Problem 1: Write down the gradient descent update rule for the objective function

$$J(\theta) = \theta^4 - 6\theta^3 + 11\theta^2 - 7\theta + 4$$

from class. Suppose the learning rate is α.

The update rule is

$$\theta := \theta - \alpha J'(\theta) = \theta - \alpha(4\theta^3 - 18\theta^2 + 22\theta - 7).$$

The recurrence relation is

$$\theta_{t+1} = \theta_t - \alpha J'(\theta_t) = \theta_t - \alpha(4\theta_t^3 - 18\theta_t^2 + 22\theta_t - 7),$$

for $t \geq 0$.

Problem 2: Consider the affine objective function

$$J(\theta) = m\theta + b,$$

for some parameters $m \neq 0$ and $b \in \mathbb{R}$.

(a) Write down the gradient descent update rule. Suppose the learning rate is α.

The update rule is

$$\theta := \theta - \alpha m.$$

The recurrence relation is

$$\theta_{t+1} = \theta_t - \alpha m,$$

for $t \geq 0$.

(b) Find a closed form expression for θ_t.

We have

$$\theta_t = \theta_0 - \alpha m t$$

for all $t \geq 1$.

(c) Using your answer to (b), discuss convergence of the gradient descent algorithm.

We have $\theta_t \to -\infty$ as $t \to \infty$.

Problem 3: Consider the quadratic objective function

$$J(\theta) = \theta^2.$$

(a) Write down the gradient descent update rule. Suppose the learning rate is α.

The update rule is

$$\theta := \theta - 2\alpha\theta.$$

The recurrence relation is

$$\theta_{t+1} = \theta_t - 2\alpha\theta_t,$$

for $t \geq 0$.

(b) Find a closed form expression for θ_t.

We have

$$\theta_t = (1 - 2\alpha)^t \theta_0.$$

(c) Using your answer to (b), discuss convergence of the gradient descent algorithm.

We have $\theta_t \to 0$ exponentially fast provided $|1-2\alpha| < 1$, which occurs if and only if $\alpha < 1$; the value θ_t orbits back and forth between $-\theta_0$ and $+\theta_0$ if $\alpha = 1$; the algorithm diverges to ∞ if $\alpha > 1$.

Problem 4: Consider again the affine objective function

$$J(\theta) = m\theta + b,$$

for some parameters $m \neq 0$ and $b \in \mathbb{R}$.

(a) Write down the gradient descent update rule with learning rate α and decay rate β.

The t-th update rule is

$$\theta := \theta - \alpha m(1-\beta)^{t+1}.$$

The recurrence relation is

$$\theta_{t+1} = \theta_t - \alpha m(1-\beta)^{t+1},$$

for $t \geq 0$.

(b) Find a closed form expression for θ_t.

Setting $\gamma = 1 - \beta$ for convenience, we have

$$\theta_t = \theta_0 - \alpha m \sum_{k=1}^{t} \gamma^k,$$

for $t \geq 1$. But

$$\sum_{k=1}^{t} \gamma^k = \frac{\gamma - \gamma^{t+1}}{1-\gamma},$$

and so

$$\theta_t = \theta_0 - \alpha m \left(\frac{\gamma - \gamma^{t+1}}{1-\gamma} \right),$$

for $t \geq 1$.

(c) Using your answer to (b), discuss convergence of the gradient descent algorithm.

As we saw in Problem 2, the algorithm diverges if $\beta = 0$. But if $\beta > 0$, then $\gamma < 1$ and

$$\lim_{t\to\infty} \theta_t = \theta_0 - \alpha m \left(\frac{\gamma}{1-\gamma} \right) = \theta_0 - \alpha m \left(\frac{1-\beta}{\beta} \right).$$

Thus, the algorithm converges (but not to a minimizer!) if the decay rate β is positive.

Problem 5: Consider the function

$$J : \mathbb{R}^2 \to \mathbb{R}, \quad J(\boldsymbol{\theta}) = 2\theta_1^2 + 4\theta_1\theta_2 + 3\theta_2^2 - 8\theta_1 - 10\theta_2 + 9.$$

(a) Compute the directional first derivative $J'_{\mathbf{v}}(1,1)$ where $\mathbf{v}^\intercal = (1,0)$.

We compute:

$$\begin{aligned} J'_{\mathbf{v}}(1,1) &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} J(1+t,1) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \left[2(1+t)^2 - 4(1+t) + 2\right] \\ &= [4(1+t) - 4]\Big|_{t=0} \\ &= 0 \end{aligned}$$

(b) Compute the directional second derivative $J''_{\mathbf{v}}(1,1)$ where $\mathbf{v}^\intercal = (1,0)$.

We compute:

$$\begin{aligned} J''_{\mathbf{v}}(1,1) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}\bigg|_{t=0} J(1+t,1) \\ &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}\bigg|_{t=0} \left[2(1+t)^2 - 4(1+t) + 2\right] \\ &= 4\Big|_{t=0} \\ &= 4 \end{aligned}$$

Problem 6: Re-do the previous problem, but use the relationship between directional first and second derivatives and gradient vectors and Hessian matrices.

