

- **How should one choose the finite-difference step h to balance truncation error $O(h^2)$ and numeric round-off / noise amplification in practice?**

Ans:

下面分三部分回答：(A) 如何選 h (包含推導／直觀公式)；(B) 實務建議 (含噪音、複數步法等替代法)；(C) 相關研究與可驗證的參考文獻 (每一項都附來源連結)。

A. 如何選 h (理論／推導 — 中心差分的一階導數為例)

對中心差分 (最常用的一階近似)

$$\hat{f}'(x) = \frac{f(x+h) - f(x-h)}{2h},$$

總誤差可以分成兩項：

- 截斷 (Taylor) 誤差： $E_T \approx \frac{f^{(3)}(\xi)}{6} h^2$ (為 $O(h^2)$)，
- 捨入 / 評估誤差 (round-off 或輸入噪音在數值上的效應)：若單次函數評估的絕對誤差約為 ε_f ，則差分的捨入項大約為 $E_R \approx \frac{\varepsilon_f}{h}$ 。

總絕對誤差 (粗略模型)：

$$E_{\text{tot}}(h) \approx Ah^2 + \frac{B\varepsilon_f}{h},$$

其中 $A \approx |f^{(3)}(x)|/6$ ， B 是常數 (通常約為 1 或 2，視誤差來源如何累積)。對 $h > 0$ 最小化 $E_{\text{tot}}(h)$ ：

$$\frac{dE_{\text{tot}}}{dh} = 2Ah - \frac{B\varepsilon_f}{h^2} = 0 \Rightarrow h_{\text{opt}} = \left(\frac{B\varepsilon_f}{2A}\right)^{1/3}.$$

直觀化簡（以機器精度 $\varepsilon_{\text{mach}}$ 為主）：

若把 ε_f 取為「可視為」 $\varepsilon_{\text{mach}}$ （double 約 2.2×10^{-16} ）且把常數合併，常見結論是

- 中心差分的一階導數最佳量級為 $h \sim \varepsilon_{\text{mach}}^{1/3}$ （常見敘述：cube-root of machine epsilon）。[維基百科+1](#)

對向前差分（forward difference，截斷誤差 $O(h)$ ）類似分析會導致 $h_{\text{opt}} \sim \varepsilon_{\text{mach}}^{1/2}$ 。（因此中心差分在許多情形下更耐舍入誤差並且達到更小的最小誤差。）[維基百科](#)

B. 實務建議

1. 實務快速規則（summary）

- 若你用中心差分且函數評估精度主要受浮點誤差限制，選 h 的量級約為 $\varepsilon_{\text{mach}}^{1/3}$ （乘上與 f 與 $f^{(3)}$ 有關的尺度因子）。若用向前差分，則約 $\varepsilon_{\text{mach}}^{1/2}$ 。[維基百科](#)
- 若資料含觀測噪音（measurement noise）或估值誤差 σ 明顯大於機器精度，則把 ε_f 換成 σ （或 $\sqrt{\text{Var}}$ ），公式變為 $h_{\text{opt}} \propto (\sigma / |f^{(3)}|)^{1/3}$ ：也就是更大噪音 → 選更大的 h （以降低差分放大的噪音）。有關 plug-in 與自適應估計步長的實作在現有工具包中也有（參見下一節）。[CRAN+1](#)

2. 不知道 $f^{(3)}$ 時怎辦？（實用做法）

- 常見做法是用「量級估計」（guesstimate）或用更粗的差分 stencil 去估算高階導數，再把它放入 plug-in 公式（這是很多數值套件實作的方式）。或用自適應方法：在一組 h 值上計算誤差曲線（log-log 看 V 形）選取靠近最低點的 h 。[CRAN+1](#)

3. 若能在複數平面評估 f （解析延拓）→ 用複數步（complex-step）

- 複數步微分 (complex-step method) 能避免相減導致的災難性抵消 (subtractive cancellation)，得到高精度且不受同樣的舍入放大問題限制 (第一導數可用 $\Im(f(x + ih))/h$ ，誤差為 $O(h^2)$ ，但不會產生前述的 $1/h$ 舍入爆炸)。若你的 f 支援複數輸入，這常常是首選。 epubs.siam.org+1

4. 高階導數或高精度情況

- 高階導數 (或高 p) 的 finite-difference 公式條件數變差，誤差模型更複雜；若 f 解析且能用複數/多複數技巧 (multicomplex / Squire–Trapp / Fornberg 方法)，數值穩定性可以大幅改善。
University of Colorado Boulder+1

5. 噪音 vs. 捨入 — 概念區別

- 捨入誤差 (machine rounding)：會出現在非常小 h 時，由於相減導致有效位數損失；模型中通常代入 ϵ_{mach} 。
- 觀測/測量噪音：如果輸入值本來就帶有隨機噪音 σ ，差分會把該噪音放大 (尺度像 σ/h)，所以選 h 時應以 $\epsilon_f \approx \max(\epsilon_{\text{mach}} \times |f|, \sigma)$ 來替代。 CRAN

C. 研究文獻

1. Nicholas J. Higham — "Differentiation With(out) a Difference" (SIAM News, 2018).

- 易讀、概念性強的討論，說明了 truncation vs rounding 與複數步法 (complex-step) 的優點。
- 連結 (SIAM News) : siam.org
<https://www.siam.org/publications/siam-news/articles/differentiation-without-a-difference/>

2. W. Squire & G. Trapp — "Using Complex Variables to Estimate Derivatives of Real Functions" (SIAM Review, 1998).

- 提出 complex-step derivative，實作且能避免相減的取消誤差。

非常常被引用。

- DOI / 出版頁面 (SIAM) : [epubs.siam.org+1](https://epubs.siam.org/doi/10.1137/S003614459631241X)
<https://epubs.siam.org/doi/10.1137/S003614459631241X>
(或作者學校存檔 PDF)
https://researchrepository.wvu.edu/faculty_publications/426/

3. **B. Fornberg — "Numerical Differentiation of Analytic Functions" (ACM TOMS, 1981) & "Generation of finite difference formulas..." (Math. Comp., 1988).**

