This quiz covers units 2 and 3. Each of the four questions is worth 25%. Please make sure to follow the hand-in instructions described in Canvas announcements and in the course website.

---

The following formulas are taken from Unit 3 and may be of use for questions 2, 3, and 4.

- **Sufficient decrease condition**:

$$f(\theta^{(k+1)}) \leq f(\theta^{(k)}) + \beta\alpha\nabla_{\mathbf{d}^{(k)}} f(\theta^{(k)}).$$

- **Updates for gradient descent with momentum** (with $g^{(k)}$ denoting the gradient):

$$\begin{aligned}\mathbf{v}^{(k+1)} &= \beta\mathbf{v}^{(k)} - \alpha g^{(k)} \\ \theta^{(k+1)} &= \theta^{(k)} + \mathbf{v}^{(k+1)}.\end{aligned}$$

- **ADAM** (with $g^{(k)}$ denoting the gradient):

$$\begin{aligned}\mathbf{v}^{(k+1)} &= \gamma_v \mathbf{v}^{(k)} + (1-\gamma_v) g^{(k)}, \quad \text{(biased momentum)} \\ \mathbf{s}^{(k+1)} &= \gamma_s \mathbf{s}^{(k)} + (1-\gamma_s)\left(g^{(k)} \odot g^{(k)}\right), \quad \text{(biased squared gradient)} \\ \widehat{\mathbf{v}}^{(k+1)} &= \frac{\mathbf{v}^{(k+1)}}{1-\gamma_v^k}, \quad \text{(unbiased momentum)} \\ \widehat{\mathbf{s}}^{(k+1)} &= \frac{\mathbf{s}^{(k+1)}}{1-\gamma_s^k}, \quad \text{(unbiased squared gradient)} \\ \theta^{(k+1)} &= \theta^{(k)} - \alpha \frac{\hat{\mathbf{v}}^{(k+1)}}{\epsilon + \sqrt{\hat{\mathbf{s}}^{(k+1)}}}, \quad \text{(next iterate)},\end{aligned}$$

  where $\odot$ denotes the element-wise multiplication of two vectors.

---

**Question 1:** Consider the following model for the binary classification task with outcome $Y \in \{0, 1\}$. There are 2 features $X_1$ and $X_2$. The model is,

$$\mathbb{P}(Y = 1 \mid X_1, X_2) = \sigma(b + w_1 X_1 + w_2 X_2),$$

where $\sigma(\cdot)$ is the sigmoid function, and $b$, $w_1$, and $w_2$ are scalars. A binary classifier is constructed by deciding $Y = 1$ if $\mathbb{P}(Y = 1 \mid X_1, X_2) > 0.5$ and otherwise deciding $Y = 0$.

(a) Show that this model leads to a linear (affine) boundary decision within the $X_1$, $X_2$ plane.

(b) Consider a variation of the previous model by using the square of each feature such that,

$$\mathbb{P}(Y = 1 \mid X_1, X_2) = \sigma(b + w_1 X_1^2 + w_2 X_2^2),$$

and the same binary classification is used as before. What is now the shape of the boundary decision (within the $X_1$, $X_2$ plane).

**Solution for Q1:**
(a) The condition $\mathbb{P}(Y = 1 \mid X_1, X_2) > 0.5$ is equivalent to,

$$\frac{1}{1 + e^{-(b+w_1X_1+w_2X_2)}} > 0.5,$$

or after isolating the exponent,

$$e^{-(b+w_1X_1+w_2X_2)} < 1$$

or,

$$w_1X_1 + w_2X_2 > -b.$$

This defines an affine decision boundary. That is, the decision if to classify as $Y = 0$ or $Y = 1$ is based on a separating hyperplane in the $(X_1, X_2)$ plane.

(b) Isolating the argument of the exponent as above we get,

$$w_1X_1^2 + w_2X_2^2 > -b.$$

This is no longer a linear decision boundary but rather a curved boundary which may be an ellipse or a hyperbolic boundary depending on $w_1$ and $w_2$.