Let's first compute the gradient vector and Hessian matrix:

$$\nabla J(\boldsymbol{\theta}) = \begin{bmatrix} 4\theta_1 + 4\theta_2 - 8 \\ 6\theta_2 + 4\theta_1 - 10 \end{bmatrix}, \quad \nabla^2 J(\boldsymbol{\theta}) = \begin{bmatrix} 4 & 4 \\ 4 & 6 \end{bmatrix}.$$

Then:

$$J'_{\mathbf{v}}(1,1) = \mathbf{v}^\intercal \nabla J(1,1) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$$

and

$$J''_{\mathbf{v}}(1,1) = \mathbf{v}^\intercal \left(\nabla^2 J(1,1)\right) \mathbf{v} = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 4 & 4 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 4.$$

Problem 7: Consider again the function J from Problem 5. Here it is, for reference:

$$J : \mathbb{R}^2 \to \mathbb{R}, \quad J(\boldsymbol{\theta}) = 2\theta_1^2 + 4\theta_1\theta_2 + 3\theta_2^2 - 8\theta_1 - 10\theta_2 + 9.$$

(a) Find the direction of maximum rate of change of J at the point $\boldsymbol{\theta}^\intercal = (0,0)$. What is the rate of change in this direction?

Theorem 11.2 tells us that the direction of maximum rate of change is in the direction of the gradient vector

$$\nabla J(0,0) = \begin{bmatrix} -8 \\ -10 \end{bmatrix}.$$

The rate of change in this direction is given by the directional first derivative $J'_{\mathbf{v}}(0,0)$, where $\mathbf{v}$ is the unit vector that points in the same direction as the gradient:

$$\mathbf{v} = \frac{\nabla J(0,0)}{|\nabla J(0,0)|}.$$

But then

$$J'_{\mathbf{v}}(0,0) = \mathbf{v}^{\intercal}\nabla J(0,0) = \frac{\nabla J(0,0)^{\intercal}\nabla J(0,0)}{|\nabla J(0,0)|} = \frac{|\nabla J(0,0)|^2}{|\nabla J(0,0)|} = |\nabla J(0,0)| = \sqrt{8^2+10^2} \approx 12.8.$$

(b) Find the direction of minimum rate of change of J at the point $\boldsymbol{\theta}^{\intercal} = (0,0)$. What is the rate of change in this direction?

Theorem 11.2 tells us that the direction of minimum rate of change is in the direction of the negative gradient vector

$$-\nabla J(0,0) = \begin{bmatrix} 8 \\ 10 \end{bmatrix}.$$

The rate of change in this direction is given by the directional first derivative $J'_{\mathbf{v}}(0,0)$, where $\mathbf{v}$ is the unit vector that points in the same direction as the negative gradient:

$$\mathbf{v} = -\frac{\nabla J(0,0)}{|\nabla J(0,0)|}.$$

But then

$$J'_{\mathbf{v}}(0,0) = \mathbf{v}^{\intercal}\nabla J(0,0) = -\frac{\nabla J(0,0)^{\intercal}\nabla J(0,0)}{|\nabla J(0,0)|} = -\frac{|\nabla J(0,0)|^2}{|\nabla J(0,0)|} = -|\nabla J(0,0)| = -\sqrt{8^2+10^2} \approx -12.8.$$

Problem 8: Yet again, consider the function J from Problem 5:

$$J : \mathbb{R}^2 \to \mathbb{R}, \quad J(\boldsymbol{\theta}) = 2\theta_1^2 + 4\theta_1\theta_2 + 3\theta_2^2 - 8\theta_1 - 10\theta_2 + 9.$$

Find and classify all extremizers of J.

We first solve the stationarity equation $\nabla J(\theta_1, \theta_2) = \mathbf{0}$ for $\boldsymbol{\theta}$, which is

$$\begin{bmatrix} 4\theta_1 + 4\theta_2 - 8 \\ 6\theta_2 + 4\theta_1 - 10 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The solution is $(\boldsymbol{\theta}^{\star})^{\intercal} = (1,1)$. Then, we consider the Hessian matrix:

$$\nabla^2 J(1,1) = \begin{bmatrix} 4 & 4 \\ 4 & 6 \end{bmatrix}.$$

Its eigenvalues are $5 \pm \sqrt{17}$, which are both positive. Hence it is positive definite, so by the Second Derivative Test, the point $\boldsymbol{\theta}^{\star}$ is a minimizer of J. (It is in fact the global minimizer.)

Problem 9: For the fifth time, consider the function J from Problem 5:

$$J : \mathbb{R}^2 \to \mathbb{R}, \quad J(\boldsymbol{\theta}) = 2\theta_1^2 + 4\theta_1\theta_2 + 3\theta_2^2 - 8\theta_1 - 10\theta_2 + 9.$$

Find the directions of extreme curvature at the point $\boldsymbol{\theta} = (0,0)$. What are the curvatures in these directions?