- 對高階導數、解析函數及複數/多複數延拓的演算法討論詳盡，適合要處理高階導數與穩定性問題的人。 [University of Colorado Boulder+1](https://www.colorado.edu/amath/sites/default/files/attached-files/acm_81_1.pdf)
https://www.colorado.edu/amath/sites/default/files/attached-files/acm_81_1.pdf
<https://www.ams.org/journals/mcom/1988-51-184/S0025-5718-1988-0935077-0/>

4. **實作 / 套件與自適應選 h 的文獻與工具** (例如 R 套件 pnd, Julia 的 FiniteDiff 等) — 這些工具說明了 plug-in 與 empirical approaches 去估 h_{opt} 。

- pnd 套件文件 (CRAN) 說明 plug-in 估算步長的做法。 [CRAN+1](https://cran.r-project.org/web/packages/pnd/pnd.pdf)
<https://cran.r-project.org/web/packages/pnd/pnd.pdf>
<https://docs.sciml.ai/FiniteDiff/dev/epsilons/>

5. **綜合介紹 (教科書/教學資料)** : Higham 的教科書《Accuracy and Stability of Numerical Algorithms》、以及大學數值分析講義，這些都系統性討論截斷 vs 捨入的 trade-off。 [بيان صندوق+1](#)

- **The error bounds shown are $O(h^2)$ etc.; what constants (dependent on σ and domain of x) control those bounds?**
How do they scale with p and $|x|$?

Ans:

1) 常數從哪裡來 — 一個一般性公式 (直觀 + 嚴格框架)

設

$$\phi(w) = \sigma(wx + b),$$

我們用以 w 為自變量的有限差分去近似 $\phi^{(p)}(0)$ 。因為 ϕ 是由 σ 與線性內函數 $w \mapsto wx + b$ 組成，連鎖律給出

$$\phi^{(p)}(w) = \sigma^{(p)}(wx + b) x^p.$$

當我們用一個二階精度 ($O(h^2)$) 的對稱中心差分配方來近似 $\phi^{(p)}(0)$ (例如用 $w = \pm h$ 的三點 stencil 或其他產生二階誤差的對稱 stencil)，Taylor 展開的餘項告訴我們誤差有形如

$$\text{誤差} = C_{p,\text{stencil}} \sigma^{(p+2)}(\xi) x^{p+2} h^2,$$

對某個 ξ 在 w 的鄰域 (具體在 $w \in [-h, h]$ 對應到 σ 的自變量區間) 內。換成上界表示可寫成

$$|\text{誤差}| \leq C_{p,\text{stencil}} \left(\max_{t \in I} |\sigma^{(p+2)}(t)| \right) |x|^{p+2} h^2,$$

其中 $I = \{b + wx : w \in [-h, h]\}$ (也就是 σ 在被取樣區間上的 $(p+2)$ -階導數之最大值)，而常數 $C_{p,\text{stencil}}$ 只取決於你使用的 finite-difference 權重 (stencil 結構) 與 p (但與 h 無關)。這是一般形式：誤差的大小被兩個主要成分控制——(A) $\sigma^{(p+2)}$ 在所用區間上的大小；(B) 因為內函是 $wx + b$ ，每多做一階導數會帶出一個 x 因子，所以導致額外的 $|x|^{p+2}$ 因子。關於 finite-difference 權重與誤差常數的系統性推導，可見 Fornberg 的工作 (見下方參考)。[University of Colorado Boulder+1](#)

重點句：

「差分誤差是『activation 的高階導數大小』乘上『輸入尺度的高次方』

x^{p+2} 』再乘上與所用差分模板有關的一個純數值常數，最後乘上 h^2 。」

2) 兩個常見具體例子（精確常數）

$p = 1$ （用三點中心差分 $\frac{\phi(h) - \phi(-h)}{2h}$ 近似 $\phi'(0)$ ）

標準結果（由 Taylor 展開）：

$$\frac{\phi(h) - \phi(-h)}{2h} - \phi'(0) = \frac{\phi^{(3)}(\xi)}{6} h^2$$

對某 $\xi \in (-h, h)$ 。把 $\phi^{(3)}(\xi) = \sigma^{(3)}(\xi)x + b$ 代入，得到

$$\text{誤差} = \frac{\sigma^{(3)}(\tilde{\xi})}{6} x^3 h^2,$$

所以可以以上界寫為

$$|\text{誤差}| \leq \frac{1}{6} (\max_{t \in I} |\sigma^{(3)}(t)|) |x|^3 h^2.$$

（這是教科書上常見的公式。） dam.brown.edu

$p = 2$ （用三點中心差分 $\frac{\phi(h) - 2\phi(0) + \phi(-h)}{h^2}$ 近似 $\phi''(0)$ ）

標準結果：

$$\frac{\phi(h) - 2\phi(0) + \phi(-h)}{h^2} - \phi''(0) = \frac{\phi^{(4)}(\xi)}{12} h^2.$$

代入 $\phi^{(4)} = \sigma^{(4)}(\cdot)x^4$ 得

$$\text{誤差} = \frac{\sigma^{(4)}(\tilde{\xi})}{12} x^4 h^2,$$

因此上界為

$$|\text{誤差}| \leq \frac{1}{12} (\max_{t \in I} |\sigma^{(4)}(t)|) |x|^4 h^2.$$

(這也是常見、可在多數數值分析教材找到的常數。)dam.brown.edu

註：上面 $1/6, 1/12$ 是對應到「三點對稱中心差分」的具體常數；若你用更多點或不同 stencil，常數 $C_{p,\text{stencil}}$ 會不同，但仍是可計算的固定數 (Fornberg 的算法可給出任意 stencil 的權重與誤差行為)。University of Colorado Boulder

3) p 與 $|x|$ 的尺度如何影響常數 (更直觀的說)

