Exercise class 9

Learning Objectives:

- Finding extreme points
- Gradient method

Demo 1: Finding minima and maxima of functions

Find the minima and/or maxima of the following functions.

a) $f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1x_2$

b) $f(x_1, x_2, x_3) = x_1^2(x_1 - 3) + (x_2 - 1)^2 + (x_3 - 1)^2$

Hint. Use the Hessian to verify necessary and sufficient conditions.

a) Solve $\nabla f(x) = 0$ to obtain the stationary points

$$\nabla f(x) = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \end{bmatrix} = \begin{bmatrix} 3x_1^2 - 3x_2 \\ 3x_2^2 - 3x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow (x_1, x_2) = (0,0) \text{ and } (1,1)$$

The Hessian matrix is given by

$$H(x_1, x_2) = \begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 \\ \partial^2 f/\partial x_1 \partial x_2 & \partial^2 f/\partial x_2^2 \end{bmatrix} = \begin{bmatrix} 6x_1 & -3 \\ -3 & 6x_2 \end{bmatrix}$$

The Hessian matrices at the stationary points are given by

$$H(0,0) = \begin{bmatrix} 0 & -3 \\ -3 & 0 \end{bmatrix} \quad \text{and} \quad H(1,1) = \begin{bmatrix} 6 & -3 \\ -3 & 6 \end{bmatrix}$$

Calculate the eigenvalues of Hessian matrices (roots of the polynomial)

$$\det(H(0,0) - \lambda \mathbf{I}) = \det\left(\begin{bmatrix} -\lambda & -3 \\ -3 & -\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow \lambda^2 - 9 = 0 \Rightarrow \lambda_1 = 3; \lambda_2 = -3 \text{ or vice-versa.}$$

Because the eigenvalues are neither all positive nor negative, the Hessian at this point is indefinite, thus $(x, y) = (0, 0)$ is local saddle point.

$$\det(H(1,1) - \lambda \mathbf{I}) = \det\left(\begin{bmatrix} 6-\lambda & -3 \\ -3 & 6-\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow (6-\lambda)^2 - 9 = 0 \Rightarrow \lambda_1 = 9, \lambda_2 = 3 \text{ or vice-versa.}$$

Because of positive eigenvalues, $H(1,1)$ is positive definite and thus $(x_1, x_2) = (1, 1)$ is local minimum.

b) Solve $\nabla f(x) = 0$ to obtain the stationary points:

$$\nabla f(x_1, x_2, x_3) = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \partial f/\partial x_3 \end{bmatrix} = \begin{bmatrix} 3x_1^2 - 6x_1 \\ 2x_2 - 2 \\ 2x_3 - 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow (x_1, x_2, x_3) = (0, 1, 1) \text{ and } (2, 1, 1)$$

The Hessian matrix is given by

$$H(x_1, x_2, x_3) = \begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 & \partial^2 f/\partial x_1 \partial x_3 \\ \partial^2 f/\partial x_2 \partial x_1 & \partial^2 f/\partial x_2^2 & \partial^2 f/\partial x_2 \partial x_3 \\ \partial^2 f/\partial x_3 \partial x_1 & \partial x_3 \partial x_2 & \partial^2 f/\partial x_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} 6x_1 - 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

The Hessian matrices at the stationary points are given by

$$H(0, 1, 1) = \begin{bmatrix} -6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \text{ and } H(2, 1, 1) = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Calculate the eigenvalues of Hessian matrices

$$\det(H(0, 1, 1) - \lambda \mathbf{I}) = \det \left(\begin{bmatrix} -6 - \lambda & 0 & 0 \\ 0 & 2 - \lambda & 0 \\ 0 & 0 & 2 - \lambda \end{bmatrix} \right) = 0$$

$$\Rightarrow (-6 - \lambda)(2 - \lambda)(2 - \lambda) = 0 \Rightarrow \lambda_1 = -6, \lambda_2 = 2 \ \text{ and } \ \lambda_3 = 2.$$

Because the eigenvalues are neither all positive nor negative, H at $(0, 1, 1)$ is indefinite and thus $(0, 1, 1)$ is a saddle point.

