Exercise Collection

Nonlinear Programming – Gradient Descent and Newton's Method

English

Wolfgang Boehme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

25 October 2026

The documentation on this website was developed with the support of the AI Claude (Anthropic).

Table of Contents

1 Level 1: Easy **2**

Exercise 1.1 2

Exercise 1.2 2

2 Level 2: Medium **4**

Exercise 2.1 4

Exercise 2.2 5

3 Level 3: Hard **7**

Exercise 3.1 7

Exercise 3.2 8

1 Level 1: Easy

Exercise 1.1 – Stationary Points and Optimality Conditions

Exercise 1.1

The function $f : \mathbb{R}^2 \to \mathbb{R}$ is given:

$$f(x_1, x_2) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2 + 10$$

(a) Calculate the gradient $\nabla f(x_1, x_2)$ and the Hessian matrix $\mathbf{H}(x_1, x_2)$.

(b) Determine all stationary points (set $\nabla f = \mathbf{0}$).

(c) Use the Hessian matrix to check whether the stationary point is a minimum, maximum, or saddle point.

(d) Calculate the function value at the optimum.

Solution: Exercise 1.1

(a) Gradient and Hessian matrix:

$$\nabla f = \begin{pmatrix} 2x_1 - 4 \\ 8x_2 - 8 \end{pmatrix}, \qquad \mathbf{H} = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}$$

(b) Stationary point: $\nabla f = \mathbf{0}$:

$$2x_1 - 4 = 0 \Rightarrow x_1^* = 2, \qquad 8x_2 - 8 = 0 \Rightarrow x_2^* = 1$$

The only stationary point: $\mathbf{x}^* = (2, 1)^\top$.

(c) Nature of the stationary point:

Eigenvalues of $\mathbf{H}$: $\lambda_1 = 2 > 0$, $\lambda_2 = 8 > 0 \Rightarrow \mathbf{H} \succ 0$ (positive definite).

$\Rightarrow \mathbf{x}^* = (2, 1)^\top$ is a **strict global minimum**.

(d) Function value:

$$f(2, 1) = 4 + 4 - 8 - 8 + 10 = 2$$

Optimum: $\mathbf{x}^* = (2, 1)^\top$, $f^* = 2$

Exercise 1.2 – One Step of Gradient Descent

Exercise 1.2

Given $f(x_1, x_2) = (x_1 - 3)^2 + 2(x_2 + 1)^2$ with starting point $\mathbf{x}^{(0)} = (1, 1)^\top$.

(a) Calculate $\nabla f(\mathbf{x}^{(0)})$ and the search direction $\mathbf{d}^{(0)} = -\nabla f(\mathbf{x}^{(0)})$.

(b) Perform one gradient descent step with fixed step size $\alpha = 0.25$.

(c) How many iterations does gradient descent with optimal step size need to reach the minimum? (Justify.)

(d) Calculate the condition number $\kappa = \lambda_{\max}/\lambda_{\min}$ of the Hessian matrix.

Solution: Exercise 1.2

(a) Gradient and search direction:

$$\nabla f = \begin{pmatrix} 2(x_1 - 3) \\ 4(x_2 + 1) \end{pmatrix}, \quad \nabla f(\mathbf{x}^{(0)}) = \begin{pmatrix} 2(1-3) \\ 4(1+1) \end{pmatrix} = \begin{pmatrix} -4 \\ 8 \end{pmatrix}$$

$$\mathbf{d}^{(0)} = -\nabla f(\mathbf{x}^{(0)}) = \begin{pmatrix} 4 \\ -8 \end{pmatrix}$$

(b) Gradient descent step with $\alpha = 0.25$:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + 0.25 \cdot \mathbf{d}^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0.25 \begin{pmatrix} 4 \\ -8 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$f(\mathbf{x}^{(1)}) = (2-3)^2 + 2(-1+1)^2 = 1 + 0 = 1$$

(c) Optimal step size:

The Hessian matrix is $\mathbf{H} = \mathrm{diag}(2, 4)$. Since $\mathbf{H}$ is diagonal and constant, we can optimize each component separately. With optimal step size $\alpha_i^* = 1/\lambda_i$:

$$x_1^{(1)} = 1 + \frac{1}{2} \cdot 4 = 3, \quad x_2^{(1)} = 1 + \frac{1}{4} \cdot (-8) = -1$$

$\mathbf{x}^{(1)} = (3, -1)^\top = \mathbf{x}^*$. Gradient descent with component-wise optimal step size reaches the minimum in **one iteration**.