- 每一階對 w 的導數會帶來一個 x 因子：因為 $\frac{d}{dw} \sigma(wx + b) = \sigma'(wx + b)$ 。因此 $\phi^{(k)}(w)$ 中會包含 x^k 。這就是為何誤差會出現 $|x|^{p+2}$ 而非單純 $|x|^p$ ：誤差是來自比 p 還高 2 階的導數項 (因為我們求二階截斷項)，所以會乘上 x^{p+2} 。
 → 結果：當 $|x|$ 增大時，誤差會以 **多項式速度** (約 $|x|^{p+2}$) 放大。
- 隨 p 增大常數怎麼變？
 - $\sigma^{(p+2)}$ 的數值大小通常會隨 p 增大而變化——對某些 activation (例如 Gaussian 或 tanh) 高階導數會快速成長/震盪，對另一類 (如多項式型 activation) 則行為不同。也就是說，實際上 p 的增大會透過兩個途徑放大誤差：一是把 $|x|^{p+2}$ 提高；二是通常 $\max |\sigma^{(p+2)}|$ 也可能增大。
 - 從理論上，若你使用固定 stencil (固定 $C_{p,\text{stencil}}$) 且能保證 $\max_I |\sigma^{(p+2)}| \leq M_{p+2}$ ，則誤差上界會像 $\sim M_{p+2} |x|^{p+2} h^2$ 。因此高 p 的情況數值不穩定性與誤差放大是常見現象。關於高階導數數值微分與多複數複雜技術的討論，可參閱 Fornberg 與 multicomplex 的文獻。University of Colorado Boulder+1

4) 研究文獻

- B. Fornberg, "Generation of finite difference formulas on arbitrarily

spaced grids” (Math. Comp., 1988).

- 描述如何產生任意 stencil 的權重與誤差行為（可用來算出 $C_{p,\text{stencil}}$ ）。 [University of Colorado Boulder](#)
- PDF（可核實）：
https://www.colorado.edu/amath/sites/default/files/attached-files/mathcomp_88_fd_formulas.pdf

- **M. S. Floater, “Error formulas for divided difference expansions and numerical differentiation” (J. Approx. Theory, 2003).**

- 給出除法差商（divided difference）與數值微分誤差的更一般界值，可用於分析高階導數近似的餘項。 [科學直接](#)
- 網頁：
<https://www.sciencedirect.com/science/article/pii/S002190450300025X>

- **R. J. LeVeque, “Finite Difference Methods for Ordinary and Partial Differential Equations” / 教科書與講義**（常見教材，系統討論截斷誤差、穩定性與誤差常數）。 [ResearchGate](#)

- **Squire & Trapp (1998), “Using Complex Variables to Estimate Derivatives of Real Functions” (SIAM Review)** — 關於 complex-step method，討論為何能避免相減取消與改善誤差常數與穩定性（當 σ 可解析時非常有用）。 [researchrepository.wvu.edu](#)

- **Higham, “Differentiation With(out) a Difference” (SIAM News, 2018)** — 概念性討論截斷 vs 捨入與 complex-step 方法的優勢。 [siam.org](#)

- **GDA 假設「每一類別條件下的特徵服從多變量常態」。**這個假設在實務資料（非高斯、heavy-tail、或有明顯偏態）時會造成哪些典型偏差？有哪些檢驗或 替代假設？

Ans:

一、違反高斯假設會造成哪些典型偏差／誤差

1. 協方差估計不穩健 (covariance misestimation)

- 如果資料是 heavy-tailed (例如具有厚尾分布、outliers、極端值頻繁出現)，則樣本協方差矩陣 (sample covariance) 對離群點非常敏感。這會導致對類別內協方差 (within-class covariance) 的估計不準，進而使 GDA (或 QDA) 決策邊界偏差。
- 如果協方差估計錯誤，Bayes 判別規則 (posterior) 也可能被扭曲，造成誤分類率提升。

2. 均值估計偏差 (mean bias)

- 若類別內分布偏態 (skewed)，樣本平均 (sample mean) 並不是類別中心 (robust center) 的好代表。平均值被極端值拉扯，可能使分類邊界位移。
- 這種偏差在類別分佈極不對稱時尤為明顯。

3. 模型錯誤 (model misspecification) 對判別性能的影響

- GDA 對生成模型 (generative) 做了強假設。如果這假設錯誤 (分布不是高斯)，其 posterior $P(y | x)$ 模型可能非常不準。相比之下，判別式模型 (如 logistic regression) 對分布假設較弱，常更穩健。這是實務中常見現象。實作者常回報，在類別資料明顯非高斯時，LDA/QDA 效果不如預期。 yunlongs.cn+1
- 當類內變異異常 (heteroscedasticity) 違反共變異數假設 (比如 LDA 假設各類 Σ 相同)，也會導致 misclassification。

4. 抗噪性差 / 通用性差

- 當資料含有雜訊 (noise)、污染 (contamination) 或離群點 (outliers)，GDA 的生成式估計 (means, covariances) 容易被破壞。這樣在新數據 (test) 上的泛化能力會下降。

5. 維度災難 (high dimensionality)

- 雖然這不完全是「非高斯」問題，但在高維時 (d 很大) 協方差估計本來就困難。如果再加上非正態 (heavy-tail) 的特性，估計

誤差更大。這可能使 GDA 判別邊界非常不穩定。

二、可以採用哪些檢驗或替代假設／方法 (robust 或非高斯版本)

為了解 GDA 假設違反的影響，或替代 GDA 在非高斯情況下使用，有許多研究與方法。以下是幾種常見方向與對應研究：

1. 橢圓對稱(Elliptically Symmetric, ES) 分布假設

- 很多 robust 判別分析 (discriminant analysis) 的擴展假設資料來自於 ES 分布 (而不是嚴格 Gaussian)。ES 分布是一類比高斯更廣泛的分布，包括厚尾 (heavy-tail) 分布 (例如 t 分布、Cauchy 分布等)。
- **Robust Generalised Quadratic Discriminant Analysis (GQDA):** Ghosh 等人在 “*Robust Generalised Quadratic Discriminant Analysis*” (arXiv, 2020) 提出，他們將 QDA 擴展到 ES 分布，並使用 robust 平均與 dispersion (散度) 矩陣估計器以處理 outlier / heavy-tail。 [arXiv+2ResearchGate+2](#)
- **FEMDA (Flexible & robust discriminant):** Pierre Houdouin 等人在 arXiv 的 “*ROBUST CLASSIFICATION WITH FLEXIBLE DISCRIMINANT ANALYSIS IN HETEROGENEOUS DATA*” 中提出，每個樣本點都可以有自己的尺度參數與 ES 分佈。這個模型對樣本之間的「尺度異質性 (scale heterogeneity)」非常有彈性。 [arXiv](#)