$$\det(H(2, 1, 1) - \lambda \mathbf{I}) = \det \left(\begin{bmatrix} 6 - \lambda & 0 & 0 \\ 0 & 2 - \lambda & 0 \\ 0 & 0 & 2 - \lambda \end{bmatrix} \right) = 0$$

$$\Rightarrow (6 - \lambda)(2 - \lambda)(2 - \lambda) = 0 \Rightarrow \lambda_1 = 6, \lambda_2 = 2 \ \text{ and } \ \lambda_3 = 2.$$

Because all eigenvalues are positive, H at $(2, 1, 1)$ is positive definite and thus $(2, 1, 1)$ is a local minimum.

Demo 2: Linear regression using gradient method

Linear regression is a key prediction technique in machine learning and statistics. It consists of obtaining the linear function

$$y = a^\top x + b$$

that best fit some m data points $(x_i, y_i)_{i=1,\ldots,m}$ available for an input x with n features (that is, $x \in \mathbb{R}^n$) and an output y. Then, given a new observation $m + 1$, we can predict y_{m+1} to be

$$\hat{y}_{m+1} = a^\top (x_{m+1}) + b$$

In these applications, the measurement of fitness of the predictor is given by the *accumulated (or sum of) squared error* $f : \mathbb{R}^{n+1} \mapsto \mathbb{R}$ for the predictions obtained for a given (a, b) for x_i and the observed y_i, for $i = 1, \ldots, m$. Notice that it simply amounts to the difference between the prediction $\hat{y}$ and the actual observation y, squared to compensate for positive and negative deviations. That is

$$f(a,b) = \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n} a_j x_{ij} + b_i \right) - y_i \right]^2 = \sum_{i=1}^{m} e_i^2 = e^\top e = ||e||_2^2.$$

Finding the best fitting (a, b) can be achieved by employing optimisation to find (a, b) that minimise the accumulate squared error, a method that is commonly referred to as the *least squared error* (LSE) estimation.

Given the data below (with $m = 7$ and $n = 1$), estimate the parameters a and b of estimate $y = ax + b$ using the LSE estimation. To find the optimal parameters, minimise the squared error function f using the gradient method, with starting point $(a, b) = (0, 0)$ and step size $\lambda = 0.01$. Use a tolerance $|\nabla f(a_k, b_k)| \leq 0.01$.

x_i	0	1	2	3	4	5	6
y_i	1	3	1.5	4	6.5	5	8

Solution

The estimate lies on the predictor (that is, the line) defined as $\hat{y} = ax + b$. The error between estimate $\hat{y}$ and the real value y is e. The objective is to minimise the squared error, so the predictor is fitted according to the data values

$$\min.\ f(a,b) = \sum_{i=0}^{6} e_i^2 = \sum_{i=0}^{6} (\hat{y}_i - y_i)^2 = \sum_{i=0}^{6} (ax_i + b - y_i)^2.$$

Let us define the following matrix, so we can work with more compact matrix notation:

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \\ 6 & 1 \end{bmatrix},\ y = \begin{bmatrix} 1 \\ 3 \\ 1.5 \\ 4 \\ 6.5 \\ 5 \\ 8 \end{bmatrix},\ \alpha = \begin{bmatrix} a \\ b \end{bmatrix}$$

We artificially append a columns of ones to the right of the features x_i, $i = 1, \ldots, n$, so we can represent our parameters (a, b) as a vector $\alpha \in \mathbb{R}^{n+1}$ and our predictions as the $m \times 1$ matrix $X\alpha$. That means that our error $e \in \mathbb{R}^m$ is the vector given by $e = X\alpha - y$ and that we need to minimise the accumulated squared error given by

$$f(a,b) = e^\top e = (X\alpha - y)^\top (X\alpha - y) = \alpha^\top X^\top X\alpha - 2y^\top X\alpha + y^\top y$$

In order to compute the gradient, we apply the following differentiation rules:

