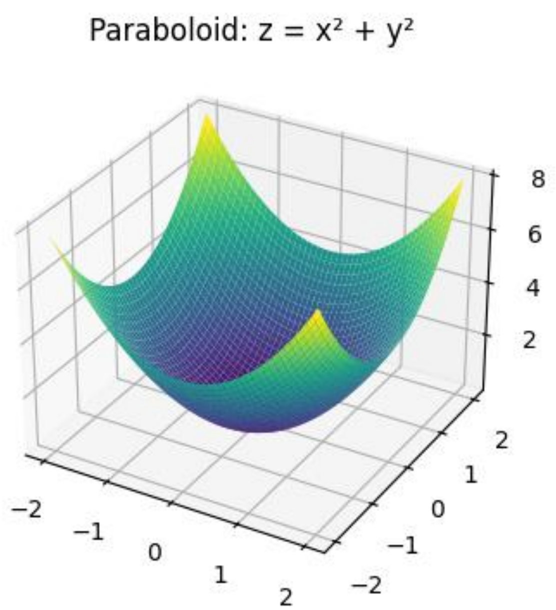
Probable Solutions for Exercise Sheet 4

Problem 1: Gradient

given the function $f(x, y) = x^2 + y^2$. This represents a paraboloid when visualized in 3D. The height z corresponds to $f(x, y)$.

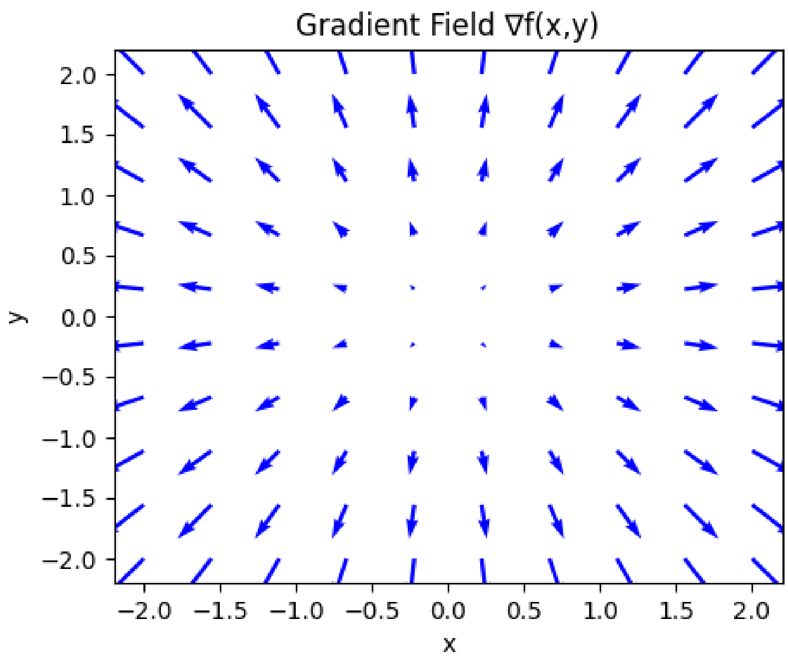
The graph of $z = x^2 + y^2$ is a bowl-shaped surface called a paraboloid. At the origin (0,0), the height is 0, and as move away from the origin, the height increases quadratically.



.

Gradient Calculation

$\nabla f(x, y) = (\partial f/\partial x, \partial f/\partial y) = (2x, 2y)$. The gradient points radially outward from the origin and its magnitude grows with distance



Height change along gradient

Moving along the gradient means going uphill on the paraboloid in the steepest direction. So the height z increases fastest. Moving opposite to the gradient decreases z fastest (gradient descent principle).

Problem 2: Chain Rule

(a) Gradient of f and Jacobian of g

Gradient of $f(y_1, y_2)$:

$\frac{\partial f}{\partial y_1} = 2y_1, \quad \frac{\partial f}{\partial y_2} = 2y_2 \; \nabla f(y_1, y_2) = (2y_1,\; 2y_2)$

Jacobian of $g(x_1, x_2, x_3)$:

$$y_1 = x_1^2 + x_2^3 + x_3, \quad y_2 = x_1 x_2 x_3$$

Compute the partial derivatives:

$$J_g = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 & 3x_2^2 & 1 \\ x_2 x_3 & x_1 x_3 & x_1 x_2 \end{bmatrix}$$

b. Gradient of the Composition $f \circ g$

By the **chain rule**:

$$\nabla_x (f \circ g)(x) = J_g^T \cdot \nabla f\big(g(x)\big)$$

Compute:

$$\nabla f\big(g(x)\big) = (2y_1, 2y_2) = \Big(2(x_1^2 + x_2^3 + x_3),\ 2(x_1 x_2 x_3)\Big)$$

Thus:

$$\nabla_x (f \circ g)(x) = \begin{bmatrix} 2x_1 & x_2 x_3 \\ 3x_2^2 & x_1 x_3 \\ 1 & x_1 x_2 \end{bmatrix} \cdot \begin{bmatrix} 2(x_1^2 + x_2^3 + x_3) \\ 2(x_1 x_2 x_3) \end{bmatrix}$$

