一、All unanswered questions

1. How to decide appropriate learning rate?

   Ans: learning rate 的選擇有以下幾種方法:

   (1) 試錯法

   通常會從幾個常見的數量級開始嘗試(例如:0.1, 0.01, 0.001, 0.0001)。

   - 觀察 Loss 曲線:

     Loss 爆炸或呈現 NaN:學習率太高。

     Loss 曲線震盪劇烈且無法下降:學習率偏高。

     Loss 下降極其緩慢:學習率太低。

     Loss 平穩下降:學習率適中。

   (2) 學習率衰減

   策略:一開始使用較大的學習率(如 0.1)以快速接近最佳解，隨著訓練進行(Epoch 增加)，逐漸降低學習率以進行微調。

   (3) 學習率搜尋器

   這是一種較為科學的自動化尋找方法(常見於 fast.ai 等庫):

   從一個非常小的學習率開始訓練，在每個 Batch 或 Step 中指數級增加學習率，記錄對應的 Loss 變化並畫圖。

   如何選擇:找出 Loss 下降最快(斜率最陡)的那一段對應的學習率，而不是 Loss 最小的點。通常會選擇下降最快點的數值再稍微除以 10 作為起始值。

   參考資料:

   https://medium.com/data-science/estimating-optimal-learning-rate-for-a-deep-neural-network-ce32f2556ce0

   https://towardsdatascience.com/how-to-choose-the-optimal-learning-rate-for-neural-networks-362111c5c783/

   https://medium.com/@sahin.samia/how-to-choose-the-right-learning-rate-in-deep-learning-with-pytorch-690de782b405

2. How does GDA perform if the true class-conditional distributions are significantly non-Gaussian?

   Ans:

   當 Gaussian Discriminant Analysis (GDA)遇到 Significantly Non-Gaussian 的類別條件機率分佈時，其表現通常會變差，甚至可能完全失效。

以下是具體的分析，解釋為什麼 GDA 在這種情況下表現不佳，以及可能會發生什麼事:

(1) 高偏差與模型錯誤設定

GDA 對數據的分佈做了極強的假設:它假設每個類別的數據都呈現「單峰」且形狀為「橢圓體」的高斯分佈。

- 數據是多峰的:

  如果 Class A 的數據分佈在兩個分開的群聚中，GDA 會強行計算出一個位於這兩群中間的「平均值」。這個平均值可能落在完全沒有數據的區域，導致模型認為該區域是 Class A 機率最高的地方，這顯然是錯誤的。

- 如果數據是 Non-convex 或流形結構:

  例如數據呈現「香蕉形」或「環形」，用一個高斯橢圓去擬合它，無法捕捉到數據的幾何特徵，導致大量的訊息丟失。

(2) 決策邊界的限制

GDA 的數學推導決定了其決策邊界的形狀:

- LDA:假設共用共變異數矩陣，邊界必定是線性的。
- QDA:允許不同的共變異數矩陣，邊界必定是二次曲線(圓、橢圓、拋物線、雙曲線)。

如果真實的類別邊界是非常複雜的、波浪狀的、或者不規則的，GDA 的線性和二次邊界根本無法將其分開，導致嚴重 Underfitting。

參考資料:

cs229-notes2.pdf

1.2. Linear and Quadratic Discriminant Analysis — scikit-learn 1.7.2 documentation

3. If the true decision boundary in real-world data is nonlinear, how much difference in classification error can typically be expected between LDA and QDA? What factors influence this error difference?

Ans:這兩者的誤差差距並非固定不變，而是由以下三個因素共同決定:

(1) 共變異數矩陣的差異程度這是最核心的因素:

$\Sigma_1 \approx \Sigma_2$:如果兩個類別的形狀(橢圓)大小和方向差不多，決策邊界會很接近直線，此時 LDA 和 QDA 表現幾乎一樣。

