

# Exercises - Solutions

## §3.5: Gradient Descent

**Problem 1**  $g(w) = (w - 5)^2$ ,  $g'(w) = 2(w - 5)$ . At  $w_0 = 1$ :  $g'(1) = -8$ .  $w_1 = 1 - 0.1(-8) = 1.8$ . At  $w_1 = 1.8$ :  $g'(1.8) = -6.4$ .  $w_2 = 1.8 - 0.1(-6.4) = 2.44$ . Costs:  $g(1) = 16$ ,  $g(1.8) = 10.24$ ,  $g(2.44) = 6.5536$ .

**Problem 2**  $g(w_1, w_2) = w_1^2 + 2w_2^2$ .  $\nabla g = \begin{bmatrix} 2w_1 \\ 4w_2 \end{bmatrix}$ . At  $(2, -1)$ :  $\nabla g = \begin{bmatrix} 4 \\ -4 \end{bmatrix}$ . Therefore

$$\mathbf{w}_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix} - 0.1 \begin{bmatrix} 4 \\ -4 \end{bmatrix} = \begin{bmatrix} 1.6 \\ -0.6 \end{bmatrix}.$$

**Problem 3**  $\Delta \mathbf{w} = -\alpha \nabla g = -0.05 \begin{bmatrix} 3 \\ -4 \end{bmatrix} = \begin{bmatrix} -0.15 \\ 0.20 \end{bmatrix}$ . Local change in cost:  $\Delta g \approx \nabla g^T \Delta \mathbf{w}$

$$= \begin{bmatrix} 3 & -4 \end{bmatrix} \begin{bmatrix} -0.15 \\ 0.20 \end{bmatrix} = -0.45 - 0.80 = -1.25. \text{ Equivalently: } -\alpha \|\nabla g\|^2 = -0.05(25) = -1.25.$$

**Problem 4**  $\nabla g(w_1, w_2) = \begin{bmatrix} 2(w_1 - 2) \\ 8(w_2 + 1) \end{bmatrix}$ . At  $(0, 0)$ :  $\nabla g = \begin{bmatrix} -4 \\ 8 \end{bmatrix}$ . First iterate:

$$\mathbf{w}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.05 \begin{bmatrix} -4 \\ 8 \end{bmatrix} = \begin{bmatrix} 0.2 \\ -0.4 \end{bmatrix}. \text{ At } (0.2, -0.4): \nabla g = \begin{bmatrix} -3.6 \\ 4.8 \end{bmatrix}. \text{ Second iterate:}$$

$$\mathbf{w}_2 = \begin{bmatrix} 0.2 \\ -0.4 \end{bmatrix} - 0.05 \begin{bmatrix} -3.6 \\ 4.8 \end{bmatrix} = \begin{bmatrix} 0.38 \\ -0.64 \end{bmatrix}.$$

### Problem 5

```
w = np.array([0.0, 0.0])
alpha = 0.05

for k in range(30):
    w = w - alpha*grad_g(w)

print(w)
print(g(w))
```

Approximately:  $\mathbf{w} \approx (1.9152, -1.0000)$ ,  $g(\mathbf{w}) \approx 0.00719$ . The first coordinate converges more slowly than the second for this cost function.

**Problem 6** The magnitude of the gradient measures how strongly the cost changes locally. Using  $-\nabla g$  preserves that information. Far from a stationary point, the gradient may be large. Near a stationary point, it naturally becomes small.