Exercise 7

Malik Mardan

1.

$g := \nabla_\theta R_{emp}$ and H_θ is the Hessian at θ_0

a)

$$q(\theta) := R_{emp}(\theta_0) + (\theta - \theta_0)^T g + \frac{1}{2}(\theta - \theta_0)^T H_\theta (\theta - \theta_0)$$

To find it's critical point we differentiate w.r.t. θ:

$$\nabla_\theta q(\theta) = g + H_\theta(\theta - \theta_0)$$

(Gradient of $\frac{1}{2}(\theta - \theta_0)^T H_\theta (\theta - \theta_0)$ is $H_\theta(\theta - \theta_0)$ when H_θ is symmetric)

Set gradient to zero for minimum:

$$g + H_\theta(\theta - \theta_0) = 0 \iff H_\theta(\theta - \theta_0) = -g \iff \theta = \theta_0 - H_\theta^{-1} g$$

If we denote the current iterate by θ_t:

$$\theta_{t+1} = \theta_t - H_\theta(\theta_t)^{-1} \nabla_\theta R_{emp}(\theta_t)$$

b)

$$f(x) = (x-1)(x-3) = x^2 - 4x + 3 \quad f'(x) = 2x - 4 \quad f''(x) = 2$$

Newton's Method in 1D is: $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$

$$f'(4) = 2 \cdot 4 - 4 = 4 \qquad f''(4) = 2$$

$$x_1 = 4 - \frac{4}{2} = 2$$

The exact minimizer of the quadratic is at:

$$x^* = -\frac{b}{2a} = -\frac{-4}{2 \cdot 1} = 2$$

After one step we are already at the minimum $x_1 = 2 = x^*$

c)

- The Hessian is a $d \times d$ matrix for d parameters. Memory and time cost scale as $O(d^2)$ just to store it, which is infeasible for large neural networks.

- Each Newton step requires solving $H_\theta \Delta\theta = -g$, which in general is $O(d^3)$ if done naively. This is far too expensive for large models.

- In practice we use mini-batches, so gradients and Hessian estimates are noisy. Accurate second-order information is hard and expensive to obtain per step.

2.

a)

x_1:

- x_1 lies near the steep, central region of the sigmoid
- the tangent line is steep
- the gradient $\sigma'(x_1)$ is relatively large.

Therefore updates to parameters at this point are meaningful.

x_2:

- x_2 lies on the right flat tail of the sigmoid
- the tangent line is almost horizontal
- the gradient $\sigma'(x_2)$ is close to zero

Therefore parameter updates become extremely small.

b)

i) Hidden pre activation: $a = \theta_1 x_1 = 4 \cdot 1{,}5 = 6$

Hidden activation: $z = \sigma(6) = \frac{1}{1+e^{-6}} \approx 0{,}9975$

Output pre activation: $b = \theta_2 z = 1{,}5 \cdot 0{,}9975 = 1{,}4963$

Output activation: $f = \sigma(1{,}4963) = \frac{1}{1+e^{-1{,}4963}} \approx 0{,}8170$

Loss: $L = \frac{1}{2}(f-y)^2 = \frac{1}{2}(0{,}8170-1)^2 = \frac{1}{2}(-0{,}1830)^2 \approx 0{,}0167$

$\sigma'(b) = f(1-f) = 0{,}8170(1-0{,}8170) = 0{,}1493$

$\sigma'(a) = z(1-z) = 0{,}9975(1-0{,}9975) = 0{,}0025$

$\frac{\partial L}{\partial \theta_2} = (f-y)\sigma'(b)z = (-0{,}1830)(0{,}1493)(0{,}9975) \approx -0{,}0273$

$\frac{\partial L}{\partial \theta_1} = (f-y)\sigma'(b)\,\theta_2\sigma'(a)x = (-0{,}1830)(0{,}1493)(1{,}5)(0{,}0025)(1{,}5)$

$= -0{,}000153$

ii)

Hidden: $a=6$, $z = ReLU(6) = 6$

Output: $b = 1{,}5 \cdot 6 = 9$, $f = ReLU(9) = 9$

Loss: $\frac{1}{2}(9-1)^2 = \frac{1}{2} \cdot 8^2 = 32$

$ReLU'(b) = ReLU'(9) = 1$, $ReLU'(a) = ReLU'(6) = 1$

$\frac{\partial L}{\partial \theta_2} = (f-y) \cdot 1 \cdot z = (9-1) \cdot 6 = 48$

$\frac{\partial L}{\partial \theta_1} = (f-y) \cdot 1 \cdot \theta_2 \cdot 1 \cdot x = 8 \cdot 1{,}5 \cdot 1{,}5 = 18$

Interpretation:

Sigmoid:

- $z=6$ means deep saturation
- $\sigma'(6) = 0{,}0025$ is almost zero and means a huge gradient shrinkage
- Gradient θ_1 is $-0{,}00015$ which means almost no learning happens

ReLU:

- ReLU' stays 1 for all positive values
- for $z=6$ we have no saturation
- Gradient on θ_1 is large (18). Therefore learning continues normally.

The reason why ReLU fixes vanishing gradients is, that it is linear with slope 1 on the positive side and back-propagated gradients are not multiplied by numbers <1. We also have no shrinking across layers.

Tanh still suffers the same problem as sigmoid, because tanh saturates for large $|x|$. $\tanh'(x) = 1 - \tanh^2(x)$ becomes very small when x is far from 0. Even though it is centered at zero, it still has flat tails and therefore vanishing gradients return.