**Question 2 (Backtracking line search):** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a Rosenbrock banana function defined by

$$f(\theta_1, \theta_2) = (1 - \theta_1)^2 + 50(\theta_2 - \theta_1^2)^2.$$

Suppose we are at $k^{th}$ iteration of a descent direction method that uses backtracking for approximate line search to find $\alpha^{(k)}$. Assume that $\theta^{(k)} = [-1, 1/2]$ and the descent direction is $d^{(k)} = -\frac{\nabla f(\theta^{(k)})}{\|\nabla f(\theta^{(k)})\|}$.

For the backtracking, take $\beta = 0.001$ and the initial value of $\alpha$ to be 3, and decrease $\alpha$ according to the rule $\alpha \leftarrow \alpha/3$ until the sufficient decrease condition is obtained.

Determine (the final) $\alpha^{(k)}$, $\theta^{(k+1)}$ and $f\left(\theta^{(k+1)}\right)$.

**Solution for Q2:** First observe that

$$\nabla f(\theta_1, \theta_2) = \left[-2(1 - \theta_1) - 200\theta_1(\theta_2 - \theta_1^2), \quad 100(\theta_2 - \theta_1^2)\right],$$

and the directional derivative

$$\nabla_{d^{(k)}} f(\theta^{(k)}) = d^{(k)} \cdot \nabla f(\theta_1, \theta_2) = -\|\nabla f(\theta^{(k)})\|.$$

Therefore, for $\theta^{(k)} = [-1, 1/2]$,

$$\begin{aligned}
f(\theta^{(k)}) &= 16.5, \\
\nabla f(\theta^{(k)}) &= [-104, \ -50], \\
\|\nabla f(\theta^{(k)})\| &= \sqrt{104^2 + 50^2}, \\
\nabla_{d^{(k)}} f(\theta^{(k)}) &= -\sqrt{104^2 + 50^2}, \\
d^{(k)} &= \left[\frac{104}{\sqrt{104^2 + 50^2}}, \ \frac{50}{\sqrt{104^2 + 50^2}}\right] \approx [0.9013, \ 0.4333].
\end{aligned}$$

Furthermore, since $\beta = 0.001$ and $\sqrt{104^2 + 50^2} \approx 115.39$, for all $0 \leq \alpha \leq 3$,

$$0 \geq \beta\alpha\nabla_{d^{(k)}} f(\theta^{(k)}) \geq -0.003 \times \sqrt{104^2 + 50^2} \approx -0.35.$$

Hence,

$$16.5 \geq f(\theta^{(k)}) + \beta\alpha\nabla_{d^{(k)}} f(\theta^{(k)}) \geq 16.15.$$

**Iterate 1:** For $\alpha = 3$,

$$\theta^{(k+1)} = \theta^{(k)} + 3d^{(k)} \approx [-1 + 3 \times 0.9013, \quad 0.5 + 3 \times 0.4333] = [1.7039,\ 1.7999].$$

For this $\theta^{(k+1)}$, we have $f(\theta^{(k+1)}) \geq 50(\theta_2 - \theta_1^2)^2 \approx 60.87$. Clearly, the sufficient decrease condition is not satisfied.

**Iterate 2:** For $\alpha = 1$,

$$\theta^{(k+1)} = \theta^{(k)} + d^{(k)} \approx [-1 + 0.9013, \quad 0.5 + 0.4333] = [-0.0987,\ 0.9333],$$

and $f(\theta^{(k+1)}) \geq 50(\theta_2 - \theta_1^2)^2 \approx 42.65$. Again the sufficient decrease condition is not satisfied.

**Iterate 3:** For $\alpha = 1/3$,

$$\theta^{(k+1)} = \theta^{(k)} + d^{(k)}/3 \approx [-1 + 0.9013/3, \quad 0.5 + 0.4333/3] = [-0.6996,\ 0.6444],$$

and $f(\theta^{(k+1)}) \approx 4.09$. Now, the sufficient decrease condition is satisfied.

As consequence, $\alpha^{(k)} = 1/3$, $\theta^{(k+1)} \approx [-0.6996,\ 0.6444]$, and $f(\theta^{(k+1)}) \approx 4.09$.

**Question 3 (Gradient descent with momentum):** Let

$$f(\theta) = \theta_1^2/2 + \theta_2^2/2$$

for $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$. Using gradient descent with momentum, take $\theta^{(1)} = [-5, 5]$, initial momentum update $\mathbf{v}^{(1)} = [0, 0]$, and $\beta = 1/2$. Assume the learning rate $\alpha$ is fixed at $1/5$.

