香港中文大學
數學系
梯度下降法練習

單變量函數的梯度下降法(Gradient Descent Method, GDM)

我們現在介紹梯度下降法，以數值方式求得最小平方問題的近似解。這裏有一個尋找可微分的單變量函數$f(x)$ 最小值的問題，如下所示。

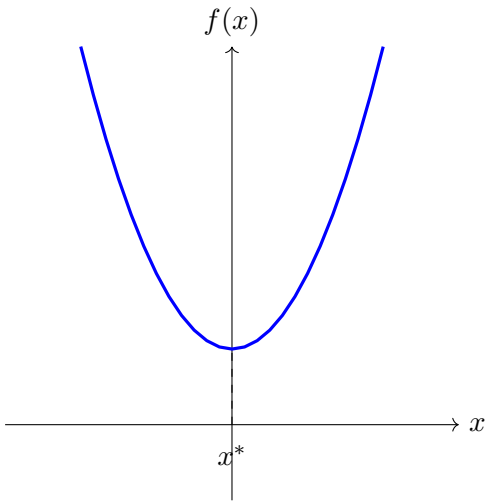

Figure 1: 函數$f(x) = x^2 + 1$ 及其最小值x^*

然後，根據費馬臨界點定理（Fermat's critical point theorem），給出函數最小值的點x^*（稱為最佳解）滿足以下條件：

$$f'(x^*) = 0.$$

因此，我們可以確定滿足方程式$f'(x) = 0$ 的解中哪一個是最佳解。然而，當函數很複雜時，透過解方程式來尋找臨界點並不容易。在這種情況下，臨界點是透過數值方法獲得的。在本節中，我們將探討梯度下降法，這是尋找給定函數最小值的最常用的數值最佳化方法。梯度下降法的基本概念是找出函數的斜率，將其向下方移動，並重複此過程直至達到極值。

梯度下降法演算法

- **步驟1** 設定一個初始疊代值x_1、容差$0 \le \epsilon \ll 1$、初始學習率η (eta)，以及疊代次數$k := 1$。
- **步驟2** 計算$g_k = f'(x_k)$。如果$|g_k| \le \epsilon$，則停止。
- **步驟3** 設定$x_{k+1} = x_k - \eta g_k$，$k := k + 1$ 並回到**步驟2**。

讓我們解釋一下這個GDM 演算法。設定x_k 並計算$g_k = f'(x_k)$。如果$|g_k| < \epsilon$（它在我們允許的誤差範圍內滿足臨界點定理），那麼演算法停止，並給出x_k 作為最佳解。這裡的ϵ（epsilon，容差）滿足$0 \le \epsilon \ll 1$。此不等式中的符號“ $\ll$ ” 表示epsilon 遠小於1。因此，我們給出一個非常接近零的微小容差$\epsilon = 10^{-6}$，並配有一個學習率η。同時設定疊代次數k 為1。如果$|g_k| > \epsilon$，則使用公式確定$x_{k+1} = x_k - \eta g_k$。然後計算$g_k = f'(x_k)$ 以確定是否滿足臨界點定理。如果不滿足，則以類似方式確定下一個值。透過這種方式，我們創建了$x_1, x_2, \ldots (\to x^*)$。如果在經過一些重複步驟$x_{k+1} = x_k - \eta f'(x_k)$ 後，$g_k = f'(x_k)$ 的值位於給定容差內，則演算法停止。我們預期某個x_k 或極限x^* 會滿足$f'(x^*) = 0$。

讓我們看看GDM 的具體運作方式。假設在第k 次疊代後，函數f 於近似解x_k 處的切線斜率為負。如下圖所示，即$g_k = f'(x_k) < 0$。這意味著當x_k 從左向右移動時，函數正在下降，因此我們可以預期x^* 位於x_k 的右側。如果x_k 從左向右移動，那麼x_{k+1} 就會更接近最佳值x^*。因此，我們沿著$-\eta g_k > 0$ 的方向，從x_k 移動到$x_{k+1} = x_k - \eta g_k$。