2. Robust 協方差 / 均值估計

- **Tyler's M-estimator:** 在 *Robust Linear Discriminant Analysis Using Tyler's Estimator* (Auguin & McKay, 2019) 中提出。Tyler's 協方差估計器 (Tyler's scatter matrix) 是對 heavy-tailed / 離群點非常穩健的估計方法。它不需要假設資料是 Gaussian，但對 ES 分布非常適合。 [pureadmin.qub.ac.uk](#)
- **Cellwise Robust Regularized Discriminant Analysis:** Stéphanie Aerts, Ines Wilms (2016) 提出的方法。當資料中存在 cellwise 異

常值 (即個別特徵值出錯) 時，傳統 **sample covariance** 估計不合理。他們把 **robust covariance** 插入正則化的 **LDA / QDA** 中。
[arXiv](#)

- **Trimmed Likelihood Discriminant Analysis:** 用裁剪 (trimming) 的似然 (likelihood) 方法，以抵抗資料污染 (contamination) / 非正態性。舉例在 *Robust Trimmed Likelihood Discriminant Analysis* 的文獻中。[jmasm.com](#)

3. 分佈自由 (distribution-free) / 深度 (data depth) 方法

- **Data depth based discriminant analysis:** Ghosh (2005) 在 “*On data depth and distribution-free discriminant analysis using separating surfaces*” 一文中探討，用資料深度 (data depth) 構造判別規則，而無需假設 **multivariate normality**。這使方法對 **heavy-tail**、非正態分佈更魯棒。[projecteuclid.org](#)

4. 高維 / 稀疏 / 空間符號 (spatial sign) 方法

- **SSQDA** (Spatial-Sign QDA): 最新的一篇是 Anqing Shen, Long Feng (2025) “*A Spatial-Sign based Direct Approach for High Dimensional Sparse Quadratic Discriminant Analysis*”。他們假設資料來自 **ES** 分布 (尤其是 **heavy-tailed**)，並使用空間中位數 (spatial median) 和空間符號協方差 (spatial sign covariance) 作為 **robust** 估計器。[arXiv](#)

5. 理論「非正態性 (non-normality) 對 **LDA/QDA** 的魯棒性研究

- 一篇碩士論文 “*Robustness Against Non-Normality*” 探討 **LDA** 和 **QDA** 在資料不滿足 **multivariate** 正態假設時的分類性能 (simulation study)。[diva-portal.org](#)
- **Linnea Lantz** (2019) 的論文 (碩士) 比較了 **LDA**、**QDA**、**GQDA** 等方法在不同維度與污染 (contamination) 下的魯棒性。[diva-portal.org](#)

三、是否有 **open problem**

其實，關於 GDA / QDA 在非高斯資料 (heavy-tail, skew, contamination) 下的問題 並不是完全 **open problem**：已有很多研究（如上面列舉）提出了解法 (robust discriminant 分析、ES 分布、Tyler estimator、trimmed likelihood、spatial-sign 等)。但問題仍有**挑戰性**：

- 如何在高維 (d 很大) 的情況下同時做到 robust、可計算 (computationally efficient)、且泛化好？
 - 雖然 ES 分布是一個很通用的假設，但資料可能不是橢圓對稱 (例如高度偏態、多峰)，這時這些 robust ES 方法仍可能失效。
 - 在實務中選擇哪種 robust 估計 (Tyler, M-estimator, trimmed, sign) 最合適？如何根據資料特性 (污染程度、維度、樣本數) 系統性地選擇？
- **GDA 常見的「共變異數相同 (shared Σ)」假設若不成立，會如何改變決策邊界 (線性 \rightarrow 二次)？實務上要如何判斷採用 LDA (線性) 還是 QDA (二次)？**

Ans:

1) 數學推導：為什麼 Σ 共享 \rightarrow 線性決策邊界，否則為二次？

假設二分類（可延展到多類）且類別先驗分別為 π_0, π_1 。類條件為多元常態：

$$p(x | y = k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k)\right), k = 0, 1.$$

對於最優 Bayes 分類，我們比較後驗（或比較 log posterior），即考慮量

$$\delta_k(x) = \ln p(x | y = k) + \ln \pi_k.$$

展開（去掉共同常數）得：

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \ln |\Sigma_k| + \ln \pi_k.$$

分類決策在於取使 $\delta_k(x)$ 最大的 k 。邊界由 $\delta_0(x) = \delta_1(x)$ 決定，展開差值：

$$\begin{aligned}\delta_1(x) - \delta_0(x) = & -\frac{1}{2}x^\top(\Sigma_1^{-1} - \Sigma_0^{-1})x + x^\top(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) - \frac{1}{2}(\mu_1^\top\Sigma_1^{-1}\mu_1 \\ & - \mu_0^\top\Sigma_0^{-1}\mu_0) - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} + \ln\frac{\pi_1}{\pi_0}.\end{aligned}$$

注意第一項包含 $x^\top(\cdot)x$ 的二次型；若 $\Sigma_1 \neq \Sigma_0$ 其係數不為零 \rightarrow 邊界含二次項 \rightarrow 二次曲面 (QDA)。

若 共享協方差： $\Sigma_1 = \Sigma_0 = \Sigma$ 。則 $\Sigma_1^{-1} - \Sigma_0^{-1} = 0$ ，二次項消掉，差值變成線性形式：

$$\delta_1(x) - \delta_0(x) = x^\top\Sigma^{-1}(\mu_1 - \mu_0) + (\text{常數}),$$

即決策邊界為 $\{x: w^\top x + b = 0\}$ (線性)，其中 $w = \Sigma^{-1}(\mu_1 - \mu_0)$ 。

(此推導與教科書推導一致，詳見教材/教科書章節。)。 [Classpages+1](#)

2) 實務上如何判斷採 LDA 還是 QDA? (檢驗、規則、替代法)

