1. Consider stochastic gradient descent method to learn the house price model

$$h(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2),$$

where σ is the sigmoid function.

Given one single data point $(x_1, x_2, y) = (1, 2, 3)$, and assuming that the current parameter is $\theta^0 = (b, w_1, w_2) = (4, 5, 6)$, evaluate θ^1.

> Just write the expression and substitute the numbers; no need to simplify or evaluate.

$$\theta^{n+1} = \theta^n - \alpha \nabla_\theta \text{Loss}, \quad \text{Loss} = \frac{1}{2}|y - h(x_1, x_2)|^2$$

$$h(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2) = \sigma(z), \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

Let $z = b + w_1 x_1 + w_2 x_2$

$$\begin{cases} \frac{\partial L}{\partial b} = |y-h| \cdot \left(\frac{-\partial h}{\partial b}\right) \\ \quad = |y-h| \cdot \left(-\frac{\partial \sigma}{\partial z} \cdot \frac{\partial z}{\partial b}\right) = -|y-h|(1-\sigma)\cdot\sigma \\ \frac{\partial L}{\partial w_1} = -|y-h|(1-\sigma)\cdot\sigma x_1 \\ \frac{\partial L}{\partial w_2} = -|y-h|(1-\sigma)\sigma x_2 \end{cases}$$

$$* \ \frac{\partial \sigma}{\partial z} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}}$$

$$= \sigma(1-\sigma)$$

$$\Rightarrow \theta' = \begin{pmatrix} b \\ w_1 \\ w_2 \end{pmatrix} + \alpha \begin{pmatrix} \partial L/\partial b \\ \partial L/\partial w_1 \\ \partial L/\partial w_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + \alpha \begin{pmatrix} -|3-\sigma(21)|(1-\sigma(21))\cdot\sigma(21) \\ -|3-\sigma(21)|(1-\sigma(21))\cdot\sigma(21)\cdot 1 \\ -|3-\sigma(21)|(1-\sigma(21))\cdot\sigma(21)\cdot 2 \end{pmatrix}$$

2. (a) Find the expression of $\frac{d^k}{dx^k}\sigma$ in terms of $\sigma(x)$ for $k = 1, \cdots, 3$ where σ is the sigmoid function.

(b) Find the relation between sigmoid function and hyperbolic function.

(a)

- $\frac{d}{dx}\sigma(x) = \underline{(1-\sigma)\sigma}$ (by 1 (*))

- $\frac{d^2}{dx^2}\sigma(x) = -1\cdot(1-\sigma)\sigma\cdot\sigma + (1-\sigma)(1-\sigma)\sigma$

$$= \sigma^3 - \sigma^2 + \sigma - 2\sigma^2 + \sigma^3$$

$$= \underline{2\sigma^3 - 3\sigma^2 + \sigma}$$

- $\frac{d^3}{dx^3}\sigma(x) = 6\sigma^2\cdot\sigma' - 6\sigma\cdot\sigma' + \sigma'$

$$= \underline{\sigma'(6\sigma^2 - 6\sigma + 1)}$$

(b) $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1} = \frac{2e^{2x}}{e^{2x}+1} - \frac{e^{2x}+1}{e^{2x}+1}$

$$= 2\sigma(2x) - 1$$

$$\Rightarrow \tanh\left(\frac{x}{2}\right) = 2\sigma(x) - 1$$

$$\Rightarrow \sigma(x) = \frac{\tanh\left(\frac{x}{2}\right) + 1}{2}$$

3. There are unanswered questions during the lecture, and there are likely more questions we haven't covered. Take a moment to think about them and write them down here.

Q：Mini-Batch SGD 的 batch size 要如何選擇？

🔷 **1. 理論面**

Mini-Batch SGD 的更新公式：

$$\theta^{t+1} = \theta^t - \eta \cdot \frac{1}{B}\sum_{i=1}^{B} \nabla_\theta L(x_i, y_i; \theta)$$

其中 B 是 **batch size**。

- **小 batch** (例如 $B = 1$)：
  - 更新方向比較「噪聲大」
  - 每次更新很快 (少量資料就能更新)
  - 但收斂路徑會抖動，可能更容易跳出局部最小值。
- **大 batch** (例如 $B = N$，全量資料)：
  - 更新方向接近「真實梯度」，比較平滑
  - 每次更新需要很多計算，收斂慢
  - 容易卡在 sharp minima (泛化能力可能較差)

👉 **理論上**，batch size 的選擇其實是「bias–variance trade-off」：

- 小 batch → 高 variance, 低 bias
- 大 batch → 低 variance, 高 bias

🔷 **2. 經驗法則**

1. **常見選擇**：
   - 在電腦資源允許下，**通常選 32 或 64**，有時也用 128、256。
   - 這些數字通常跟 GPU/TPU 的向量化計算有關 (2 的倍數快)。
2. **學習率關係**：
   - Batch size 大時，可以適度 **提高學習率**，因為梯度估計更穩定。
   - 有個經驗公式叫 **linear scaling rule**：

$$\eta_{new} \approx \eta_{old} \cdot \frac{B_{new}}{B_{old}}$$

3. **資料集大小影響**：
   - 如果資料集很小，可以用全量 batch (batch = dataset)。
   - 如果資料集很大，通常會選擇「不大不小」的 batch (32~256)。
4. **研究觀察**：
   - 論文 (例如 Keskar et al. 2017) 發現：大 batch 可能收斂到「sharp minima」，泛化比較差。
   - 小 batch 帶來的「梯度噪聲」反而有助於找到「flat minima」，泛化比較好。

🔷 **3. 總結**

- 沒有一個唯一正解，但常見範圍是 **32~256**。
- 小 batch → 泛化好，但更新不穩定。
- 大 batch → 計算效率好，但可能泛化差。
- 實務上通常從 **32 或 64** 開始試，然後依 GPU 記憶體調整。

OpenAI. (2025)。ChatGPT (GPT-5) [大型語言模型]。取自 https://chat.openai.com/