50.021 Artificial Intelligence
Theory Homework 2

Due: every Monday, 4PM before class starts

[Q1]. Gradient descent is used to find a local minimum of a function, by taking steps proportional to the negative gradient. In class we learned that we do gradient descent to solve the decision boundary for logistic regression, by solving,

$$w* = \text{argmin}_{\boldsymbol{w}}\ L(\boldsymbol{w}) = \text{argmin}_{\boldsymbol{w}}(-1)\sum_{i=1}^{n}\log\left(h(\boldsymbol{x}_i)^{y_i}(1-h(\boldsymbol{x}_i))^{1-y_i}\right), \tag{1}$$

where $h(\boldsymbol{x})$ is,

$$h(\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

The update equation for the gradient descent is respectively,

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \eta\nabla L(\boldsymbol{w}^k), \tag{2}$$

where η is the step size.
Theres a counterpart method for gradient descent, called gradient ascent. It is used to find a local maximum of a function, by taking steps proportional to the gradient. How can we modify equation 1 and the update equation 2 if we want to do **gradient ascent** instead?

Solution:

Equation 1 becomes:

$$w* = \text{argmax}_w L'(\boldsymbol{w}) = \text{argmax}_w \log\sum_{i=1}^{n}\left(h(\boldsymbol{x}_i)^{y_i}(1-h(\boldsymbol{x}_i))^{1-y_i}\right),$$

The update equation for the gradient ascent is respectively,

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \eta\nabla L'(\boldsymbol{w}^k),$$

[Q2]. In class we learn to update weights using various learning rates, one of them is the momentum term,

$$w^{t+1} = w^t - m^{t+1}$$
$$m^{t+1} = \alpha m_t + \eta\nabla_w\hat{E}(w^t, L)$$
$$m_0 = 0, \alpha \in [0,1]$$

and another one is the exponential moving average,

$$w^{t+1} = w^t - \text{EMA}(\nabla_w\hat{E}(w^t, L))$$

Show that the momentum term and the EMA updates differ by only a constant c, i.e:

$$m^{t+1} = c\cdot\text{EMA}(\nabla_w\hat{E}(w^t, L))$$

and write down the expression for c.

Solution:
The general rule for the momentum term is,

$$m_{t+1} = \eta\left(\sum_{s=0}^{t}\alpha^t\nabla_w\hat{E}(w^s, L)\right)$$

The general rule for the EMA is,

$$EMA(\nabla_w \hat{E}(w^t, L)) = \sum_{s=0}^{t} \alpha^t (1-\alpha) \nabla_w \hat{E}(w^s, L)$$

Hence,

$$c = \frac{\eta}{1-\alpha}$$

Note: In the lecture notes $\nabla_w \hat{E}(w^s, L)$ *is simplified as* g_s