

4. (MATLAB/Python experiment) Consider the following nonlinear optimization problem:

$$\min_{x,y \in \mathbb{R}} f(x, y) = x^2 - 5xy + y^4 - 25x - 8y \qquad (1)$$

Implement Gradient Descent with Armijo's rule for line search, and use it to find a relative minimum point. Explain the parameters of the algorithm you used. Plot (i) $||\nabla f(\mathbf{x}_k)||$ versus k (use the log scale for the y-axis so that will be able to recognize the small difference such as the one between 10^{-4} and 10^{-6}), and (ii) $f(\mathbf{x}_k)$ versus k and provide interpretation. Check the Hessian at the algorithm output to verify whether the algorithm output is indeed a relative minimum point.

(i) Explain the parameters of the algorithm you used

The parameters I used is in Figure 1. For Armijo's rule algorithm, there are several parameters to control the convergence: α, ϵ, η, and tol (stop condition when the gradient norm is small enough). The initial value of α is 1.0, ϵ is 0.001 (1e-3), which should be in the range between 0 and 1, and η is 2.0 because it should be greater than 1.

(ii) *Plot $||\nabla f(\mathbf{x}_k)||$ versus k (use the log scale for the y-axis so that will be able to recognize the small difference such as the one between 10^{-4} and 10^{-6})*

Figure 3 (top) shows the result plot of $||\nabla f(\mathbf{x}_k)||$ versus k. The graph of $||\nabla f(\mathbf{x}_k)||$ versus iteration (in log scale) shows a steady exponential decrease. After around 450 iterations, the gradient norm drops sharply, indicating that the iterates are entering near the local minimum. Beyond iteration 460, the gradient norm remains below the tolerance threshold 10^-6, meaning convergence.

(iii) *Plot $f(\mathbf{x}_k)$ versus k*

As shown in Figure 3 (bottom), the graph of $f(x_k)$ versus iteration k illustrates rapid initial decrease in the objective function. After around 50 iterations, the decrease slows down as the iterates approach a local minimum. The curve flattens out after iteration 100. This means the iterates approach near a local minimum.

(iv) *Check the Hessian at the algorithm output to verify whether the algorithm output is indeed a relative minimum point*

For a function $f(x, y)$, the Hessian is the matrix of second derivatives:

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

The second derivatives are:

$$\frac{\partial^2 f}{\partial x^2} = 2, \frac{\partial^2 f}{\partial x \partial y} = -5, \frac{\partial^2 f}{\partial y^2} = 12y^2$$

So, the Hessian matrix is:

$$H(x, y) = \begin{bmatrix} 2 & -5 \\ -5 & 12y^2 \end{bmatrix}$$

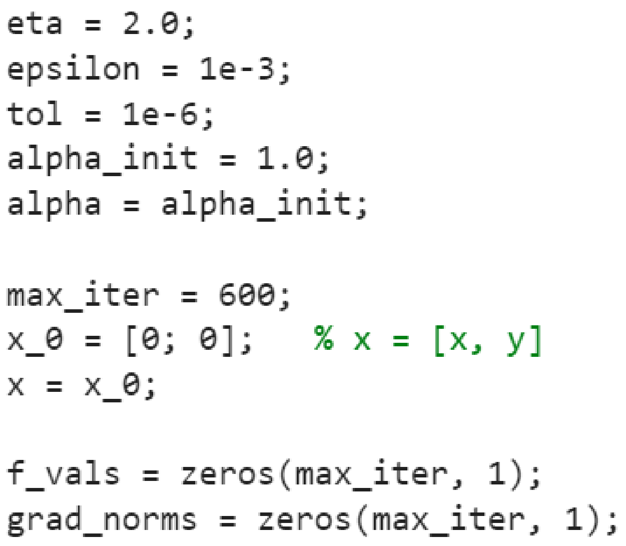
As can be seen in Figure 2, a relative minimum point (x^*, y^*) is (20, 3). Applying it to the Hessian matrix:

$$H(x, y) = \begin{bmatrix} 2 & -5 \\ -5 & 12 \cdot (3)^2 \end{bmatrix} = \begin{bmatrix} 2 & -5 \\ -5 & 108 \end{bmatrix}$$

To verify that x^* is a local minimum, the Hessian at that point must be positive definite. The leading minor $H_{11} = 2 > 0$, and the determinant of the matrix is:

$$det(H) = 2 * 108 - (-5)^2 = 216 - 25 = 191 > 0$$

Thus, the Hessian matrix is positive definite, meaning that the point is a local minimum.

