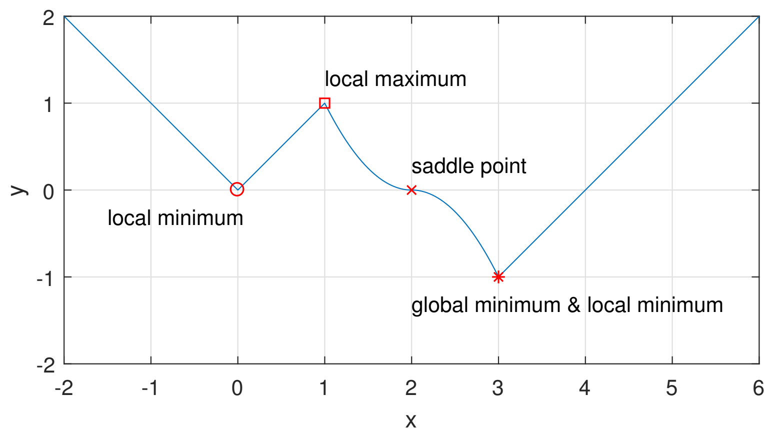
Mid-term exam solutions

1. (Local minimum, global minimum and saddle point)



Note that the global minimum is also a local minimum since the function value only increases in a small neighborhood of $x = 3$. $x = 2$ is a saddle point since $f'(2) = 0$ but it is neither local minimum or local maximum. Note that $f(x)$ is differentiable once at $x = 2$, but not twice.

2. (Gradient and Hessian)

Solution)

$$\nabla_{(x_1,x_2)} f(x_1,x_2) = \begin{pmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} \\ \frac{\partial f(x_1,x_2)}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2x_1 - 5x_2 + 3x_1^2 x_2^2 \\ 6x_2 - 5x_1 + 2x_1^3 x_2 \end{pmatrix}$$

$$H = \begin{pmatrix} \frac{\partial^2 f(x_1,x_2)}{\partial x_1^2} & \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_2^2} \end{pmatrix} = \begin{pmatrix} 2 + 6x_1 x_2^2 & -5 + 6x_1^2 x_2 \\ -5 + 6x_1^2 x_2 & 6 + 2x_1^3 \end{pmatrix}$$

3. (Unbiased estimator)

Solution)

$$\hat{\theta}(x) = |x|$$

will do the job since $\mathbb{E}[\hat{\theta}(x)] = \mathbb{E}[|x|] = \theta$ and the bias is $\mathbb{E}[\hat{\theta}(x)] - \theta = \theta - \theta = 0$.

4. (ReLU)

Solution)

a	b	c	d	e
1	$-\frac{1}{2}$	$-\frac{7}{2}$	4	-1

5. (XOR problem)

Solution) There are many solutions for this problem. One possibility is given as follows:

a	b	c	$\phi(x)$
1	1	-1	$1-\lvert x\rvert$

6. (Backpropagation)

Solution)

(a)

$$\begin{aligned}
\frac{\partial J}{\partial z_i} &= -\frac{\partial}{\partial z_i}\sum_{i'}\log\sigma((2y_{i'}-1)z_{i'}) \\
&= -\frac{\partial}{\partial z_i}\log\sigma((2y_i-1)z_i) \\
&= (1-2y_i)(1-\sigma((2y_i-1)z_i)) \\
&= \begin{cases} \hat{y}_i & \text{if } y_i = 0 \\ \hat{y}_i - 1 & \text{if } y_i = 1 \end{cases} \\
&= \hat{y}_i - y_i,
\end{aligned}$$

where the third equality follows since $\frac{\partial}{\partial z}\log\sigma(\alpha z) = \frac{\partial}{\partial z}\log\frac{1}{1+\exp(-\alpha z)} = \frac{\alpha\exp(-\alpha z)}{1+\exp(-\alpha z)} = \alpha(1-\sigma(\alpha z))$ and the fourth equality follows by considering two cases $y_i = 0$ or 1 and by using $1-\sigma(-z) = \sigma(z)$.