Theorem 11.4 tells us that the directions of extreme curvature are given by the eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$ corresponding to the eigenvalues

$$\lambda_1 = 5 - \sqrt{17} \approx 0.88 \quad \text{and} \quad \lambda_2 = 5 + \sqrt{17} \approx 9.12.$$

Using technology, we compute these eigenvectors

$$\mathbf{e}_1 \approx \begin{bmatrix} -0.79 \\ 0.62 \end{bmatrix}, \quad \mathbf{e}_2 \approx \begin{bmatrix} -0.62 \\ -0.79 \end{bmatrix}.$$

The theorem also tells us that the curvatures are the eigenvalues themselves.

Problem 10: One last time, consider the function J from Problem 5:

$$J : \mathbb{R}^2 \to \mathbb{R}, \quad J(\boldsymbol{\theta}) = 2\theta_1^2 + 4\theta_1\theta_2 + 3\theta_2^2 - 8\theta_1 - 10\theta_2 + 9.$$

Set $\mathbf{H} = \nabla^2 J(0,0)$. Compute the spectrum $\sigma(\mathbf{H})$, the spectral radius $\rho(\mathbf{H})$, and the condition number $\kappa(\mathbf{H})$.

We already computed the spectrum of the Hessian to be (approximately) $\{0.88, 9.12\}$. Thus, the spectral radius and condition number are given by

$$\rho(\mathbf{H}) \approx 9.12, \quad \kappa(\mathbf{H}) \approx \frac{9.12}{0.88} \approx 10.36.$$

Problem 11: Consider the objective function

$$J : \mathbb{R}^2 \to \mathbb{R}, \quad J(\boldsymbol{\theta}) = 2\theta_1^2 + 4\theta_2^2.$$

(a) Write down the gradient descent update rule. Suppose the learning rate is α, while the decay rate is $\beta = 0$.

We have

$$\nabla J(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta},$$

where

$$\mathbf{H} = \begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix}.$$

Thus, the update rule is

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha\mathbf{H}\boldsymbol{\theta}.$$

In the form of a recurrence relation, this is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\mathbf{H}\boldsymbol{\theta}_t.$$

(b) Find a closed form expression for $\boldsymbol{\theta}_t$.

We have

$$\boldsymbol{\theta}_t = (I - \alpha\mathbf{H})^t\boldsymbol{\theta}_0 = \begin{bmatrix} (1-4\alpha)^t & 0 \\ 0 & (1-8\alpha)^t \end{bmatrix} \boldsymbol{\theta}_0$$

for all $t \geq 1$.

(c) Using your answer to (b), discuss convergence of the gradient descent algorithm.

For all $t \geq 1$, we have

$$(\boldsymbol{\theta}_t)_1 = (1-4\alpha)^t(\boldsymbol{\theta}_0)_1 \quad \text{and} \quad (\boldsymbol{\theta}_t)_2 = (1-8\alpha)^t(\boldsymbol{\theta}_0)_2.$$

Then the algorithm will converge to $\boldsymbol{\theta}^\star = (0,0)$ if and only if both $|1-4\alpha| < 1$ and $|1-8\alpha| < 1$. But this happens if and only if $0 < \alpha < 1/4$.

Problem 12: Consider the stochastic objective function

$$J : \mathbb{R}^2 \to \mathbb{R}, \quad J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}|\mathbf{x}_i - \boldsymbol{\theta}|^2$$

from class, where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^2$ is an observed dataset.

(a) Compute the update rule in the batch gradient descent algorithm with learning rate α and decay rate β.

Note that

$$\frac{1}{2}|\mathbf{x} - \boldsymbol{\theta}|^2 = \frac{1}{2}(x_1-\theta_1)^2 + \frac{1}{2}(x_2-\theta_2)^2,$$

so

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}}\left(|\mathbf{x}_i - \boldsymbol{\theta}|\right) = \frac{1}{m}\sum_{i=1}^{m} \begin{bmatrix} \theta_1 - x_{i1} \\ \theta_2 - x_{i2} \end{bmatrix} = \boldsymbol{\theta} - \bar{\mathbf{x}},$$

where $\bar{\mathbf{x}}$ is the empirical mean of the dataset. Thus, the update rule is

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha(1-\beta)^{t+1}(\boldsymbol{\theta} - \bar{\mathbf{x}}).$$

The recurrence relation is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha(1-\beta)^{t+1}(\boldsymbol{\theta}_t - \bar{\mathbf{x}})$$

for $t \geq 0$.

(b) Assuming $\beta = 0$, discuss convergence of the batch gradient descent algorithm.

From part (a), we compute

$$\boldsymbol{\theta}_t - \bar{\mathbf{x}} = (1-\alpha)^t(\boldsymbol{\theta}_0 - \bar{\mathbf{x}})$$

for all $t \geq 1$. Thus, the algorithm converges to the empirical mean $\mathbf{x}$ if the learning rate is $\alpha < 1$.