(d) Condition number:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{4}{2} = 2$$

A condition number of 2 is very good – gradient descent converges quickly.

Optimum: $\mathbf{x}^* = (3, -1)^\top$, $f^* = 0$, $\kappa = 2$

2 Level 2: Medium

Exercise 2.1 – Newton's Method: Two Iterations

Exercise 2.1

The following nonlinear function is given:

$$f(x_1, x_2) = x_1^4 + x_2^4 - 4x_1x_2 + 1$$

with starting point $\mathbf{x}^{(0)} = (2, 2)^\top$.

(a) Calculate ∇f and $\mathbf{H}f$ (in general form).

(b) Perform **two** complete Newton iterations (with step size $\alpha = 1$).

(c) Calculate $\|\nabla f(\mathbf{x}^{(k)})\|$ after each iteration and assess convergence.

(d) Justify why Newton's method converges quickly here.

Solution: Exercise 2.1

(a) Gradient and Hessian matrix:

$$\nabla f = \begin{pmatrix} 4x_1^3 - 4x_2 \\ 4x_2^3 - 4x_1 \end{pmatrix}, \qquad \mathbf{H} = \begin{pmatrix} 12x_1^2 & -4 \\ -4 & 12x_2^2 \end{pmatrix}$$

(b) Newton iteration 1 (at $\mathbf{x}^{(0)} = (2, 2)^\top$):

$$\nabla f(\mathbf{x}^{(0)}) = \begin{pmatrix} 4 \cdot 8 - 8 \\ 4 \cdot 8 - 8 \end{pmatrix} = \begin{pmatrix} 24 \\ 24 \end{pmatrix}, \qquad \mathbf{H}(\mathbf{x}^{(0)}) = \begin{pmatrix} 48 & -4 \\ -4 & 48 \end{pmatrix}$$

$\det(\mathbf{H}^{(0)}) = 48^2 - 16 = 2304 - 16 = 2288$.
Newton system of equations $\mathbf{H}^{(0)}\mathbf{d}^{(0)} = -\nabla f(\mathbf{x}^{(0)})$:

$$\begin{pmatrix} 48 & -4 \\ -4 & 48 \end{pmatrix} \mathbf{d}^{(0)} = \begin{pmatrix} -24 \\ -24 \end{pmatrix}$$

By symmetry (∇f has equal components): $\mathbf{d}^{(0)} = (d, d)^\top$ with $(48 - 4)d = -24$, so $d = -24/44 = -6/11$.

$$\mathbf{d}^{(0)} = \begin{pmatrix} -6/11 \\ -6/11 \end{pmatrix} \approx \begin{pmatrix} -0.5455 \\ -0.5455 \end{pmatrix}$$

$$\mathbf{x}^{(1)} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} + \begin{pmatrix} -6/11 \\ -6/11 \end{pmatrix} = \begin{pmatrix} 16/11 \\ 16/11 \end{pmatrix} \approx \begin{pmatrix} 1.4545 \\ 1.4545 \end{pmatrix}$$

Newton iteration 2 (at $\mathbf{x}^{(1)} \approx (1.4545, 1.4545)^\top$):

$$\nabla f(\mathbf{x}^{(1)}) \approx \begin{pmatrix} 4(1.4545)^3 - 4(1.4545) \\ 4(1.4545)^3 - 4(1.4545) \end{pmatrix} = \begin{pmatrix} 4(3.0787) - 5.818 \\ 4(3.0787) - 5.818 \end{pmatrix} \approx \begin{pmatrix} 6.497 \\ 6.497 \end{pmatrix}$$

$$\mathbf{H}(\mathbf{x}^{(1)}) \approx \begin{pmatrix} 12(2.1156) & -4 \\ -4 & 12(2.1156) \end{pmatrix} = \begin{pmatrix} 25.387 & -4 \\ -4 & 25.387 \end{pmatrix}$$