Determine $\theta^{(2)}$, $\theta^{(3)}$, $\mathbf{v}^{(2)}$, and $\mathbf{v}^{(3)}$.

**Note:** Do not normalize the gradient vectors for updating the vectors $\mathbf{v}$.

**Solution for Q3:** Observe that

$$\nabla f(\theta) = [\theta_1, \theta_2].$$

**Iterate 1:** Since $\mathbf{v}^{(1)} = [0, 0]$ and $\theta^{(1)} = [-5, 5]$,

$$\mathbf{v}^{(2)} = -\left[-5, 5\right]/5 = [1, -1],$$

and hence,

$$\theta^{(2)} = [-5, 5] + [1, -1] = [-4, 4].$$

**Iterate 2:** Since $\nabla f(\theta^{(2)}) = [-4, 4]$,

$$\mathbf{v}^{(3)} = [1, -1]/2 - [-4, 4]/5 = [13/10,\ -13/10],$$

and hence,

$$\theta^{(3)} = [-4, 4] + [13/10, -13/10] = [-27/10,\ 27/10].$$

**Question 4 (ADAM optimizer):** Again consider the function

$$f(\theta) = \theta_1^2/2 + \theta_2^2/2, \quad \theta \in \mathbb{R}^2.$$

For the ADAM (Adaptive Moment Estimation) algorithm, take $\gamma_v = \gamma_s = 0.9$, $\alpha = 1$, and $\epsilon = 0$. Assume $\theta^{(1)} = [-5, 5]$ and initialise $\mathbf{v}^{(1)}$ and $\mathbf{s}^{(1)}$ both at $[0, 0]$.

Determine $\widehat{\mathbf{v}}^{(2)}$ and $\widehat{\mathbf{v}}^{(3)}$, as well as $\widehat{\mathbf{s}}^{(2)}$ and $\widehat{\mathbf{s}}^{(3)}$, and the points $\theta^{(2)}$ and $\theta^{(3)}$.

**Solution for Q4:** Again, observe that

$$\nabla f(\theta) = [\theta_1, \theta_2].$$

**Iterate 1:** Since $\mathbf{v}^{(1)} = [0, 0]$, $\mathbf{s}^{(1)} = [0, 0]$ and $\theta^{(1)} = g^{(1)} = \nabla f(\theta^{(1)}) = [-5, 5]$,

$$\mathbf{v}^{(2)} = (1 - 0.9)\,[-5, 5] = [-0.5, 0.5],$$
$$\mathbf{s}^{(2)} = (1 - 0.9)\,[25, 25] = [2.5, 2.5].$$

Hence,

$$\widehat{\mathbf{v}}^{(2)} = [-5, 5]\,,$$
$$\widehat{\mathbf{s}}^{(2)} = [25, 25]\,.$$

Consequently,

$$\theta^{(2)} = [-5, 5] - \left[-5/\sqrt{25},\ 5/\sqrt{25}\right] = [-4, 4].$$

**Iterate 2:** Using the vectors from the first iteration and $g^{(2)} = \nabla f(\theta^{(2)}) = [-4, 4]$,

$$\mathbf{v}^{(3)} = 0.9[-0.5, 0.5] + (1 - 0.9)\,[-4, 4] = [-17/20,\ 17/20],$$
$$\mathbf{s}^{(3)} = 0.9[2.5, 2.5] + (1 - 0.9)\,[16, 16] = [77/20,\ 77/20].$$

Hence,

$$\widehat{\mathbf{v}}^{(3)} = \frac{1}{1-0.9^2}\,[-17/20,\, 17/20] \approx [-4.47,\, 4.47],$$
$$\widehat{\mathbf{s}}^{(3)} = \frac{1}{1-0.9^2}\,[-77/20, 77/20] \approx [-20.26,\, 20.26].$$

Consequently,

$$\theta^{(3)} = [-4, 4] - \frac{\widehat{\mathbf{v}}^{(3)}}{\sqrt{\widehat{\mathbf{s}}^{(3)}}} \approx [-3.006,\, 3.006].$$