1. $\nabla(a^\top x) = a$ (note that the multiplier $-2y^\top X$ is a row vector)
2. $\nabla(x^\top Ax) = A^\top x + Ax$ (note that $X^\top X$ is symmetric and thus $(X^\top X)^\top\alpha + X^\top X\alpha = 2X^\top X\alpha$)

The gradient of $f(a,b)$ is therefore given by $\nabla f(a,b) = \nabla f(\alpha) = 2X^\top X\alpha - 2X^\top y$, which is equal to

$$2\underbrace{\begin{bmatrix} 91 & 21 \\ 21 & 7 \end{bmatrix}}_{X^\top X}\begin{bmatrix} a_k \\ b_k \end{bmatrix} - 2\underbrace{\begin{bmatrix} 117 \\ 29 \end{bmatrix}}_{X^\top y}$$

To solve this minimisation problem, we have to employ the gradient method, which consists of repeatedly taking steps of the form

$$(a_{k+1}, b_{k+1}) = (a_k, b_k) - \lambda \nabla f(a_k, b_k),$$

which is the same as solving the following recursion:

$$\begin{bmatrix} a_{k+1} \\ b_{k+1} \end{bmatrix} = \begin{bmatrix} a_k \\ b_k \end{bmatrix} - 0.01\left(2\begin{bmatrix} 91 & 21 \\ 21 & 7 \end{bmatrix}\begin{bmatrix} a_k \\ b_k \end{bmatrix} - 2\begin{bmatrix} 117 \\ 29 \end{bmatrix}\right)$$

The solution is $a = 1.072$ and $b = 0.926$, which should be obtained after approximately 140 iterations.

Problem 1: Finding minima and maxima of functions

Find the minima and/or maxima of the following functions.

a) $f(x_1, x_2) = x_1^3(x_1 - 4) + (x_2 - 5)^2$

b) $f(x_1, x_2, x_3) = (1 - x_2)(1 - x_3) + x_1^2 - 1$

Hint. Use the Hessian.

Solution

a) The first-order conditions are satisfied by stationary points

$$\nabla f(x_1, x_2) = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \end{bmatrix} = \begin{bmatrix} 4x_1^3 - 12x_1^2 \\ 2x_2 - 10 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow (x_1, x_2) = (3, 5) \text{ and } (0, 5).$$

Form the Hessian matrix

$$H(x_1, x_2) = \begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 \\ \partial^2 f/\partial x_1 \partial x_2 & \partial^2 f/\partial x_2^2 \end{bmatrix} = \begin{bmatrix} 12x_1^2 - 24x_1 & 0 \\ 0 & 2 \end{bmatrix}$$

Calculate the values of Hessian matrices for the stationary points

$$H(3, 5) = \begin{bmatrix} 36 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad H(0, 5) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$$

Solve the eigenvalues of Hessian matrix

$$\det(H(3, 5) - \lambda \mathbf{I}) = \det\left(\begin{bmatrix} 36 - \lambda_1 & 0 \\ 0 & 2 - \lambda_2 \end{bmatrix}\right) = 0$$

$$\Rightarrow (36 - \lambda_1)(2 - \lambda_2) = 0 \Rightarrow \lambda_1 = 36 \text{ and } \lambda_2 = 2$$

and

$$\det(H(0, 5) - \lambda \mathbf{I}) = \det\left(\begin{bmatrix} -\lambda_1 & 0 \\ 0 & 2 - \lambda_2 \end{bmatrix}\right) = 0$$

$$\Rightarrow (-\lambda_1)(2 - \lambda_2) = 0 \Rightarrow \lambda_1 = 0 \text{ and } \lambda_2 = 2$$

Because the eigenvalues are non-negative, $(x_1, x_2) = (3, 5)$ and $(x_1, x_2) = (0, 5)$ are local minima.

b) The first-order conditions are satisfied by stationary points

$$\nabla f(x_1, x_2, x_3) = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \partial f/\partial x_3 \end{bmatrix} = \begin{bmatrix} 2x_1 \\ x_3 - 1 \\ x_2 - 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow (x_1, x_2, x_3) = (0, 1, 1)$$