下面把常用做法、各自優缺點與文獻來源整理成清單與具體建議：

A — 直接檢驗：檢查各類協方差是否相等

- **Box's M 檢定 (Box's M test)**：常用的多變量檢定來檢驗多個類的協方差矩陣是否相等 (homogeneity of covariances)。若檢定拒絕同質性，表示 Σ 不同 \rightarrow QDA 可能更合適。注意 Box's M 對常態性敏感，且在樣本量大或過小時可能不穩定。 [維基百科+1](#)

實務提醒：如果資料嚴重偏離常態，Box's M 可能會錯誤拒絕或過度保守；因此通常不要只靠單一檢定作決策。

B — 交叉驗證 (cross-validation)：實務上最可靠的做法

- 用 **交叉驗證 (k-fold CV)** 比較 LDA 與 QDA 的預測誤差。當 QDA 的 CV 誤差比 LDA 明顯低，且差異穩定 (不只是隨機波動)，可選 QDA；反之選 LDA。教科書與實務建議都把 CV 視為首選判定方法。 [卡尔林网站+1](#)

C — 樣本大小 / 參數數量考量 (偏差-變異 tradeoff)

- QDA 要估計每類一個協方差矩陣 \rightarrow 參數數量較多 (在 d 維下每個 Σ 有 $\sim d(d+1)/2$ 個自由參數)。如果每類的樣本數 n_k 太小, Σ 的估計會很不穩定, 導致 QDA 的方差很大 (過擬合)。
- 常見經驗法則與建議: 當 n 比 p (特徵數) 小很多時, 偏好 LDA (或正則化方法)。有些簡單的 rule-of-thumb 建議 $n \gtrsim 5p$ 才較安全用 LDA/QDA, 但這只是 heuristic (參考資料會提到類似建議)。UC R+1

D — 正則化 / 折衷方法 (bridge between LDA & QDA)

- **Regularized Discriminant Analysis (RDA)**: Friedman (1989) 提出把 pooled Σ 與 class-specific Σ_k 線性/混合地平滑, 與對角收縮, 形成一個連續家族, 能在 LDA 與 QDA 之間折衷, 並透過 CV 選參數。RDA 在實務上常用來避免 LDA 的偏差或 QDA 的高變異。SCI2S+1
- 其他現代變體: 高維 RDA、HDRDA、shrinkage covariance estimators (Ledoit-Wolf 等) 在 p 大時廣泛應用。這些方法能在小樣本或高維情形下穩健估計協方差, 讓 QDA / RDA 更可行。arXiv+1

E — 檢視資料 (視覺化) 與診斷工具

- 透過 **投影** (PCA、LDA 投影) 或繪製類內散布 (scatter plots、contour of class covariances) 觀察類內形狀是否相似 (若形狀明顯不同、方向與尺度差異大 \rightarrow QDA)。也可使用可視化工具檢查 Box's M 不適用情況。arXiv

F — 若資料非常不符合常態 (heavy-tail、偏態)

- 若非正態、或出現離群點, LDA/QDA 的 MLE 協方差估計會受影響。可採 **robust covariance estimators** (Tyler's M-estimator、M-estimators、spatial-sign) 或使用 **distribution-free** 類別器 (如 tree-based、SVM、或判別式模型 logistic/regularized) 作比較。若堅持用 GA-DA 類方法, 建議先用 robust 方法估協方差, 再做判別。相關研究與方法可見文獻 (下方列出)。arXiv+1

3) 總結性實務建議 (step-by-step)

1. 先視覺化、檢查類內差異（PCA/LDA 投影、散布圖）。
 2. 進行 **Box's M**（若你相信近似常態）；若 **Box's M** 顯著 \rightarrow 表示 Σ 有差異（偏向 QDA），但同時檢查常態性。[維基百科](#)
 3. 用交叉驗證 比較 LDA、QDA（與 RDA）之預測誤差；以 CV 表現為最直接的判準。[卡尔林网站](#)
 4. 若資料維度高或樣本小，優先考慮正則化/收縮方法（RDA、Ledoit-Wolf shrinkage、HDRDA 等），而不是直接 QDA。[SCI2S+1](#)
 5. 若資料 heavy-tailed 或有離群點，先用 **robust** 協方差估計 或考慮非參數/判別式替代模型（例如 logistic、SVM、tree）比較。[Semantic Scholar](#)
-

4) 研究/文獻（已查證、附連結）

下面列出與上面論點最緊要、且實務或理論上常被引用的資源（我已查證這些為真實出版物或可取得的教材/論文）：

教科書 / 綜述（理論與推導）

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* — Chapter on Discriminant Analysis（LDA/QDA 推導與比較）。可線上下載（pdf）。[國立臺灣大學統計系](#)
<https://www.stat.ntu.edu.tw/download/%E6%95%99%E5%AD%B8%E6%96%87%E4%BB%B6/bigdata/The%20Elements%20of%20Statistical%20Learning.pdf>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. *An Introduction to Statistical Learning* (ISLR) — LDA/QDA 的直觀說明與實務指南（含 CV）。[卡尔林网站](#)
<https://www.statlearning.com/>
- Bishop, C. M. *Pattern Recognition and Machine Learning* — Gaussian discriminant analysis 與分類模型推導（教材章節）。[Classpages](#)（可見教材章節或教學頁面）
<https://www.di.fc.ul.pt/~jpn/r/PRML/chapter4.html>

關鍵論文 / 方法 (bridge & regularization)

- Friedman, J. H. (1989). *Regularized Discriminant Analysis*. Journal of the American Statistical Association, 84(405):165–175. (提出 RDA，介於 LDA 與 QDA 之間的折衷與正則化)。PDF 可取得 (例如大學/研究中心 鏡像)。 [SCI2S+1](#)
<https://sci2s.ugr.es/keel/pdf/algorithm/articulo/friedman1989.pdf>
- Ramey, J. A., Stein, C. K., Young, P. D., & Young, D. M. (2016). *High-Dimensional Regularized Discriminant Analysis* (HDRDA). arXiv. (高維情況的改良)。 [arXiv](#)
<https://arxiv.org/abs/1602.01182>

檢定 / 實務測試

- Box's M test — 說明與參考 (百科與軟體實作說明)。注意其對常態性敏感性。 [維基百科+1](#)
https://en.wikipedia.org/wiki/Box%27s_M_test
(R 語言 heplots / boxM 文件)
<https://friendly.github.io/heplots/reference/boxM.html>

