梯度下降法

姓名：李紘宇

1 極值的特徵

一個點是多變數函數的極值，其必要條件是該點梯度爲零。當考慮單變數函數時，若該點微分爲零，且點的左邊斜率爲負、右邊斜率爲正，則該點爲局部極小值。

2 梯度下降法

梯度下降法是一種尋找函數局部極小值的方法，其思路是透過從一個起始點 $(x_0, f(x_0))$ 出發，以逐步逼近的方式，來尋找局部極小值 (如 Figure 1)。根據泰勒展開式的一階近似，我們可以得到

$$f(x + \Delta x) - f(x) \approx f'(x)\Delta x \tag{1}$$

由此我們可以計算，當在橫軸上移動 Δx 時，函數値的變化。如果我們設 $\Delta x = -\alpha f'(x)$，其中 α 爲學習率，是一個微小的值，將 Δx 代入式 (1) 可以得到函數値會根據以下公式變化

$$f(x + \Delta x) - f(x) = -\alpha f'(x)^2 \tag{2}$$

因此我們會發現，當我們不斷以下面的方法更新 x 值。如果斜率爲正時，更新後的點會往左下移動；斜率爲負時，更新後的點會往右下移動。如此便可逐步逼近最小值，這便是梯度下降法。

$$x = x - \alpha f'(x) \tag{3}$$

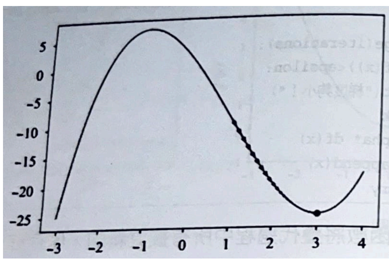

Figure 1: 梯度下降法逼近局部最小值的示意圖

多變量函數也可以使用梯度下降法，只需要稍微修改一下，其更新點的方式如下：

$$\mathbf{x} = \mathbf{x} - \alpha \nabla f(\mathbf{x}) \tag{4}$$

梯度下降法有許多變體，以下就來一一介紹它們。

3 Momentum 法

Momentum 法，更新點的方式如下：

$$\mathbf{x} = \mathbf{x} - \mathbf{v_t} \tag{5}$$

$$\mathbf{v_t} = \gamma \mathbf{v_{t-1}} + \alpha \nabla \mathbf{f}(\mathbf{x}) \tag{6}$$

其中 $\mathbf{v_t}$ 是更新向量。Momentum 法認爲 $\mathbf{v_t}$ 值的更新應該是有慣性的，因此在第 t 次的更新向量 $\mathbf{v_t}$ 中添加了前一次的更新向量 $\mathbf{v_{t-1}}$ 的貢獻。

優點：保留了之前的運動慣性，在平坦處保有運動速度，也不會因梯度突然變大而過衝。

4 AdaGrad 法

AdaGrad 法，更新點的方式如下：

$$x_{t+1,i} = x_{t,i} - \alpha \frac{1}{\sqrt{\sum_{t'=1}^{t} g_{t',i}^2} + \epsilon} g_{t,i} \tag{7}$$

其中 $x_{t,i}$ 是第 t 次迭代時 $\mathbf{x}$ 的 i 分量，$g_{t,i}$ 是第 t 次迭代時 f 在 i 方向上的偏導數 $\frac{\partial f}{\partial x_i}$，$\epsilon$ 是一個微小的常數，用以避免除數爲零情況。

函數在每個方向的偏導數可能差距過大，因此更新向量的每個分量使用相同的學習率不一定適合尋找極小值。AdaGrad 透過將每個梯度分量都除以該梯度分量的歷史累加值，解決這個問題

優點：消除各梯度分量差異的影響。
缺點：隨著累加值不斷增大，學習會變得緩慢。且更新方向可能會偏離最佳解。

5 AdaDelta 法

AdaDelta 法，更新點的方法如下：

$$x = x + \alpha \Delta x_t \tag{8}$$

$$\Delta x_t = -\sqrt{\frac{E[\Delta x^2]_{t-1} + \epsilon}{E[g^2]_t + \epsilon}} g_t \tag{9}$$

其中

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2 \tag{10}$$

$$E[\Delta x^2]_t = \gamma E[\Delta x^2]_{t-1} + (1-\gamma) \Delta x_t^2 \tag{11}$$

γ 爲衰減率參數，通常設爲 0.9。AdaDelta 法對更新向量改用移動平均法，解決了 AdaGrad 累加值不斷增大，使收斂速度越來越慢的問題。
優點：不會有收斂速度越來越慢的問題。且更新點的路徑更爲平滑。