Form the Hessian matrix

$$H(x_1, x_2, x_3) = \begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 & \partial^2 f/\partial x_1 \partial x_3 \\ \partial^2 f/\partial x_2 \partial x_1 & \partial^2 f/\partial x_2^2 & \partial^2 f/\partial x_2 \partial x_3 \\ \partial^2 f/\partial x_3 \partial x_1 & \partial x_3 \partial x_2 & \partial^2 f/\partial x_3^2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Calculate the eigenvalues of Hessian matrix

$$\det(H(0, -1, -1) - \lambda \mathbf{I}) = \det\left(\begin{bmatrix} 2 - \lambda & 0 & 0 \\ 0 & -\lambda & 1 \\ 0 & 1 & -\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow (2 - \lambda)((-\lambda)(-\lambda) - 1^2) = 0 \Rightarrow (2 - \lambda)(\lambda^2 - 1) = 0$$

$$\Rightarrow \lambda_1 = 2, \lambda_2 = 1 \text{ and } \lambda_3 = -1.$$

Because the eigenvalues are all neither negative nor positive, the point (0,-1,-1) is a saddle point.

Problem 2: Extreme points

Determine the nature of the extreme points of the following function:

$$f(\mathbf{x}) = 2x_1^2 + x_2^2 + x_3^2 + 6(x_1 + x_2 + x_3) + 2x_1x_2x_3$$

Examine the points $(1,\ -4.2,\ 1.2)$, $(1,\ 1.2,\ -4.2)$, and $(-2.82,\ 1.65,\ 1.65)$.

Solution

First form the Hessian:

$$H(x_1, x_2, x_3) = \quad = \begin{bmatrix} 4 & 2x_3 & 2x_2 \\ 2x_3 & 2 & 2x_1 \\ 2x_2 & 2x_1 & 2 \end{bmatrix}$$

Calculate the eigenvalues of the Hessian at the given points.

First, point (1,-4.2,1.2):

$$\det(H(1, -4.2, 1.2) - \lambda\mathbf{I}) = \det\left(\begin{bmatrix} 4-\lambda & 2.4 & -8.4 \\ 2.4 & 2-\lambda & 2 \\ -8.4 & 2 & 2-\lambda \end{bmatrix}\right) = 0$$

$$\begin{aligned}
&\Rightarrow (4-\lambda)(\lambda^2 - 4\lambda) - 2.4(-2.4\lambda + 21.6) - 8.4(-8.4\lambda + 21.6) = 0 \\
&\Rightarrow -\lambda^3 + 8\lambda^2 + 60.32\lambda - 233.28 = 0 \\
&\Rightarrow \lambda_1 = 11.48, \lambda_2 = -6.57 \text{ and } \lambda_3 = 3.09.
\end{aligned}$$

Thus, the Hessian at this point is indefinite (not all negative/positive eigenvalues), and the point (1,-4.2,1.2) is a saddle point.

Similarly, the Hessian at (1,1.2,-4.2) gives the same eigenvalues, and thus is also a saddle point.

Finally, the point (-2.82,1.65,1.65):

$$\det(H(1, -4.2, 1.2) - \lambda\mathbf{I}) = \det\left(\begin{bmatrix} 4-\lambda & 2(1.65) & 2(1.65) \\ 2(1.65) & 2-\lambda & 2(-2.82) \\ 2(1.65) & 2(-2.82) & 2-\lambda \end{bmatrix}\right) = 0$$

$$\begin{aligned}
&\Rightarrow -\lambda^3 + 8\lambda^2 + 33.5896\lambda - 277.6376 = 0 \\
&\Rightarrow \lambda_1 = 7.64, \lambda_2 = -5.85 \text{ and } \lambda_3 = 6.21.
\end{aligned}$$

Once again, the Hessian at this point is indefinite, thus the point (-2.82,1.65,1.65) is a saddle point.

Problem 3: Stationary and extreme points

Verify that the function

$$f(x_1, x_2, x_3) = 2x_1x_2x_3 - 4x_1x_3 - 2x_2x_3 + x_1^2 + x_2^2 + x_3^2 - 2x_1 - 4x_2 + 4x_3$$

has the stationary points (0, 3, 1), (0, 1, -1), (1, 2, 0), (2, 1, 1), and (2, 3, -1). Use the sufficiency condition to identify the extreme points.