Newton step: $d = -6.497/(25.387 - 4) = -6.497/21.387 \approx -0.3038$.

$$\mathbf{x}^{(2)} \approx \begin{pmatrix} 1.4545 - 0.3038 \\ 1.4545 - 0.3038 \end{pmatrix} = \begin{pmatrix} 1.1507 \\ 1.1507 \end{pmatrix}$$

(c) Gradient norm:

k	$\mathbf{x}^{(k)}$	$f(\mathbf{x}^{(k)})$	$\|\nabla f(\mathbf{x}^{(k)})\|$
0	$(2, 2)$	$2 \cdot 16 - 4 \cdot 4 + 1 = 17$	$\sqrt{2} \cdot 24 \approx 33.94$
1	$(1.455, 1.455)$	≈ 3.82	≈ 9.19
2	$(1.151, 1.151)$	≈ 1.48	≈ 3.14

Stationary point at $\mathbf{x}^* = (1, 1)^\top$: $4 - 4 = 0$ ✓, $f^* = 2 - 4 + 1 = -1$.
(d) Fast convergence: The function is smooth ($\mathcal{C}^\infty$) and the starting point is relatively close to the minimum. Near the minimum, $\mathbf{H} \succ 0$, so Newton's method converges quadratically.
Optimum: $\mathbf{x}^* = (1, 1)^\top$, $f^* = -1$

Exercise 2.2 – Comparing Gradient Descent and Newton: Condition Number

Exercise 2.2

Consider the function $f(x_1, x_2) = x_1^2 + 100x_2^2$ (a strongly elongated ellipse).

(a) Calculate the gradient, Hessian matrix, and condition number κ.

(b) Perform **three** gradient descent steps with optimal step size $\alpha^* = 2/(\lambda_{\max} + \lambda_{\min})$, starting point $\mathbf{x}^{(0)} = (1, 1)^\top$.

(c) Perform **one** Newton step and compare.

(d) Explain why gradient descent is slow for a large condition number (the zig-zag phenomenon).

Solution: Exercise 2.2

(a) Gradient, Hessian matrix, condition number:

$$\nabla f = \begin{pmatrix} 2x_1 \\ 200x_2 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} 2 & 0 \\ 0 & 200 \end{pmatrix}, \quad \kappa = \frac{200}{2} = 100$$

(b) Gradient descent, $\alpha^* = 2/(200 + 2) = 1/101 \approx 0.0099$:
$k = 0$: $\nabla f(1, 1) = (2, 200)^\top$, $\mathbf{d} = (-2, -200)^\top$.

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{101} \begin{pmatrix} -2 \\ -200 \end{pmatrix} = \begin{pmatrix} 1 - 2/101 \\ 1 - 200/101 \end{pmatrix} \approx \begin{pmatrix} 0.9802 \\ -0.9802 \end{pmatrix}$$

$f(\mathbf{x}^{(1)}) = (0.9802)^2 + 100(0.9802)^2 = 101 \cdot (0.9802)^2 \approx 96.93$. (Initial function value: $1 + 100 = 101$.)

$k = 1$: $\nabla f(\mathbf{x}^{(1)}) \approx (1.960, -196.04)^\top$.

$$\mathbf{x}^{(2)} \approx \begin{pmatrix} 0.9608 \\ 0.9608 \end{pmatrix}, \quad f(\mathbf{x}^{(2)}) \approx 93.97$$

$k = 2$: analogously: $\mathbf{x}^{(3)} \approx (0.9417, -0.9417)^\top$, $f \approx 89.06$.
Reduction factor per iteration: $\left(\frac{\kappa-1}{\kappa+1}\right)^2 = \left(\frac{99}{101}\right)^2 \approx 0.9610$ – very slow convergence!

(c) Newton step (one step):

$$\mathbf{d}^{(0)} = -\mathbf{H}^{-1}\nabla f(\mathbf{x}^{(0)}) = -\begin{pmatrix} 1/2 & 0 \\ 0 & 1/200 \end{pmatrix}\begin{pmatrix} 2 \\ 200 \end{pmatrix} = -\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\mathbf{x}_N^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \mathbf{x}^* \quad \checkmark$$

Newton reaches the minimum in **one step**; gradient descent needs > 10000 iterations!