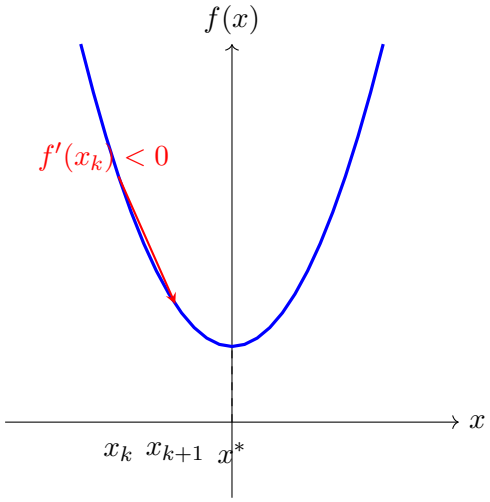

Figure 2: 當$f'(x_k) < 0$ 時的梯度下降步驟

同理，如果在第k 次疊代後，函數f 於近似解x_k 處的切線斜率為正。如下圖所示，即$g_k = f'(x_k) > 0$。這意味著當x_k 從左向右移動時函數在增加，因此我們可以預期x^* 位於x_k 的左側。如果x_k 向左移動，那麼x_{k+1} 會更接近最佳值x^*。因此，我們沿著$-\eta g_k < 0$ 的方向，從x_k 移動到$x_{k+1} = x_k - \eta g_k$。如果我們重複這個過程直到x_k 移動至x^*，那麼$f'(x_k)$ 將會收斂至0。透過這種方式，我們得到了近似解x_k。

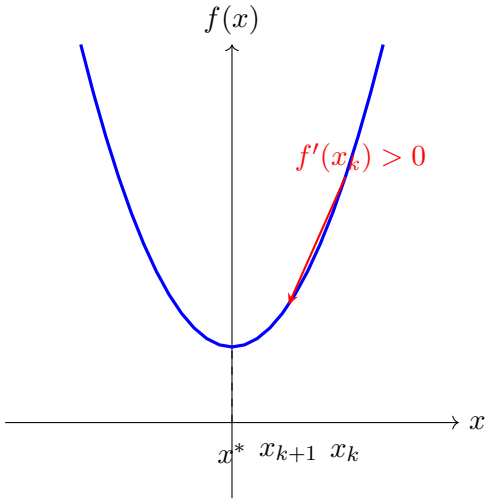

Figure 3: 當$f'(x_k) > 0$ 時的梯度下降步驟

這顯示了GDM 將會找到滿足

$$f(x_1) > f(x_2) > f(x_3) > \dots$$

的$x_1, x_2, x_3, \dots$。這裡的η 決定了移動的幅度。我們稱此為「學習率」（learning rate）。如果學習率過大，x_{k+1} 可能會越過x^*，甚至函數值可能會增加。因此我們應該小心選擇一個合理的