Solution

First derivatives are:

$$\frac{\partial f}{\partial x_1}(x_1, x_2, x_3) = 2x_2x_3 - 4x_3 + 2x_1 - 2$$

$$\frac{\partial f}{\partial x_2}(x_1, x_2, x_3) = 2x_1x_3 - 2x_3 + 2x_2 - 4$$

$$\frac{\partial f}{\partial x_3}(x_1, x_2, x_3) = 2x_1x_2 - 4x_1 - 2x_2 + 2x_3 + 4$$

Form the Hessian matrix

$$H(x_1, x_2, x_3) = \begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1\partial x_2 & \partial^2 f/\partial x_1\partial x_3 \\ \partial^2 f/\partial x_2\partial x_1 & \partial^2 f/\partial x_2^2 & \partial^2 f/\partial x_2\partial x_3 \\ \partial^2 f/\partial x_3\partial x_1 & \partial x_3\partial x_2 & \partial^2 f/\partial x_3^2 \end{bmatrix}$$

The Hessian for this function is:

$$H = \begin{bmatrix} 2 & 2x_3 & 2x_2 - 4 \\ 2x_3 & 2 & 2x_1 - 2 \\ 2x_2 - 4 & 2x_1 - 2 & 2 \end{bmatrix}$$

First, the point (0,3,1):

$$\det(H(0,3,1) - \lambda\mathbf{I}) = \det\left(\begin{bmatrix} 2-\lambda & 2 & 2 \\ 2 & 2-\lambda & -2 \\ 2 & -2 & 2-\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow (2-\lambda)^3 - 12(2-\lambda) - 16 = 0$$
$$\Rightarrow \lambda_1 = -2, \lambda_2 = 4 \text{ and } \lambda_3 = 4$$

As the eigenvalues are not neither all nonnegative nor nonpositive, the Hessian is indefinite and (0,3,1) is a saddle point.

Then, the point (0,1,-1):

$$\det(H(0,1,-1) - \lambda\mathbf{I}) = \det\left(\begin{bmatrix} 2-\lambda & -2 & -2 \\ -2 & 2-\lambda & -2 \\ -2 & -2 & 2-\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow (2-\lambda)^3 - 12(2-\lambda) - 16 = 0$$
$$\Rightarrow \lambda_1 = -2, \lambda_2 = 4 \text{ and } \lambda_3 = 4$$

As the eigenvalues are not neither all nonnegative nor nonpositive, the Hessian is indefinite and (0,1,-1) is a saddle point.

Then, the point (1,2,0):

$$\det(H(1,2,0) - \lambda\mathbf{I}) = \det\left(\begin{bmatrix} 2-\lambda & 0 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow (2-\lambda)^3 = 0$$
$$\Rightarrow \lambda_1 = 2, \lambda_2 = 2 \text{ and } \lambda_3 = 2$$

As the eigenvalues are all positive, the Hessian is positive definite and (1,2,0) is a minimal point.

Then, the point (2,1,1):

$$\det(H(0,1,-1)-\lambda\mathbf{I}) = \det\left(\begin{bmatrix} 2-\lambda & 2 & -2 \\ 2 & 2-\lambda & 2 \\ -2 & 2 & 2-\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow (2-\lambda)^3 - 12(2-\lambda) - 16 = 0$$
$$\Rightarrow \lambda_1 = -2, \lambda_2 = 4 \text{ and } \lambda_3 = 4$$

As the eigenvalues are not neither all nonnegative nor nonpositive, the Hessian is indefinite and (1,2,0) is a saddle point.

Then, the point (2,3,-1):

$$\det(H(0,1,-1)-\lambda\mathbf{I}) = \det\left(\begin{bmatrix} 2-\lambda & -2 & 2 \\ -2 & 2-\lambda & 2 \\ 2 & 2 & 2-\lambda \end{bmatrix}\right) = 0$$

$$\Rightarrow (2-\lambda)^3 - 12(2-\lambda) - 16 = 0$$
$$\Rightarrow \lambda_1 = -2, \lambda_2 = 4 \text{ and } \lambda_3 = 4$$

As the eigenvalues are not neither all nonnegative nor nonpositive, the Hessian is indefinite and (2,3,-1) is a saddle point.