教學與 tutorial

- Ghojogh, B., & Crowley, M. (2019). *Linear and Quadratic Discriminant Analysis: Tutorial* (arXiv). (清楚展示 LDA/QDA 推導與例子)。 [arXiv](#)
<https://arxiv.org/pdf/1906.02590>

其他實務參考 / 規則

- UC-R “Linear & Quadratic Discriminant Analysis” practical notes (rule-of-thumb 與使用情境)。 [UC R](#)
https://uc-r.github.io/discriminant_analysis
- **GDA** 的理論保證與泛化誤差：在哪種資料/參數條件下 (樣本數 n 、維度 d 、類別數 k) **GDA** 的 **MLE** 才是有意義且穩定的？有沒有明確的樣本複雜度 界限？

Ans:

一、結論總覽 (Quick answers)

1. 古典漸近 (固定維度 d):

- 若 $n \rightarrow \infty$ 而 d 固定, GDA 的 MLE (樣本均值、樣本協方差) 一致且漸近常態, plug-in 分類器會收斂到 Bayes 決策, 誤差趨近其極限 (standard asymptotic theory)。(經典多變量統計結果)

[ENS de Lyon](#)

2. 可逆性/估計有意義的最基本條件:

- 為了直接用樣本協方差 $\hat{\Sigma}$ 做 inverse (例如 LDA 的 $w = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)$), 最少需要每類樣本數 n_k 大於維度 d (否則 $\hat{\Sigma}$ 奇異)。實務上若 n 與 d 相近, $\hat{\Sigma}$ 會非常病態 (ill-conditioned), 使分類器數值與統計表現不穩定。[ENS de Lyon](#)

3. 高維速率 / 樣本複雜度 (定量) 的大致判準:

- 在次高維/高維情形, 樣本協方差的 operator-norm 誤差通常尺度為 $\|\hat{\Sigma} - \Sigma\|_2 = O_p\left(\sqrt{\frac{d}{n}}\right)$ (在 sub-Gaussian 或有界二階矩假設下, 並可用矩陣濃縮不等式證明)。因此若希望在譜範數 (operator norm) 上一致, 需要 $d/n \rightarrow 0$ (也就是 $n \gg d$)。若只需元素級或弱範數的一致性, 條件可能較鬆, 但總體結論是: 若 d 與 n 同階或 $d > n$, 樣本協方差的誤差不會收斂到 0, 直接使用 MLE 會不穩定。這一類矩陣濃縮與速率可由矩陣 Bernstein / Vershynin 的高維概率工具導出。[arXiv+1](#)

4. 結論對 LDA / QDA 的實務意義:

- LDA 使用 pooled (共享) 協方差, 需估一次 $\Sigma \rightarrow$ 需要總樣本量 n 相對於 d 足夠大。
- QDA 每類估一個 $\Sigma_k \rightarrow$ 每個類別需各自滿足 $n_k \gg d$ 才可靠, 否則 QDA 的方差 (估計不穩) 會比 LDA 高得多。
- 當 d 不是很小而 n 不足時, 要麼用正則化/收縮 (Ledoit-Wolf,

thresholding, RDA) 或稀疏方法 (sparse LDA / direct estimators) 來穩定估計，要麼放棄 GDA 類插件方法改用判別式/非參數方法。 [ENS de Lyon+2stat.berkeley.edu+2](http://ENS.de.Lyon+2stat.berkeley.edu+2)

5. 可得的更精確條件 / 範例 (來自文獻):

- 若資料為 sub-Gaussian, operator norm 誤差量級給出 $n \gtrsim Cd$ 才能保證 $\|\hat{\Sigma} - \Sigma\|_2 = o(1)$ 。
- 若 Σ 具有稀疏結構或可 threshold 化則可在 $(\log d)/n \rightarrow 0$ 類條件下一致 (Bickel & Levina 的 thresholding 結果)。若想在 $p \gg n$ 下仍有好分類效果, 可用稀疏 LDA 類方法 (Cai & Liu 等) 得到在稀疏性條件下的一致性與速率。 stat.berkeley.edu+1

二、理論依據 (每一重要結論對應的研究與可查網址)

我把最關鍵的、可用以支撐上面結論的文獻列出 (每條你都能打開閱讀)。我在每處結論已標注相應的引用 ID; 下列是完整引用與連結。

1. **Ledoit, O. & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices.**

- 這篇提出 shrinkage (收縮) 協方差估計, 強調 sample covariance 在 high-dim 下往往 ill-conditioned, 需收縮以穩定逆矩陣運算。適合 d 與 n 同階的情況。
- PDF: https://perso.ens-lyon.fr/patrick.flandrin/LedoitWolf_JMA2004.pdf. [ENS de Lyon](http://ENS.de.Lyon)

2. **Bickel, P. J. & Levina, E. (2008). Covariance regularization by thresholding. Annals of Statistics.**

- 證明若真實協方差矩陣在某種稀疏度下可 threshold, 則 thresholded estimator 在 operator norm 上一致, 條件為 $(\log p)/n \rightarrow 0$ 。這給出了當 p 很大但矩陣結構稀疏時的樣本複雜度條件。
- DOI/PDF: <https://www.stat.berkeley.edu/~bickel/BL2008-thresholding.pdf>. stat.berkeley.edu

3. Tropp, J. A. (2015). *An introduction to matrix concentration inequalities (survey)*.

- 提供矩陣 Bernstein 等工具，常用於推導 $\|\hat{\Sigma} - \Sigma\|_2$ 的集中界（導出 $\sqrt{d/n}$ 型的速率在適當假設下成立）。實作與理論證據都引用這些濃縮不等式。
- PDF: <https://arxiv.org/pdf/1501.01571.pdf>. [arXiv](#)

4. Vershynin, R. — *High-Dimensional Probability (book / lecture notes)*.

- 系統整理高維概率工具、樣本協方差譜收斂的結果；在 sub-Gaussian 情形下可得 $\|\hat{\Sigma} - \Sigma\|_2 = O_p(\sqrt{d/n})$ 。參考書或 lecture notes。
- Notes/PDF: <https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-2.pdf>. [math.uci.edu](https://www.math.uci.edu)