學習率η。

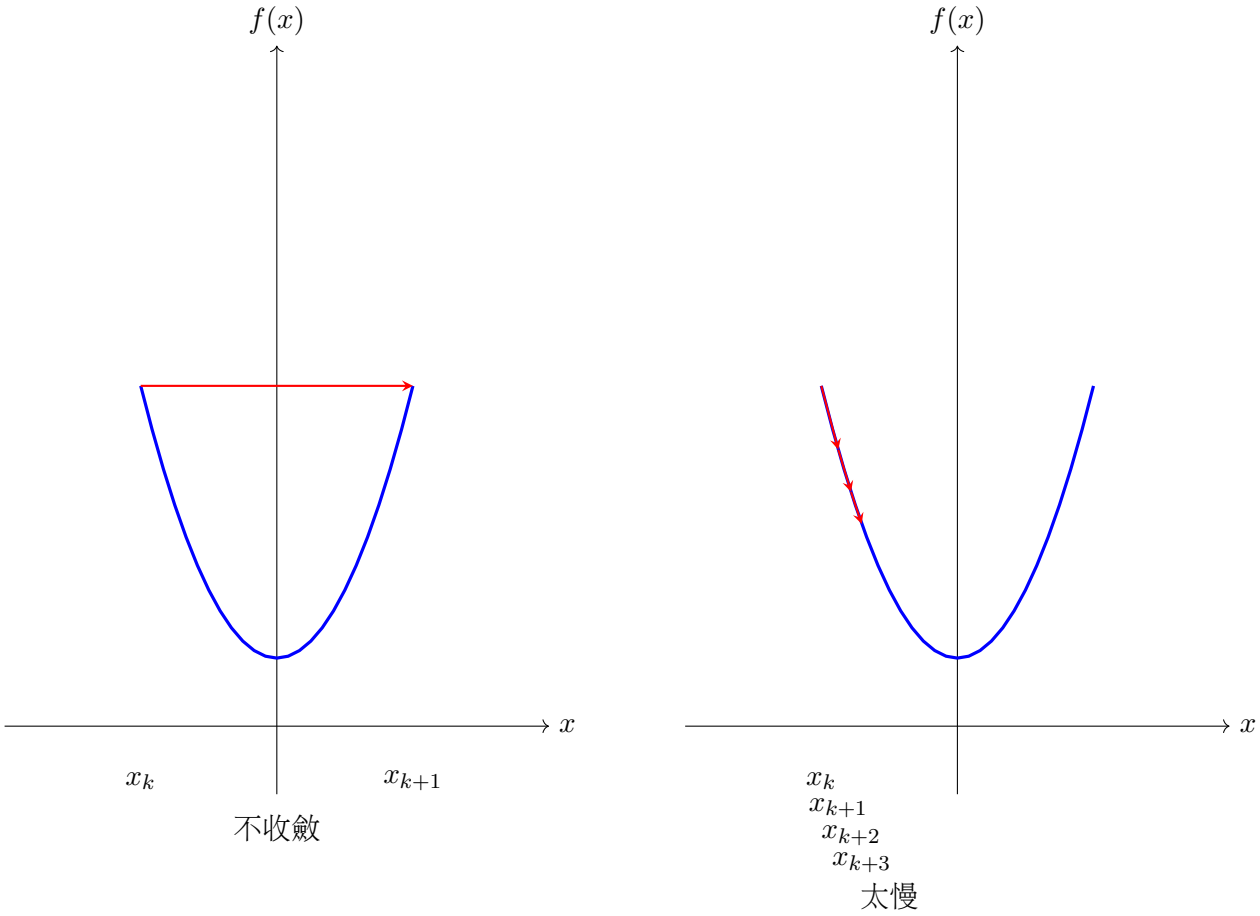

Figure 4: 學習率的影響：過大（左）與過小（右）

參數	**說明**
初始疊代值x_1	演算法的起始點
容差ϵ	$\lvert f'(x_k)\rvert$ 的停止閾值（通常為10^{-6}）
學習率η	更新的步長大小；控制收斂速度/穩定性
更新規則	$x_{k+1} = x_k - \eta f'(x_k)$
收斂條件	$\lvert f'(x_k)\rvert \leq \epsilon$

Table 1: 梯度下降法(GDM) 參數

例子1：單變量函數的梯度下降法

求$f(x) = 2x^2 - 3x + 2$ 的最小值。在以下程式碼中取$x_1 = 0$, $\eta = 0.1$, 以及$\epsilon = 10^{-6}$。

解答

從$f'(x) = 4x - 3 = 0$，得到臨界點$x = \frac{3}{4} = 0.75$。由於$f(x)$ 是凸函數（開口向上），因此在$x = \frac{3}{4}$ 處獲得最小值

$$f\left(\frac{3}{4}\right) = \frac{7}{8} = 0.875$$

。演算法的輸出為：

- 演算法成功！
- $x^* = 0.749999834194560$
- $|g^*| = 6.63221759289456e-7$
- $f(x^*) = 0.875000000000055$
- 疊代次數= 31

在上述例子中，由梯度下降法產生的點$(x_1, f(x_1))$, $(x_2, f(x_2))$, $(x_3, f(x_3))$, ... 連同函數$y = f(x)$ 的圖形一起顯示在坐標平面上。從圖中很容易看出，該數列收斂到了曲線上具有最小值的點$(x^*, f(x^*))$。可以看出該數列收斂至最佳解x^*。

例子1 梯度下降結果摘要

參數	**數值**
函數	$f(x) = 2x^2 - 3x + 2$
導數	$f'(x) = 4x - 3$
初始疊代值 x_1	0.0
容差 ϵ	10^{-6}
學習率 η	0.1
收斂解 x^*	0.75
最小值 $f(x^*)$	0.875
收斂所需疊代次數	31