Problem 4: The Gradient method

Calculate by hand the first two steps (x_1 and x_2) of the gradient method for the minimization of the function f. Initial value is $x_0 = (0,0)$. Compute optimal step sizes at each iteration.

$$f(x_1, x_2) = (1-x_1)^2 + (1-x_2-x_1)^2$$

Hint. The optimal step size can be obtained from first-order optimality conditions, namely $\min\limits_{\alpha\in\mathbb{R}} f(x_{k+1}) = \min\limits_{\alpha\in\mathbb{R}} f(x_k - \alpha\nabla f(x_k))$.

Solution

In the gradient method you move a step (sized α_k) from the iteration point in the direction of the steepest descent.

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \text{ in which } x_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix}$$

The gradient of the function is

$$\nabla f(x_k) = \begin{bmatrix} -2(1-x_k) - 2(1-y_k-x_k) \\ -2(1-y_k-x_k) \end{bmatrix} = \begin{bmatrix} -4+4x_k+2y_k \\ -2+2y_k+2x_k \end{bmatrix}$$

The value of the gradient in the point $(x_0, y_0) = (0,0)$ is

$$\nabla f\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} -4 \\ -2 \end{bmatrix}.$$

The optimal step size α_0 is given by

$$\min_\alpha f\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \alpha\begin{bmatrix} -4 \\ -2 \end{bmatrix}\right) = f(4\alpha, 2\alpha) = (1-4\alpha)^2 + (1-6\alpha)^2$$
$$\Rightarrow f'(\alpha) = -8(1-4\alpha) - 12(1-6\alpha) = -20 + 104\alpha = 0 \Rightarrow \alpha = \tfrac{5}{26}.$$

Next iteration point is

$$x_1 = x_0 - \alpha_0 \nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \frac{5}{26} \begin{bmatrix} -4 \\ -2 \end{bmatrix} = \begin{bmatrix} 10/13 \\ 5/13 \end{bmatrix}$$

The second iteration is carried out in the same way, giving (note, values may be slightly different due to rounding but within our tolerance):

k	x y	$\nabla f(x, y)$	α
0	0.000 0.000	-4.000 -2.000	0.192
1	0.769 0.385	-0.154 0.308	1.250
2	0.962 -0.000	-0.154 -0.077	0.192

Problem 5: Analytical LSE estimation*

Linear regression, as presented in Demo 2, can be alternatively performed by finding a point $\alpha = (a, b) \in \mathbb{R}^{n+1}$ that satisfies the optimality conditions of the accumulated squared error function $f(\alpha) = e^\top e$, where e is defined as in Demo 2.

Formulate the minimisation problem for the LSE estimation in a general manner and provide its optimality conditions.

Hint. You might need the following differentiation rules:

1. $\nabla(a^\top x) = a$
2. $\nabla(x^\top A x) = A^\top x + Ax$

Solution

Form a matrix

$$X = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix},$$

so you can calculate the estimates from $\hat{y} = X\alpha$, in which $\hat{y}$ are the estimates for each input value

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}.$$

Thus the error is $e = \hat{y} - y = X\alpha - y$. The sum of squared errors $f(\alpha)$ is

$$f(\alpha) = \sum_{i=1}^{n} e_k^2 = e^\top e = (\hat{y} - y)^\top (\hat{y} - y) = (X\alpha - y)^{\mathrm{T}} (X\alpha - y)$$

The objective function is

$$\min_{\alpha} e^\top e$$

The first degree optimality condition is

$$\nabla_\alpha e^\top e = 0.$$

Replace the e with $X\alpha - y$:

$$
\begin{aligned}
\nabla f(\alpha) = \nabla \left[(X\alpha - y)^\top (X\alpha - y)\right] &= \nabla \left[(\alpha^\top X^\top - y^\top)(X\alpha - y)\right] \\
&= \nabla \left[\alpha^\top X^\top X\alpha - y^\top X\alpha - \alpha^\top X^\top y - y^\top y\right]
\end{aligned}
$$