(b) Using the result in (a), we get

$$\nabla_{\mathbf{z}} J = \hat{\mathbf{y}} - \mathbf{y}.$$

(c) $\frac{\partial J}{\partial u_i}$ can be derived as follows:

$$\frac{\partial J}{\partial u_i} = \sum_{i'}\frac{\partial z_{i'}}{\partial u_i}\frac{\partial J}{\partial z_{i'}} = \phi'(u_i)w_2\frac{\partial J}{\partial z_i} = \phi'(u_i)w_2 g_i.$$

(d) Using the result in (c), we get

$$\nabla_{\mathbf{u}} J = w_2\phi'(\mathbf{u}) \odot \nabla_{\mathbf{z}} J = w_2\phi'(\mathbf{u}) \odot \mathbf{g}.$$

(e)

$$
\begin{aligned}
\frac{\partial J}{\partial w_2} &= \sum_i \frac{\partial z_i}{\partial w_2}\frac{\partial J}{\partial z_i} = \sum_i h_i \frac{\partial J}{\partial z_i} = \mathbf{h}^T \nabla_{\mathbf{z}} J = \mathbf{h}^T \mathbf{g} \\
\frac{\partial J}{\partial b_2} &= \sum_i \frac{\partial z_i}{\partial b_2}\frac{\partial J}{\partial z_i} = \sum_i \frac{\partial J}{\partial z_i} = \mathbf{1}^T \nabla_{\mathbf{z}} J = \mathbf{1}^T \mathbf{g} \\
\frac{\partial J}{\partial w_1} &= \sum_i \frac{\partial u_i}{\partial w_1}\frac{\partial J}{\partial u_i} = \sum_i x_i \frac{\partial J}{\partial u_i} = \mathbf{x}^T \nabla_{\mathbf{u}} J = \mathbf{x}^T \mathbf{f} \\
\frac{\partial J}{\partial b_1} &= \sum_i \frac{\partial u_i}{\partial b_1}\frac{\partial J}{\partial u_i} = \sum_i \frac{\partial J}{\partial u_i} = \mathbf{1}^T \nabla_{\mathbf{u}} J = \mathbf{1}^T \mathbf{f}
\end{aligned}
$$

7. (L^1 regularization)

Solution) The Hessian H of $J(x_1, x_2)$ is given by

$$H = \begin{pmatrix} 6 & 0 \\ 0 & 14 \end{pmatrix}.$$

Therefore,

$$
\begin{aligned}
\alpha_1 &= H_{1,1}|w_1^*| = 6 \times 5 = 30 \\
\alpha_2 &= H_{2,2}|w_2^*| = 14 \times 4 = 56
\end{aligned}
$$

where $H_{i,j}$ is the (i,j)-th element of H and $(w_1^*, w_2^*) = (5, 4)$ is the local minimum of J. See page 8 of lecture notes #7 to see why α_1 and α_2 shown above give the minimum α for which x_1^* and x_2^* are zero, respectively.

8. (Overfitting or underfitting)

(a) *An estimator $\hat{\theta}(x) = \frac{0.95+0.1x}{2}$ that estimates $0 \le \theta \le 1$ from x following Bernoulli(θ)*

Solution) This is an example of underfitting since $\hat{\theta}(x)$ is either 0.475 (when $x = 0$) or 0.525 (when $x = 1$), which is close to a trivial estimator that always outputs 0.5. Such a trivial estimator and the estimator in this problem would be in the underfitting regime whereas $\hat{\theta}(x) = x$ would be in the overfitting regime. $\hat{\theta}(x) = \frac{1+2x}{4}$ would achieve a good balance between underfitting and overfitting.

(b) *Someone thinks Earth is sitting on four giant elephants.*

Solution)

Overfitting. Why four? Why elephants? There are infinitely many other possibilities and you are simply believing what you chose to believe based on your limited experience.