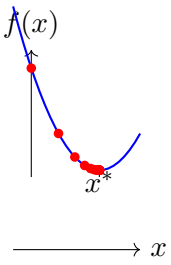

Figure 5: 在$f(x) = 2x^2 - 3x + 2$ 上的梯度下降路徑

例子2：單變量函數的梯度下降法

求$f(x) = 9x^2 - 7x + 6$ 的最小值。利用上述GDM 程式碼，取$x_1 = 0$, $\eta = 0.1$, 以及$\epsilon = 10^{-6}$。

解答 首先，計算導數：

$$f'(x) = 18x - 7.$$

設定$f'(x) = 0$ 得到臨界點：

$$18x - 7 = 0 \implies x^* = \frac{7}{18} \approx 0.3889.$$

由於$f''(x) = 18 > 0$，該函數為凸函數，因此$x^* = \frac{7}{18}$ 是極小值點。最小值為：

$$f\left(\frac{7}{18}\right) = 9\left(\frac{7}{18}\right)^2 - 7\left(\frac{7}{18}\right) + 6 = \frac{167}{36} \approx 4.6389.$$

例子2 梯度下降結果摘要

參數	數值
函數	$f(x) = 9x^2 - 7x + 6$
導數	$f'(x) = 18x - 7$
初始疊代值x_1	0.0
容差ϵ	10^{-6}
學習率η	0.1
真實最小值x^*	$7/18 \approx 0.3889$
最小值$f(x^*)$	$167/36 \approx 4.6389$

梯度下降法的應用

在第一節介紹的GDM 是用於最小化單變量函數$f(x)$。讓我們將該演算法推廣至多變量函數的最小化。我們學過的最小平方問題是一個具有至少兩個自變量的多變量函數$f(\mathbf{x})$。最小平方問題是利用多變量轉換後的誤差函數$E(\mathbf{u})$ 來解決的。最小平方問題也可以利用梯度下降法(GDM) 來解決。

多變量函數最小化的GDM 演算法

- **步驟1** 設定一個初始疊代值$\mathbf{u}_1 \in \mathbb{R}^n$、容差$0 \leq \epsilon < 1$、初始學習率$\eta$ (eta)，以及疊代次數$k := 1$。
- **步驟2** 計算$\mathbf{g}_k = \nabla E(\mathbf{u}_k)$。如果$\|\mathbf{g}_k\| \leq \epsilon$，則停止。
- **步驟3** 設定$\mathbf{u}_{k+1} = \mathbf{u}_k - \eta \mathbf{g}_k$，$k := k + 1$ 並回到**步驟2**。

用於多變量函數最小化的GDM 演算法所有步驟都與單變量函數相同。

單變量與多變量函數的對應關係

單變量函數	多變量函數
純量x	向量$\mathbf{u}$
絕對值$\|a\|$	向量的範數$\|\mathbf{g}\|$
導數$f'(x)$	梯度$\nabla E(\mathbf{u})$

設x, y 為自變量，z 為第三個變量。現在我們可以考慮一個函數$z = f(x, y)$，它依賴於x和y。多變量函數可以同樣的方式定義。在三維（坐標）空間中，$z = f(x, y)$ 可以視為一個點(x, y, z)。當x 和y 變動時，$z = f(x, y)$ 的圖形會變成一個曲面。例如，雙變量函數$z = f(x, y) = -xye^{-x^2-y^2}$ 的圖形可以利用以下程式碼繪製。如圖所示，f 的圖形直觀地顯示f 在波峰處有局部極大值，在波谷處有局部極小值。要找出這些臨界點，需要用到多變量函數的梯度（gradient of a multi-variable function）這個概念。雙變量函數$f(x, y)$ 的梯度定義如下：

$$\operatorname{grad} f(x, y) = \nabla f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}.$$

$f(x, y)$ 的梯度是一個2×1 向量，其中$\frac{\partial f}{\partial x}$ 表示f 對x 的偏導數（partial derivative）。對x 的偏導數是將除了x 以外的其他變量視為常數。同理，$\frac{\partial f}{\partial y}$ 表示f 對y 的偏導數。$z = f(x, y) = -xye^{-x^2-y^2}$ 的梯度可求得如下。