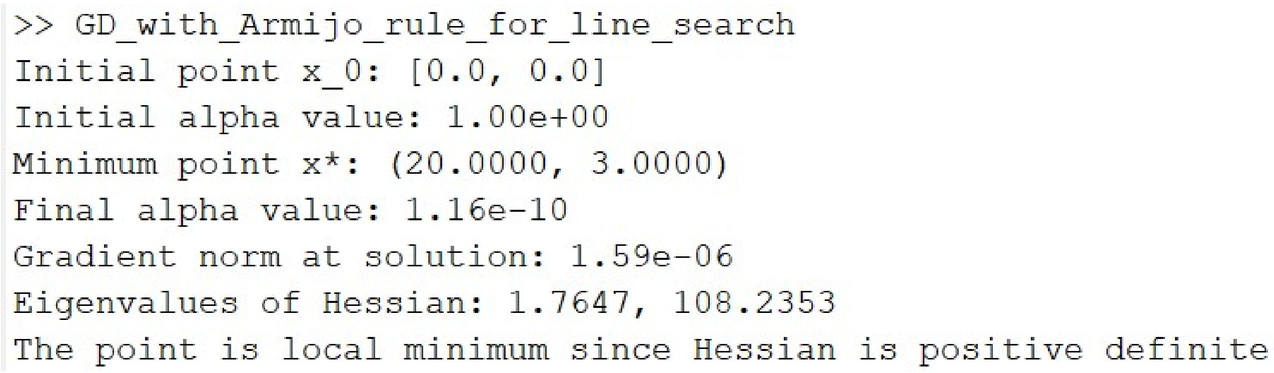


```
eta = 2.0;
epsilon = 1e-3;
tol = 1e-6;
alpha_init = 1.0;
alpha = alpha_init;

max_iter = 600;
x_0 = [0; 0];   % x = [x, y]
x = x_0;

f_vals = zeros(max_iter, 1);
grad_norms = zeros(max_iter, 1);
```

Figure 1: Parameters Initialization



```
>> GD_with_Armijo_rule_for_line_search
Initial point x_0: [0.0, 0.0]
Initial alpha value: 1.00e+00
Minimum point x*: (20.0000, 3.0000)
Final alpha value: 1.16e-10
Gradient norm at solution: 1.59e-06
Eigenvalues of Hessian: 1.7647, 108.2353
The point is local minimum since Hessian is positive definite
```

Figure 2: Result

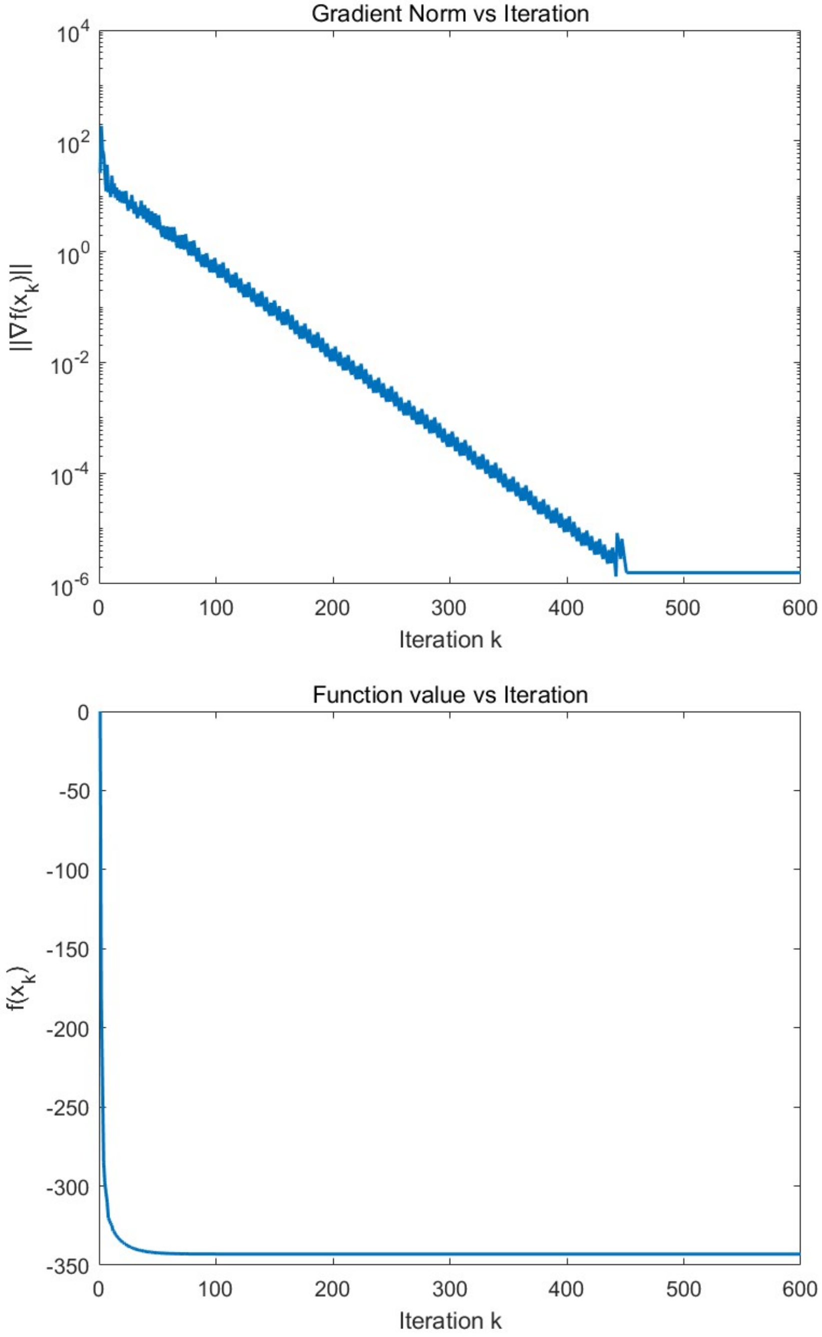


Figure 3: $||\nabla f(\mathbf{x}_k)||$ vs. Iteration k (top) and $f(\mathbf{x}_k)$ vs. Iteration k (bottom)