Because $y^\top X\alpha$ is scalar, it is the same as its transpose: $y^\top X\boldsymbol{\alpha} = (y^\top X\alpha)^\top = \alpha^\top X^\top y$. This can be replaced in the above, yielding

$$
\begin{aligned}
\nabla f(\alpha) &= \nabla \left[\alpha^\top X^\top X\alpha - 2\alpha^\top X^\top y - y^\mathrm{T} y\right] \\
&= 2X^\top X\alpha - 2X^\top y
\end{aligned}
$$

and the first-order conditions are given by:

$$
\begin{array}{rcl}
2X^\top X\alpha - 2X^\top y & = & 0 \\
X^\top X\alpha & = & X^\top y \\
\alpha & = & (X^\top X)^{-1} X^\top y
\end{array}
$$

Now the optimal parameters α are solved. Notice that the matrix $X^\top X$ has to have an inverse matrix so that the estimate can be calculated. This applies if columns of the matrix X are independent. Also, notice that first-order conditions are also sufficient since the Hessian is given by $2X^\top X$ which is positive definite. We know that the Hessian is positive definite as the columns of X are linearly independent, X has full column rank. So for any non-zero vector v we can define $y := Xv \neq 0$ as a linear combination of the columns of X. Thus, we have:

$$
\begin{aligned}
v^\top X^\top Xv &= (Xv)^\top Xv \\
&= y^\top y = \sum_i y_i^2 > 0
\end{aligned}
$$

Therefore, the Hessian is positive definite.

Home Exercise 9: Gradient method with line search

Perform one iteration of the gradient method to solve

$$\max .f(x_1, x_2) = 2x_1x_2 + 2x_2 - x_1^2 - 2x_2^2$$

from the initial point $x_0 = (0.5, 0.5)$. Use the bisection method to find the optimal step size with interval [0,2] and tolerance $\varepsilon = 0.01$. Is the new point obtained optimal (considering the tolerance of $\varepsilon = 0.01$)?

Hint. Do it by hand and notice it is a maximisation.

Solution

The gradient of f is given by $\nabla f(x) = \begin{bmatrix} 2x_2 - 2x_1 \\ 2x_1 - 4x_2 + 2 \end{bmatrix}$.

At $x_0 = (0.5, 0.5)$, $\nabla f(x_0) = [0, 1]^\top$.

To find the optimal step size, we need to use the bisection method to maximise in λ:

$$\begin{aligned}
\bar{\lambda} &= argmax.\{f(x_0 + \lambda \nabla f(x_0))\} \\
&= f(0.5, 0.5 + \lambda) \\
&= 2(0.5)(0.5 + \lambda) + 2(0.5 + \lambda) - (0.5)^2 - 2(0.5 + \lambda)^2 \\
&= 0.75 + \lambda - 2\lambda^2
\end{aligned}$$

To maximise, we have $\bar{\lambda}' = 1 - 4\lambda$, we have:

1. For interval $[a_0, b_0] = [0, 2] : \lambda_0 = (2 + 0)/2 = 1 \Rightarrow \bar{\lambda}' = -3$
2. For the next interval, as we are maximising, $[a_0, \lambda_0] = [0, 1] : \lambda_1 = (0 + 1)/2 = 0.5 \Rightarrow \bar{\lambda}' = -1$
3. For the following interval $[a_1, \lambda_1] = [0, 0.5] : \lambda_2 = 0.25 \Rightarrow \bar{\lambda}' = 0$, thus λ_2 is the maximiser.

Optimal step size is $\bar{\lambda} = 0.25$.

The gradient step is: $x_1 = x_0 + \bar{\lambda} \nabla f(x_0) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} + 0.25 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.75 \end{bmatrix}$

The gradient at x_1 is $\nabla f(x_1) = [0.5, 0]^\top$, with norm $||\nabla f(x_1)|| = 0.5$. Thus, x_1 is not optimal. It takes roughly 17 iterations to reach the optimal $[2, 2]$ under this setting.