(d) Zig-zag phenomenon:
For large κ, the level curves are strongly elongated ellipses. The gradient is perpendicular to the level curves and points in the direction of the short ellipse axis (here the x_2-direction, with large $\lambda = 200$). Every gradient descent step overshoots in the x_2-direction and must correct in the next iteration $\Rightarrow$ zig-zag. Newton corrects for the different curvatures via the Hessian matrix.
Newton: 1 iteration. Gradient descent: $\approx$ thousands. Condition number $\kappa = 100$ is problematic.

3 Level 3: Hard

Exercise 3.1 – Saddle Point and Convergence Analysis

Exercise 3.1

The function $f(x_1, x_2) = x_1^2 - x_2^2 + 2x_1x_2$ is given.

(a) Calculate all stationary points.

(b) Classify the stationary point using the Hessian matrix and explain what the sign of $\det(\mathbf{H})$ indicates.

(c) What problems arise when Newton's method is applied to this problem? Calculate the Newton step from $\mathbf{x}^{(0)} = (1, 0)^\top$ and explain the result.

(d) How does gradient descent behave at a saddle point? Formulate a general statement.

Solution: Exercise 3.1

(a) Stationary points:

$$\nabla f = \begin{pmatrix} 2x_1 + 2x_2 \\ -2x_2 + 2x_1 \end{pmatrix} = \mathbf{0} \Rightarrow \begin{cases} x_1 + x_2 = 0 \\ x_1 - x_2 = 0 \end{cases} \Rightarrow x_1 = x_2 = 0$$

The only stationary point: $\mathbf{x}^* = (0, 0)^\top$.

(b) Classification:

$$\mathbf{H} = \begin{pmatrix} 2 & 2 \\ 2 & -2 \end{pmatrix}, \quad \det(\mathbf{H}) = -4 - 4 = -8 < 0$$

Since $\det(\mathbf{H}) < 0$, $\mathbf{H}$ has one positive and one negative eigenvalue $\Rightarrow$ $\mathbf{H}$ is **indefinite** $\Rightarrow$ $(0, 0)$ is a **saddle point** (neither a minimum nor a maximum).

Eigenvalues: $\lambda_{1,2} = \frac{(2-2) \pm \sqrt{(2+2)^2 + 4 \cdot 4}}{2} = \pm\sqrt{4+4} \cdot \frac{1}{1} = \pm 2\sqrt{2}$.

f is unbounded both above *and* below.

(c) Newton step at the saddle point:

$$\nabla f(1, 0) = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \mathbf{H}^{-1} = \frac{1}{-8} \begin{pmatrix} -2 & -2 \\ -2 & 2 \end{pmatrix} = \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & -1/4 \end{pmatrix}$$

$$\mathbf{d} = -\mathbf{H}^{-1} \nabla f = -\begin{pmatrix} 1/4 & 1/4 \\ 1/4 & -1/4 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix} = -\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

$$\mathbf{x}^{(1)} = (1, 0)^\top + (-1, 0)^\top = (0, 0)^\top$$

Newton converges to the saddle point in one step! This is a problem: Newton finds **stationary points** but does not distinguish between minima, maxima, and saddle points. Checking $\det(\mathbf{H})$ is essential.

(d) Gradient descent at saddle points:

Gradient descent theoretically does **not** converge to saddle points (since $-\nabla f$ points in a descent direction and a saddle point is not a local minimum). However,

for certain starting points it can drift slowly toward the saddle point before diverging away from it. In practice: starting exactly on the stable subspace $\rightarrow$ convergence to the saddle point; otherwise, divergence from it.
$(0,0)$ **is a saddle point. Newton finds it in 1 step – but it is not a minimum!**

Exercise 3.2 – Rosenbrock: Numerical Analysis

Exercise 3.2

The Rosenbrock function $f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$ is a classic test problem.

(a) Show that $\mathbf{x}^* = (1,1)^\top$ is the unique global minimum with $f^* = 0$.

(b) Calculate $\mathbf{H}$ at the minimum $\mathbf{x}^* = (1,1)^\top$ and determine the condition number.

(c) Explain why gradient descent is extremely slow on the Rosenbrock function.