5. Cai, T. and Liu, W. (2011). *A direct estimation approach to sparse linear discriminant analysis*. JASA / arXiv.

- 提出在高維下直接估計 $\Omega(\mu_1 - \mu_0)$ 的方法 (LPD)，在稀疏條件下允許 $p \gg n$ 時仍能得到一致和良好分類效果，說明在稀疏結構下樣本複雜度可大幅改善。
- arXiv / PDF: <https://arxiv.org/abs/1107.3442> (或 <https://stat.wharton.upenn.edu/~tcai/paper/Sparse-LDA-LPD.pdf>). [arXiv+1](#)

6. 綜述 (近年 review)

- 例如 2024/2025 年的高維判別分析綜述整理了 LDA/QDA 在現代高維背景下的挑戰與方法 (regularization, sparse LDA, robust variants)。(參考：Qu et al., 2024 review) [MDPI](#)

三、常見精確界與直觀速率 (數學化描述)

- **Operator-norm 條件 (sub-Gaussian)**：在常見次高維條件 (向量為 sub-Gaussian)，有常數 C 使得

$$\Pr \left[\|\hat{\Sigma} - \Sigma\|_2 \geq C(\sqrt{d/n} + t) \right] \leq 2e^{-cnt^2},$$

因而為了讓 $\|\hat{\Sigma} - \Sigma\|_2 \rightarrow 0$ 需 $d/n \rightarrow 0$ 。來源為矩陣 Bernstein / Vershynin 的不等式 (Tropp, Vershynin)。arXiv+1

- **稀疏情形**：若 Σ 或 precision matrix 有稀疏結構（或 $\Omega(\mu_1 - \mu_0)$ 稀疏），就可以使用 **thresholding** 或直接稀疏估計，在更嚴格的 $(\log d)/n \rightarrow 0$ 或更寬鬆稀疏條件下得到一致性 (Bickel&Levina, Cai&Liu)。stat.berkeley.edu+1
- **QDA 更嚴格**：QDA 每類估計一個 Σ_k ，要保證所有 $\hat{\Sigma}_k$ 都一致，通常需要每類樣本數 n_k 滿足類似的 $n_k \gg d$ 或在稀疏/結構假設下 $(\log d)/n_k \rightarrow 0$ 等條件；否則 QDA 的高變異會使其泛化不佳。關於 QDA 的高維修正 (regularized QDA) 也有相關研究 (見綜述)。MDPI

四、哪些問題仍是 open / active research ?

- 在最少假設（非高斯、非稀疏、無先驗結構）下的最小樣本複雜度 (**tight minimax** 下界)：某些情形已有下界/上界，但在完全一般性 setting（例如非橢圓、heavy-tailed、混合模態）下，關於 plug-in GDA 的最小樣本複雜度仍有許多 open 問題。研究者通常在限制模型族（sub-Gaussian、ES 分布、稀疏結構）下給出明確結果。綜述文獻會指出不同假設下的區別。MDPI
- 對 **heavy-tailed** 或非橢圓分布的精確最小樣本界：這仍是活躍議題；已經有 robust estimator 的研究 (Tyler 等)，但完整的 minimax 標準化理論在所有情形尚未完全定型。
- 多標籤 (multi-label) 問題中 "one-hot" (實際是 multi-hot) 與多類別 (mutually exclusive) 的一致性、損失設計與校準差異是什麼？

Ans:

1. 多標籤 (multi-hot) vs 多類別 (multi-class) — 基本差異與一致性

差異 (consistency)

- 在多類別 (multi-class) 問題中，每樣本僅屬於一個類別 (one-hot)，類別互斥 (mutually exclusive)。
- 在多標籤 (multi-label) 問題中，每樣本可能屬於多個類別 (multi-hot)；這表示不同類別之間不互斥。這種結構差異意味著：你不能用 **softmax + categorical cross-entropy** (那會強迫“總機率 = 1”) 來表示每個類別是否存在，而是更常用 **sigmoid + 二元交叉熵 (binary cross-entropy, BCE)**，對每個 label 做獨立預測。

這種差異導致「一致性 (consistency)」的問題：多類別的損失和訓練目標假設 labels 互斥；多標籤假設互相獨立 (或至少不互斥)，如果誤用了多類別 (softmax + cross entropy) 或簡單的 multi-hot 設置，有可能模型學到錯誤的概率分布或不合理的 confidence。

2. 損失設計 (Loss) 的差異與挑戰

在多標籤分類中，設計損失比純多類別更複雜，有幾個挑戰：

1. 正負不平衡 (positive-negative imbalance)

多標籤任務中，對大多數樣本來說，多數標籤是 0 (negative)；只有少數是 1 (positive)。這會造成 BCE 的梯度幾乎被負樣本主導。

- 有研究提出 **asymmetric loss**，專門對正負樣本做差別處理 (down-weight 簡單負樣本，hard-threshold easy negatives)。例如 Emanuel Ben-Baruch 等人在 *Asymmetric Loss For Multi-Label Classification* 中提出 ASL (Asymmetric Loss) 來解決這樣的失衡。
- 進一步的改進還有 **Asymmetric Polynomial Loss (APL)**，對 BCE 進行多項式擴展並調整正負貢獻。

2. 遺失標籤 (missing labels)

在許多實際多標籤資料集中，註釋可能不完整 (標籤遺漏)，即真是 positive 卻被標記成 0。Zhang 等人在 *Simple and Robust Loss Design*

for Multi-Label Learning 中指出，好的 loss 應該對遺漏 (false negatives) 有一定魯棒性。他們設計了一種簡單但穩健的損失函式來處理遺漏標籤問題。 [arXiv](#)

3. 類別共現 (label co-occurrence)

多標籤中 labels 之間可能有相關性 (共現)：某些標籤經常一起出現。傳統 BCE 假設各 label 預測獨立，但忽略這些關係可能降低效能。

- 在 *Multi-Label Classification with High-rank and High-order Label Correlations* (Si, Jia, Zhang, 等人) 中，他們提出既保留高階 (high-order) 標籤相關性，又不強制把標籤矩陣做低秩降維的方法。 [PALM](#)
- 還有方法在訓練中加入正則項或結構來建模 label correlation。