$\Sigma_1 \neq \Sigma_2$:如果 Class A 是一個窄長的橢圓，而 Class B 是一個寬圓，真

實邊界會是一條拋物線或雙曲線，差異越大，LDA 的線性假設誤差會更大，誤差差距就越大。

(2) 樣本數

這是 Bias-Variance Tradeoff 的體現。

- 小樣本:即使真實邊界是非線性的，LDA 有時反而表現得比 QDA 好，因為 QDA 需要估計更多的參數(每個類別都有獨立的矩陣 Σ_k)，在數據不足時，QDA 的變異數會爆增，導致過度擬合。此時 LDA 的「穩定性」勝過 QDA 的「靈活性」。
- 大樣本:QDA 的優勢會完全展現，隨著數據量增加，QDA 能精準描繪出彎曲的邊界，而 LDA 則會受限於其線性偏差，誤差率會停留在較高水平無法下降。

(3) 維度

QDA 的參數數量隨著維度平方增長($O(Kp^2)$)，在高維度(p很大)情況下，除非你有海量的數據，否則 QDA 極易失效，此時 LDA 雖然模型是錯的(線性的)，但因為參數少、夠強健，反而可能誤差更低。

參考資料:

An Introduction to Statistical Learning (ISL)，Chapter 4 (Classification), Section 4.4.4 "Quadratic Discriminant Analysis"。

Linear and Quadratic Discriminant Analysis with covariance ellipsoid — scikit-learn 1.7.2 documentation

4. score-based generative models 是透過 Score Function 來引導生成預測數據。但目前方法類似於梯度上升，每一步都採用預設的、保守的小步長，有沒有方法也找到或訓練出曲率資訊，來增加預測效率。

Ans: 透過 Probability Flow ODE 使用高階數值解算器

Score-based models (SDE)可以轉換為一個確定性的 Probability Flow ODE (PF-ODE)，其邊際分佈與原 SDE 相同。既然轉換成了 ODE，我們就可以利用高階自適應解算器(Adaptive Solvers)。

- 原理:不要用固定步長的 Euler-Maruyama (類似 SGD)，改用 Runge-Kutta 系列方法(如 RK45, DOPRI)。
- 如何利用曲率:這些解算器內部會計算「局部誤差估計」。

  如果局部曲率(Curvature)很大(誤差估計高)，解算器會自動縮小步長以保證精確。

如果局部非常平坦(線性度高)，解算器會自動加大步長，直接跨過一大段距離。

- 結果:這變相實現了「利用曲率調整效率」。比起固定步長需要 1000 步，RK45 可能只需要 20-50 步就能生成高品質圖像。

參考資料:

Score-Based Generative Modeling through Stochastic Differential Equations (ICLR 2021)

DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Models Sampling in Around 10 Steps (NeurIPS 2022)

5. 不同的加噪過程，會訓練出不同的解噪模型，那在不同的實務應用中，該如何選擇出適當的加噪過程?

Ans:

選擇適當的加噪過程(Forward SDE)，本質上是在數據結構、數值穩定性與採樣效率之間做權衡。

以下是根據不同實務應用場景，選擇 SDE 類型(如 VE, VP, sub-VP 等)的決策框架:

(1) 根據數據的邊界特性

- 有界數據:傾向使用 VP-SDE (Variance Preserving)

  應用場景: 圖像生成(像素值通常歸一化在$[-1, 1]$或$[0, 1]$)、影片生成。

  原因:VP-SDE (對應 DDPM)的特性是隨著噪聲增加，數據的訊號會同時衰減，使得總變異數保持在固定範圍，這對於有界數據非常自然，因為它不會讓數值無限制地發散，訓練神經網絡時輸入值的範圍也較為穩定。

- 無界數據:傾向使用 VE-SDE (Variance Exploding)

  應用場景:音訊波形合成、某些物理訊號模擬、分子結構生成。

  原因:VE-SDE (對應 SMLD/NCSN)只加噪聲而不縮放數據，對於音頻或某些物理量，數值範圍可能很大或沒有明確邊界，VP 的強制縮放可能會破壞訊號原本的物理意義或動態範圍。VE 允許變異數隨時間變大，能夠更好地覆蓋高動態範圍的數據流形。

(2) 根據任務目標:生成質量 vs.似然估計

- 追求生成質量: VE-SDE 或 VP-SDE

目前最先進的圖像生成模型(如 Stable Diffusion, Imagen)多基於 VP-SDE 或其變體。VE-SDE 在某些高解析度紋理生成的任務上表現也非常優異。