(d) Calculate the first Newton step from $\mathbf{x}^{(0)} = (0,0)^\top$ completely.

(e) Compare qualitatively: how many iterations do gradient descent and Newton's method each require?

Solution: Exercise 3.2

(a) Global minimum:
$f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ (a sum of squares). $f(\mathbf{x}) = 0$ if and only if:

$$x_2 - x_1^2 = 0 \quad \text{and} \quad 1 - x_1 = 0 \Rightarrow x_1 = 1,\ x_2 = 1$$

So $(1,1)^\top$ is the **unique global minimum** with $f^* = 0$.
(b) Hessian matrix at the minimum:
In general (from the theory):

$$\mathbf{H} = \begin{pmatrix} -400(x_2 - x_1^2) + 800x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}$$

At $\mathbf{x}^* = (1,1)^\top$: $x_2 - x_1^2 = 0$:

$$\mathbf{H}(\mathbf{x}^*) = \begin{pmatrix} 800 + 2 & -400 \\ -400 & 200 \end{pmatrix} = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$$

$\det(\mathbf{H}^*) = 802 \cdot 200 - 400^2 = 160400 - 160000 = 400 > 0$, trace $= 1002 > 0 \Rightarrow$ $\mathbf{H}^* \succ 0$ (minimum confirmed).
Eigenvalues: characteristic polynomial $(\lambda - 802)(\lambda - 200) - 160000 = 0$:

$$\lambda^2 - 1002\lambda + 400 = 0 \Rightarrow \lambda_{1,2} = \frac{1002 \pm \sqrt{1002^2 - 1600}}{2}$$

$$\lambda_1 \approx 1001.6, \quad \lambda_2 \approx 0.3996$$

$$\kappa(\mathbf{H}^*) = \frac{\lambda_1}{\lambda_2} \approx \frac{1001.6}{0.3996} \approx \mathbf{2506}$$

An extremely high condition number!

(c) Slow convergence of gradient descent:
The Rosenbrock function has a narrow, curved valley along the parabola $x_2 = x_1^2$. The gradient points across the valley (since the level curves are very closely spaced in the valley direction). Gradient descent:

- Approaches the valley floor quickly (descent in the x_2-direction).
- Then progresses extremely slowly along the valley (very small gradients in the valley direction).
- Convergence rate $\approx \left(\frac{2506-1}{2506+1}\right)^2 \approx 0.9984$ – only a 0.16% reduction per step!

(d) Newton step from $\mathbf{x}^{(0)} = (0,0)^\top$:

$$\nabla f(0,0) = \begin{pmatrix} -400 \cdot 0 \cdot (0-0) - 2(1-0) \\ 200(0-0) \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}$$

$$\mathbf{H}(0,0) = \begin{pmatrix} -400(0) + 0 + 2 & 0 \\ 0 & 200 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 200 \end{pmatrix}$$

Newton step:

$$\mathbf{d} = -\mathbf{H}^{-1}\nabla f = -\begin{pmatrix} 1/2 & 0 \\ 0 & 1/200 \end{pmatrix}\begin{pmatrix} -2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\mathbf{x}^{(1)} = (0,0)^\top + 1 \cdot (1,0)^\top = (1,0)^\top$$

$f(1,0) = 100(0-1)^2 + 0 = 100$. Not yet at the minimum – and in fact worse than the start.
Indeed: $f(0,0) = 100 \cdot 0 + 1 = 1$. After the Newton step: $f(1,0) = 100$. Here Newton actually made things worse! The Hessian matrix at the starting point is not sufficiently positive definite for global convergence with a full step.
With a line search: $\alpha < 1$ would reduce f. With $\alpha \approx 0.01$: $\mathbf{x}^{(1)} \approx (0.01, 0)^\top$, $f \approx 0.98$ – a better starting point for subsequent Newton steps.

(e) Iteration comparison (typical):

Method	Typical iterations	Achievable accuracy
Gradient descent	10 000–100 000	$\|\nabla f\| \approx 10^{-3}$
Newton (with line search)	20–50	$\|\nabla f\| \approx 10^{-10}$

$\mathbf{x}^* = (1,1)^\top$, $\kappa \approx 2506$, Newton: ≈ 30 iterations, gradient descent: $> 10\,000$ iterations.