Computing Each Component

$$\frac{\partial}{\partial x_1} = 2x_1 \cdot 2(x_1^2 + x_2^3 + x_3) + (x_2 x_3) \cdot 2(x_1 x_2 x_3)$$

$$\frac{\partial}{\partial x_2} = 3x_2^2 \cdot 2(x_1^2 + x_2^3 + x_3) + (x_1 x_3) \cdot 2(x_1 x_2 x_3)$$

$$\frac{\partial}{\partial x_3} = 1 \cdot 2(x_1^2 + x_2^3 + x_3) + (x_1 x_2) \cdot 2(x_1 x_2 x_3)$$

Problem 3: Geometric Interpretation of L2 Regularization

Given Loss Function

$$\tilde{L}(\theta_1, \theta_2) = 0.25(\theta_1 - 4)^2 + 5(\theta_2 - 3)^2$$

(a) the Minimum

Since the function is quadratic, the minimum occurs where the gradient is zero.

$$\frac{\partial \tilde{L}}{\partial \theta_1} = 0.5(\theta_1 - 4), \quad \frac{\partial \tilde{L}}{\partial \theta_2} = 10(\theta_2 - 3)$$

Setting both derivatives equal to zero:

$$\theta_1 = 4, \quad \theta_2 = 3$$

Thus, the optimal parameters are:

$$\hat{\theta} = (4,3)$$

(b) Hessian and Eigenvalues

the second derivatives:

$$H = \begin{bmatrix} \frac{\partial^2 \tilde{L}}{\partial \theta_1^2} & 0 \\ 0 & \frac{\partial^2 \tilde{L}}{\partial \theta_2^2} \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0 & 10 \end{bmatrix}$$

The eigenvalues are:

$$\sigma_1 = 0.5, \quad \sigma_2 = 10$$

and the eigenvectors are:

$$e_1 = (1,0), \quad e_2 = (0,1)$$

(c) Effect of L2 Regularization

When adding L2 regularization, each parameter is scaled (shrunk) by:

$$\text{Shrinkage Factor} = \frac{\sigma_i}{\sigma_i + \lambda}$$

For different values of $\lambda \in \{0.1,1,10\}$:

λ	**Factor for θ_1 (σ_1 = 0.5)**	**Factor for θ_2 (σ_2 = 10)**
0.1	$\frac{0.5}{0.5 + 0.1} = 0.833$	$\frac{10}{10 + 0.1} \approx 0.99$
1	$\frac{0.5}{0.5 + 1} = 0.333$	$\frac{10}{10 + 1} \approx 0.91$
10	$\frac{0.5}{0.5 + 10} \approx 0.047$	$\frac{10}{10 + 10} = 0.5$

Original optimal parameters:

$$\hat{\theta} = (4,3)$$

Eigenvalues:

$$\sigma_1 = 0.5 \quad (\text{for } \theta_1), \quad \sigma_2 = 10 \quad (\text{for } \theta_2)$$

After applying L2 regularization, each parameter is scaled by its corresponding shrinkage factor:

$$\theta_1^{\text{new}} = 4 \times \frac{0.5}{0.5+\lambda}, \quad \theta_2^{\text{new}} = 3 \times \frac{10}{10+\lambda}$$

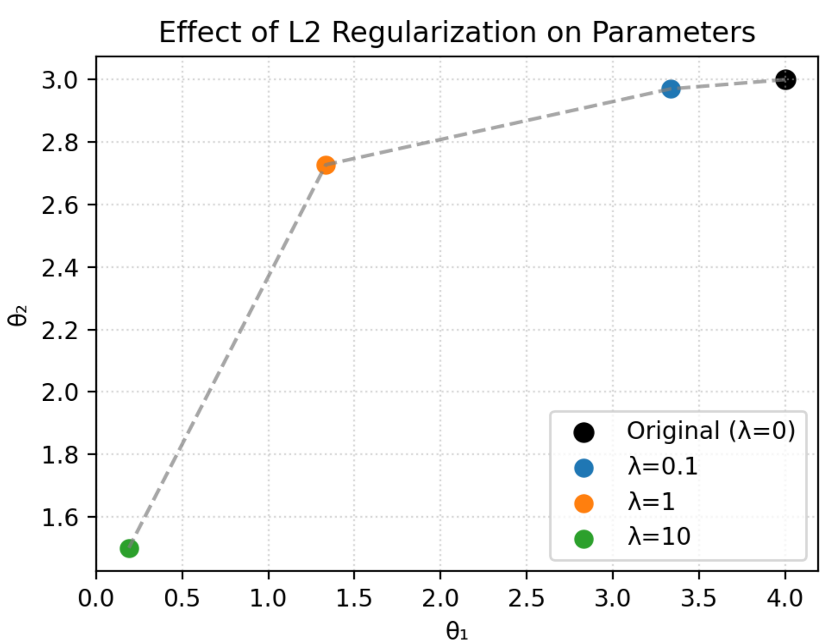
λ	**θ_1 factor**	**θ_2 factor**	**θ_1 new**	**θ_2 new**
0.1	0.8333	0.9901	3.33	2.97
1	0.3333	0.9091	1.33	2.73
10	0.0476	0.5	0.19	1.5

Increasing λ leads to stronger shrinkage.

Shrinkage is stronger in directions with smaller eigenvalues.
→ Here, θ_1 (with $\sigma_1 = 0.5$) shrinks much more than θ_2 ($\sigma_2 = 10$).

Visual Sketch

The plot below shows anisotropic shrinkage under L2 regularization: θ_1 shrinks faster (smaller curvature) than θ_2 (larger curvature).



.