解答

計算出的梯度結果為：

$$(2x^2ye^{-x^2-y^2} - ye^{-x^2-y^2},\ 2xy^2e^{-x^2-y^2} - xe^{-x^2-y^2})$$

$$\operatorname{grad} f(x,y) = \nabla f(x,y) = \left(2x^2ye^{-(x^2+y^2)} - ye^{-(x^2+y^2)},\ 2xy^2e^{-(x^2+y^2)} - xe^{-(x^2+y^2)}\right)$$

計算過程展示

雙變量純量函數$f(x,y)$ 的梯度定義為其一階偏導數向量：

$$\nabla f(x,y) = \left(\frac{\partial f}{\partial x},\ \frac{\partial f}{\partial y}\right)$$

其中∇ 表示nabla 算子，而grad f 是等效的梯度符號。

我們運用微分的**乘積法則**（product rule）來計算$f(x,y) = -xye^{-x^2-y^2}$ 的偏導數：

$$\frac{\partial}{\partial x}(u\cdot v) = \frac{\partial u}{\partial x}\cdot v + u\cdot\frac{\partial v}{\partial x},\quad \frac{\partial}{\partial y}(u\cdot v) = \frac{\partial u}{\partial y}\cdot v + u\cdot\frac{\partial v}{\partial y}.$$

1: 計算對x 的偏導數

設$u = -xy$ 及$v = e^{-x^2-y^2}$。則：

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{\partial}{\partial x}(-xy)\cdot e^{-x^2-y^2} + (-xy)\cdot\frac{\partial}{\partial x}\left(e^{-x^2-y^2}\right)\\ &= -ye^{-x^2-y^2} + (-xy)\cdot e^{-x^2-y^2}\cdot(-2x)\\ &= 2x^2ye^{-(x^2+y^2)} - ye^{-(x^2+y^2)}.\end{aligned}$$

2: 計算對y 的偏導數

設$u = -xy$ 及$v = e^{-x^2-y^2}$。則：

$$\begin{aligned}\frac{\partial f}{\partial y} &= \frac{\partial}{\partial y}(-xy)\cdot e^{-x^2-y^2} + (-xy)\cdot\frac{\partial}{\partial y}\left(e^{-x^2-y^2}\right)\\ &= -xe^{-x^2-y^2} + (-xy)\cdot e^{-x^2-y^2}\cdot(-2y)\\ &= 2xy^2e^{-(x^2+y^2)} - xe^{-(x^2+y^2)}.\end{aligned}$$

多變量函數的梯度 同樣地，可以求得三個或更多變量的多變量函數f 的梯度。一般而言，n 變量函數$f(x_1,x_2,\ldots,x_n)$ 的梯度定義為：

$$\operatorname{grad} f(x_1,x_2,\ldots,x_n) = \nabla f(x_1,x_2,\ldots,x_n) = \begin{bmatrix}\frac{\partial f}{\partial x_1}\\ \frac{\partial f}{\partial x_2}\\ \vdots\\ \frac{\partial f}{\partial x_n}\end{bmatrix}.$$

例子3：梯度下降法求最小平方問題

已知：

x_i	u_i
0	1
1	3
2	4
3	4

最小平方問題的第一個例子如下。對於每個數據(x_i, u_i)，設$\hat{u}_i$ 為將x_i 代入線性函數$u = a + bx$ 所獲得的值。因此：

$$\hat{u}_i = a + bx_i.$$

如果這個方程的解不存在，則可以尋找能使誤差平方$(u_i - \hat{u}_i)^2$ 最小化的a, b。誤差函數$E(\mathbf{u})$ 為平方誤差的總和$(u_i - \hat{u}_i)^2$，其定義如下（為了計算方便而加入了係數$\frac{1}{2}$，且不影響結論）：

$$\begin{aligned} E(\mathbf{u}) = E(a,b) &= \frac{1}{2}\left\{(a-1)^2 + (a+b-3)^2 + (a+2b-4)^2 + (a+3b-4)^2\right\} \\ &= \frac{1}{2}\|A\mathbf{u} - \mathbf{y}\|^2. \end{aligned}$$