- 追求最大似然估計:sub-VP SDE

  定義:sub-VP SDE 的加噪過程介於 VP 和確定性過程之間，它的變異數增長比 VP 更受到抑制。

  原因:研究顯示(如 Song et al., ICLR 2021)，sub-VP SDE 在計算 Log-likelihood 時往往能取得比 VP 和 VE 更好的分數。如果你是用擴散模型來做異常檢測或壓縮，這通常需要精確的 NLL，此時 sub-VP 是好選擇。

(3) 特殊領域:非歐幾里得幾何與逆問題

- 幾何數據:

  如果你要生成的數據位於黎曼流形上(例如:地球科學數據在球面上、蛋白質骨架在 SO(3)群上)，標準的高斯噪聲 SDE 就不適用了。

  選擇:必須使用 Riemannian Diffusion Models，其加噪過程被定義為流形上的布朗運動。

- 逆問題(去模糊、超解析度):

  雖然標準 SDE 可用，但最近出現了 Blurring Diffusion (Cold Diffusion)。

  選擇:使用基於 Heat Dissipation 的加噪過程，而不是加高斯白噪聲。這在做去模糊時，能更好地保留圖像的結構語義。

  參考資料:

  Score-Based Generative Modeling through Stochastic Differential Equations

  Riemannian Score-Based Generative Modelling

  Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise

6. 該如何評鑑一個 SDE 使否適合某一個實務應用，且如果不適合該如何改進?

   Ans:

   評鑑的方法有以下幾種:

   (1) 殘差分析

   如果模型是正確的，將真實數據代入模型後，剝離掉漂移項並標

準化後的「殘差」應該要符合標準常態分佈的特性($dW_t \sim \mathcal{N}(0, dt)$)。

- 檢查方法:計算$\widehat{\epsilon_t} = \frac{X_{t+1} - X_t - f(X_t)\Delta t}{g(X_t)\sqrt{\Delta t}}$ 。
- 判斷標準:
  - $\widehat{\epsilon_t}$的直方圖是否接近常態分佈？(使用 Q-Q Plot 或 Shapiro-Wilk test)。
  - $\widehat{\epsilon_t}$是否存在自相關？(使用 ACF 圖檢查)，如果殘差之間有關聯，代表模型漏掉了某些動力學特徵(Memory effect)。

(2) 邊際分佈與穩態分佈

- 長期行為:實務數據在長時間下是否會有一個穩定的分佈？
- 檢查方法:模擬 SDE 生成長路徑，畫出其機率密度函數，看是否與真實數據的 Histogram 重疊，例如，Ornstein-Uhlenbeck (OU)過程會有高斯穩態分佈，Cox-Ingersoll-Ross (CIR)過程則是 Gamma 分佈。

(3) 樣本路徑的定性特徵

- 邊界條件:數據是否有自然邊界？(例如濃度不能為負)。如果 SDE 經常穿過不該穿過的邊界，則不適合。
- 跳躍:真實數據是否有瞬間的暴漲暴跌？標準 SDE 是連續路徑，無法捕捉「黑天鵝」事件。
- 波動叢聚: 波動率是恆定的，還是會隨時間忽大忽小？

(4) 擬合優度指標

- AIC / BIC:如果在比較多個 SDE 模型，使用赤池信息量準則(AIC)來平衡模型的準確度與複雜度。
- Likelihood Ratio Test:計算模型在測試數據上的 Log-Likelihood，數值越高越好。

改進的方法有以下幾種:

(1) 修正漂移項$f(X_t)$(針對趨勢與回歸特性)

- 問題: 模型無法捕捉數據的「回歸」或「趨勢」行為。
- 改進:

  增加均值回歸:如果數據傾向回到某個水準，引入像 OU process 的項:$-\lambda(X_t - \mu)$。

  引入非線性:如果系統有雙穩態特徵(例如開關切換)，可能需要多項式漂移項，如$X_t(a - X_t^2)$。

(2) 修正擴散項$g(X_t)$(針對波動結構)

- 問題:殘差變異數不穩定，或者波動率隨狀態改變。
- 改進:

  狀態相依波動:設定$g(X_t) = \sigma X_t^{\gamma}$(如 CEV 模型)。當$X_t$變大時波動變大。

  隨機波動:如果波動率本身的變化與X_t無關，則需要將波動率建模為另一個 SDE(例如 Heston Model)。

(3) 改變噪聲源(針對厚尾與記憶性)

- 問題:數據有極端值或殘差有長記憶性。
- 改進:

  引入跳躍:在 SDE 中加入 Poisson Process 項dN_t，變成$dX_t = fdt + gdW_t + JdN_t$，用來模擬突發事件。

  分數布朗運動:如果數據有長期依賴性(Hurst exponent $H \neq 0.5$)，將標準dW_t替換為dB_t^H。

(4) 現代方法-神經隨機微分方程

- 問題:物理機制太複雜，無法寫出明確的數學式f或g。
- 改進:利用深度學習參數化漂移和擴散項。

$$dX_t = NN_f(X_t)dt + NN_g(X_t)dW_t$$

  作法:使用 Adjoint Method 或 Score Matching 來訓練神經網絡，讓它從數據中學出最佳的微分方程結構。

  參考資料:

  Testing Continuous-Time Models of the Spot Interest Rate

  Evaluating methods for approximating stochastic differential equations

  Neural Stochastic Differential Equations: Deep Latent Gaussian Models in the Diffusion Limit

二、Toy model/Solvable Model Problem in final project

合成的運動意圖腦波數據解碼模型

1. 簡化模型概念與最終能力的連結

在此 Toy Model 中，將目標中「連續的多維度控制(如手指細微動作)」簡化為「離散的二元分類問題(左手 vs.右手)」，並將難以取得的真實大腦訊號，替換為「基於生物物理機制生成的合成 EEG 數據」，這個簡化模型在概念上代表了最終能力的核心邏輯：「AI 能否透過分析電訊號的頻譜能量變化，逆向推導出隱藏在大腦中的運動意圖？」。

2. 生物物理機制與數據生成策略

為了確保合成數據具有科學意義，不使用隨機亂數，而是模擬真實大腦運動皮質區的「事件相關去同步化(ERD)」現象。

- 生理原理：人體運動受大腦「對側支配」，當一個人想像移動右手時，左腦運動區(C3 通道)的 Mu 波(8-13Hz)振幅會顯著下降，反之，想像左手時，右腦(C4 通道)的振幅會下降。
- 數據生成模型：我們利用 Python 構建雙通道訊號模擬器，公式如下：

$$Signal(t) = Noise(t) + A(intent) \cdot \sin(2\pi \cdot 10Hz \cdot t)$$

  - 我們設定 $10Hz$為 Mu 波的中心頻率。
  - $A(intent)$是振幅控制變數：系統會根據當下的「意圖標籤」，動態抑制對側通道的振幅(模擬 ERD 現象)，產生具備特徵的虛擬腦波。

3. 解決方案所需的數學與機器學習工具

針對上述生成的合成數據，我們設計了一套標準的解碼流程：

(1) 訊號處理(數學工具)：

- 帶通濾波：雖然數據是我們生成的，但為了模擬真實情境，仍會使用濾波器鎖定 8-13Hz 頻段，濾除刻意加入的高斯白雜訊。
- 能量特徵計算：計算 C3 與 C4 通道在時間窗內的訊號功率，將時間序列資料轉化為數值特徵。

(2) 機器學習分類器：

由於這是二元分類且特徵明確(能量高低)，我們採用輕量級的「邏輯迴歸」或簡單的「閾值判別法」，這模擬了未來植入式晶片因功耗限制，必須使用高效演算法的情境。

4. 模型的測試性與成功指標

此 Toy Model 具有極高的可測試性，因為數據是由我們生成的，因此擁有正確的標準答案。

- 驗證方法：生成 100 筆模擬試驗數據(50 筆左手意圖、50 筆右手意圖)，並按 8:2 分割為訓練集與測試集。
- 成功指標：
  - 分類準確率：在測試集中，AI 判斷意圖的準確率需達到 90% 以上(考慮到我們加入的雜訊干擾)。
  - 特徵驗證：繪製出訊號波形圖，確認「想像右手」時，左腦通道的波形確實呈現扁平狀(被抑制)，視覺化驗證我們的物理模型假設是否成立。