Quiz 7

1. Once the analytic gradient is computed with backpropagation, the gradients are used to perform a parameter update. Please describe the algorithm and principle of the following techniques:

   (a) SGD (Vanilla update).

   (b) Momentum update

   (c) Nesterov Momentum

   (d) Adagrad

   (e) RMSprop.

(a). $x_{t+1} = x_t - \alpha \nabla f(x_t)$

最簡單的方法，找出參數的梯度並沿梯度的反方向更新

(b) $v_{t+1} = \rho v_t - \alpha \nabla f(x_t)$ $\quad \rho = 0.9$ or 0.99

$x_{t+1} = x_t + v_{t+1}$

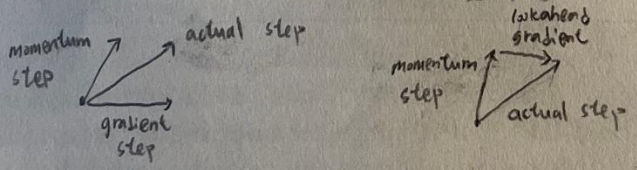
利用物理"動量"的概念，會隨著梯度方向產生速度，如果梯度保持相同則更新梯度的速度會越來越快。

(c)

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

$x_{t+1} = x_t + v_{t+1}$

在 momentum 的動量概念中加入校正因子，校正因子為"lookahead gradient"，是以現在位置加上 momentum 後去算梯度所得到的值。



(d)

$$n = \sum_{i=1}^{t} \left(\nabla f(x_t)\right)^2$$

$$x_{t+1} = x_t - \eta \frac{1}{\sqrt{n+\epsilon}} \nabla f(x)$$

ϵ 是為了避免分母為 0 所加入之極小數值的數

Adagrad 會依照梯度去調整 learning rate，前期梯度小放大 learning rate，後期梯度大約束學習率

(e)

$$\sigma_t = \sqrt{\beta (\sigma_{t-1})^2 + (1-\beta)(\nabla f(x))^2} \qquad 1$$

$$x_{t+1} = x_t - \frac{\eta}{\sigma_t} \nabla f(x)$$

如同 Adagrad 會去調整 learning rate，β 通常為 0.9、0.99、0.999 等數值，可避免提前結束訓練

2. We can begin to train the network using the following initialization approaches:

   (a) **all constant initialization**,
   (b) **small random numbers**, 一常態分佈
   (c) **calibrating the variances with 1/sqrt(n) or sqrt(2/n)**.

n = input 的數量

Please explain the pro and cons of each initialization approach.

(a) pros: 設置簡單
cons: hidden layers 內的節點在做的計算都一樣，成效不好

(b) pros: 各節點都有參與到產生出不同的輸出
cons: initialization 的數值太小可能導致最後幾層的 gradient 接近 0 (梯度消失)

(c) pros: 能使輸出與輸入的變異數保持一致 使輸出不會太集中或太過飽合
cons: 不能通用在所有情況，如使用 ReLU 為 activation function 時會導致輸出的 mean 為正的