1. 矩陣構建 線性模型$u = a + bx$ 給出了設計矩陣A、參數向量$\mathbf{u}$ 和目標向量$\mathbf{y}$：

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}.$$

誤差函數為：

$$E(\mathbf{u}) = \frac{1}{2}\|A\mathbf{u} - \mathbf{y}\|^2.$$

2. 誤差函數的梯度 利用向量微積分，梯度為：

$$\nabla E(\mathbf{u}) = A^T(A\mathbf{u} - \mathbf{y}).$$

按分量展開：

$$\begin{aligned} \frac{\partial E}{\partial a} &= (a-1) + (a+b-3) + (a+2b-4) + (a+3b-4), \\ \frac{\partial E}{\partial b} &= 0\cdot(a-1) + 1\cdot(a+b-3) + 2\cdot(a+2b-4) + 3\cdot(a+3b-4). \end{aligned}$$

3. GDM 更新規則 迭代更新參數向量：

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \eta\,\nabla E(\mathbf{u}_k).$$

當$\|\nabla E(\mathbf{u}_k)\| < \epsilon$ 時停止。利用`Python` 使用GDM 來尋找最小化$E(\mathbf{u})$ 的近似解$\hat{\mathbf{u}}$。設定初始疊代值$\mathbf{u}_1 = (-2.5, -2.5)$、容差$\epsilon = 10^{-6}$ 及初始學習率$\eta = 0.1$。

GDM 最小平方線性擬合摘要

參數	數值
初始疊代值$\mathbf{u}_1$	$(-2.5, -2.5)$
容差ϵ	10^{-6}
學習率η	0.1
收斂解$\mathbf{u}^*$	$(1.5, 1.0)$
最小誤差$E(\mathbf{x}^*)$	0.5
收斂所需疊代次數	118
最小平方線	$y = x + \frac{3}{2}$

因此，我們從最小平方解獲得的直線為

$$y = \frac{3}{2} + x.$$

例子4：梯度下降法求最小平方問題

利用Python 使用GDM 和上述相同的數據，來尋找最能擬合給定數據的二次函數$y = a + bx + cx^2$。我們採用梯度下降法(GDM) 尋找最適合給定數據的二次函數$y = a + bx + cx^2$。誤差函數$E(\mathbf{u})$ 定義為：

$$\begin{aligned} E(a,b,c) &= \frac{1}{2}\left\{(a-1)^2 + (a+b+c-3)^2 + (a+2b+4c-4)^2 + (a+3b+9c-4)^2\right\} \\ &= \frac{1}{2}\|A\mathbf{u} - \mathbf{y}\|^2, \end{aligned}$$

其中$\mathbf{u} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ 是參數向量。

1. 矩陣形式構建 二次模型為$y = a + bx + cx^2$。從誤差函數得知，數據點為：$(x, y) = (0, 1), (1, 3), (2, 4), (3, 4)$。設計矩陣$A$、參數向量$\mathbf{u}$ 和目標向量$\mathbf{y}$ 分別為：

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}.$$

誤差函數為均方誤差（按1/2 縮放）：

$$E(\mathbf{u}) = \frac{1}{2}\|A\mathbf{u} - \mathbf{y}\|^2.$$

2. 誤差函數的梯度（對GDM 至關重要） 用於GDM 更新的梯度$\nabla E(\mathbf{u})$ 為：

$$\nabla E(\mathbf{u}) = A^T(A\mathbf{u} - \mathbf{y}).$$

展開(a, b, c) 的梯度分量：

$$\nabla E = \begin{bmatrix} \frac{\partial E}{\partial a} \\ \frac{\partial E}{\partial b} \\ \frac{\partial E}{\partial c} \end{bmatrix} = A^T(A\mathbf{u} - \mathbf{y}).$$

3. GDM 更新規則 以學習率η 迭代更新參數：

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \eta \cdot \nabla E(\mathbf{u}_k).$$

結果

GDM 收斂至最佳最小平方參數：

$$a \approx 1.05, \quad b \approx 1.45, \quad c \approx -0.25.$$

最佳擬合二次函數為：

$$y = 1.05 + 1.45x - 0.25x^2.$$