Exercise 8

1.

a)

$f'(\theta) = \theta^2 + 40\theta - 50$

Update rules (velocity form):

$v_{t+1} = \rho v_t - \alpha \nabla f(\theta_t)$

$\theta_{t+1} = \theta_t + v_{t+1}$

Iteration 1 ($t=0 \to 1$):

Gradient at $\theta_0 = 20$: $\nabla f(\theta_0) = f'(20) = 20^2 + 40 \cdot 20 - 50 = 400 + 800 - 50 = 1150$

Velocity: $v_1 = \rho v_0 - \alpha \nabla f(\theta_0) = 0{,}5 \cdot 0 - 0{,}01 \cdot 1150 = -11{,}5$

Parameter: $\theta_1 = \theta_0 + v_1 = 20 - 11{,}5 = \underline{8{,}5}$

Iteration 2 ($t=1 \to 2$):

Gradient at $\theta_1 = 8{,}5$: $\nabla f(\theta_1) = f'(8{,}5) = 8{,}5^2 + 40 \cdot 8{,}5 - 50 = 72{,}25 + 340 - 50 = 362{,}25$

Velocity: $v_2 = \rho v_1 - \alpha \nabla f(\theta_1) = 0{,}5(-11{,}5) - 0{,}01 \cdot 362{,}25 = -5{,}75 - 3{,}6225 = -9{,}3725$

Parameter: $\theta_2 = \theta_1 + v_2 = 8{,}5 - 9{,}3725 = \underline{-0{,}8725}$

b)

First vanilla gradient descent:

Update Rule:

$\theta_{t+1} = \theta_t - \alpha \nabla f(\theta_t)$

Iteration 1: $\theta_1^{(GD)} = 20 - 0{,}01 \cdot 1150 = 8{,}5$

Iteration 2: $\theta_2^{(GD)} = 8{,}5 - 0{,}01 \cdot 362{,}25 = 4{,}8775$

The true minimum satisfies $f'(\theta) = 0$: $\theta^2 + 40\theta - 50 = 0 \iff \theta \approx 1{,}21$

Distance to the minimizer:

- Vanilla GD after 2 steps: $|4{,}88 - 1{,}21| \approx 3{,}67$
- Momentum after 2 steps: $|-0{,}87 - 1{,}21| \approx 2{,}09$

Function values:

- $f(4{,}88) \approx 265{,}6$ (Vanilla GD)
- $f(-0{,}87) \approx 53{,}6$ (momentum)

Using momentum makes the effective step in the 2. Iteration much larger in the same descent duration (step size $-9{,}37$ instead of $-3{,}62$). After only 2 Iteration we are closer to the true minimum and have a much lower objective value with momentum than with GD. Momentum therefore speeds up convergence for this problem.

2.

Iteration 1 ($t=0$): $g_0 = (4,2,3,1)$

First moment: $s_1 = 0{,}9 \cdot 0 + 0{,}1\, g_0 = 0{,}1(4,2,3,1) = (0.4, 0.2, 0.3, 0.1)$

Second moment: $g_0^2 = (16,4,9,1)$ $\quad r_1 = 0{,}99 \cdot 0 + 0{,}01\, g_0^2 = 0{,}01(16,4,9,1) = (0.16, 0.04, 0.09, 0.01)$

Bias-corrected moments: $\hat{s}_1 = \frac{s_1}{1-0{,}9^1} = \frac{s_1}{0{,}1} = (4,2,3,1)$ $\quad \hat{r}_1 = \frac{r_1}{1-0{,}99^1} = \frac{r_1}{0{,}01} = (16,4,9,1)$

Update step: $\Delta\theta_0 = -0{,}01 \frac{(4,2,3,1)}{(\sqrt{16},\sqrt{4},\sqrt{9},\sqrt{1})} = -0{,}01 \frac{(4,2,3,1)}{(4,2,3,1)} = -0{,}01(1,1,1,1) = (-0.01,-0.01,-0.01,-0.01)^T$

Iteration 2 ($t=1$): $g_1 = (1,2,1,2)$

First moment: $s_2 = 0{,}9 s_1 + 0{,}1 g_1 = (0.36, 0.18, 0.27, 0.09) + (0.1, 0.2, 0.1, 0.2) = (0.46, 0.38, 0.37, 0.29)$

Second moment: $g_1^2 = (1,4,1,4)$ $\quad r_2 = (0.1584, 0.0396, 0.0891, 0.0099) + (0.01, 0.04, 0.01, 0.04) = (0.1684, 0.0796, 0.0991, 0.0499)$

Bias corrections: $\hat{s}_2 = \frac{s_2}{1-0{,}9^2} = \frac{s_2}{0{,}19}$ $\quad \hat{s}_2 \approx (2.421, 2.000, 1.947, 1.526)$

$\hat{r}_2 = \frac{r_2}{1-0{,}99^2} = \frac{r_2}{0{,}0199}$ $\quad \hat{r}_2 \approx (8.46, 4.00, 4.98, 2.51)$ $\quad \sqrt{\hat{r}_2} \approx (2.91, 2.00, 2.23, 1.58)$

Update step: $\Delta\theta_1 = -0{,}01 \frac{\hat{s}_2}{\sqrt{\hat{r}_2}}$ $\quad \Delta\theta_1 = \left(-0.01 \cdot \left(\frac{2.421}{2.91}\right), -0.01\left(\frac{2.000}{2.00}\right), -0.01 \cdot \left(\frac{1.947}{2.23}\right), -0.01\left(\frac{1.526}{1.58}\right)\right)$

$\approx (-0.00832, -0.01000, -0.00873, -0.00966)^T$

3.

Initializing all parameters with the same constant c is not a good idea because it causes symmetry among neurons, leading to identical gradients and preventing the network from learning meaningful features. Better strategies Xavier/Glorot initialization for tanh/sigmoid networks to keep forward and backward signal variance stable, Kaiming initialization for ReLU networks and orthogonal initialization. All maintain proper variance and allow neurons to learn different representations.