3. 校準 (Calibration) 差異

校準是指模型輸出分數（機率估計）與真實標籤機率的一致性 (confidence 校對)。在多標籤中，比單標籤 (multi-class) 更具挑戰性：

- 多標籤過度自信 (overconfidence) 問題：由於每個 label 都是二元預測 (sigmoid)，傳統 calibration 方法 (如 temperature scaling、label smoothing) 不一定直接適用。
- 有新的研究專門針對多標籤校準 (multi-label calibration)：
 - Tianshui Chen 等 (2024) 在 *Dynamic Correlation Learning and Regularization for Multi-Label Confidence Calibration* 中提出 DCLR 方法，透過學習不同類別間的語義相關性 (semantic correlation) 來進行正則化，以校準信心 (confidence)。 [arXiv](#)
 - J. Cheng 等 (CVPR 2024) 提出了 *Towards Calibrated Multi-label Deep Neural Networks*，發現在多標籤任務中加入 SPA (something) 和 LPR (label-probability regularization) 可以顯著改善校準。 [CVF 開放取用](#)

4. 是否為 open problem ?

並不是完全 open problem — 有很多活躍研究在設計更合適的 multi-label 損失 (loss) 和校準 方法 (calibration)。上述 ASL、APL、robust loss、DCLR 等都是針對 multi-hot 特性提出的方法。但也有一些挑戰仍然持續存在：

- 標籤之間共現關係非常複雜 (高階相關性) 時，如何設計既高效又準確的 loss / 模型是一個活躍領域。
- 多標籤資料的校準 (confidence calibration) 還比單標籤更難，而且尚無統一標準。不同應用 (圖像、多標籤文本、大型語言模型) 的校準需求不同。

5. 口語化重述 (方便對外解釋)

- 什麼是 “multi-hot” vs “mutually exclusive” ?

想像一張圖裡可以有「貓」也可以有「狗」；這是多標籤 (multi-hot)，你可以把貓和狗都標成 1。而互斥多類別 (multi-class) 就像從一堆動物中只選一種，比如「這張圖是貓、狗還是鳥，只能有一個答案」。

- 為什麼損失不同？

如果你讓模型預測每個標籤 (貓、狗、鳥) 是否存在 (multi-hot)，你常用 Sigmoid + 二元交叉熵 (BCE)，但 BCE 有問題：大多數標籤是 0 (沒有那種動物)，所以模型可能被“沒標籤”的信號壓倒。為了解決，有人設計有正負不對稱處理 (asymmetric loss)、還有針對缺失標籤 (missing labels) 的 robust 損失。

- 校準 (confidence) 怎麼辦？

模型可能對每個標籤給出很高信心 (probability)，但實際上是錯的。於是新的研究進來：在訓練時不只看單標籤的誤差，還學習標籤之間語義 (semantic) 的關係 (哪些標籤容易一起出現)，然後把這些關係加到校準正則項裡。這樣模型給的「這張圖有貓」信心會更可信。

6. 重要參考文獻 (整理)

主題	文獻	連結
校準 (Calibration) in multi-label	T. Chen 等, <i>Dynamic Correlation Learning and Regularization for Multi-Label Confidence Calibration (DCLR)</i>	arXiv arXiv: https://arxiv.org/abs/2407.06844
	J. Cheng 等, <i>Towards Calibrated Multi-Label Deep Neural Networks</i>	CVF 開放取用 CVPR 2024 (論文 PDF)
不平衡 & 非對稱損失	E. Ben-Baruch 等, <i>Asymmetric Loss For Multi-Label Classification</i>	arXiv: https://arxiv.org/abs/2009.14119
	Y. Huang 等, <i>Asymmetric Polynomial Loss (APL)</i>	arXiv: https://arxiv.org/abs/2304.05361
缺失標籤 / 魯棒損失	Y. Zhang 等, <i>Simple and Robust Loss Design for Multi-Label Learning</i>	arXiv arXiv: https://arxiv.org/pdf/2112.07368.pdf
標籤相關性	C. Si, Y. Jia, Y. Zhang 等, <i>Multi-Label Classification with High-rank and High-order Label Correlations</i>	PALM TKDE 論文 PDF

● GDA（生成式）與 logistic（判別式）在 model

misspecification 下哪個更 穩健？在有限樣本下哪個更泛化？有沒有明確理論或實驗界限？

Ans:

A — 結論摘要（要點先看這裡）

1. 短樣本/小 n 時，生成式（**GDA/Naive-Bayes** 等）往往更有效率（**sample-efficient**），因為它做更強的分布性假設（例如條件高斯），把「額外資訊」編進模型；在樣本很少時可得到較低的誤差。這是 Ng & Jordan (2001) 的核心觀察與大量實驗所支持。ai.stanford.edu
 2. 當樣本量很大（**asymptotic**），判別式（**logistic / discriminative**）通常有更低的極限誤差（更好的漸近性能），且在模型錯誤（**model misspecification**）時通常更魯棒（**discriminative** 不需要假設 $p(x|y)$ 的具體形式，只直接建模 $p(y|x)$ ）。因此在 n 很大時 **logistic** 往往會勝出。Ng & Jordan 與隨後理論文獻都指出了這種「兩階段」現象。
ai.stanford.edu+1
 3. 收斂速率的 **trade-off (finite-sample vs asymptotic)**：generative 模型「學得快」但如果假設錯了其極限誤差會比 **discriminative** 差；**discriminative** 模型學得慢（需要更多資料）但在大樣本下能達到更優極限誤差。這個“快學慢收斂 / 慢學好收斂”的現象有明確實驗與理論分析支持。ai.stanford.edu+1
 4. 在 **model misspecification**（生成模型的分布假設不成立）的情況下：
 - 若生成模型（例如每類為高斯）嚴重不符合真實資料，GDA 的估計會有偏差，且該偏差可能導致較差的 **asymptotic risk**；而 **logistic**（判別式）只需正確的條件概率形式或能擬合 $p(y|x)$ 的決策邊界，通常更穩健。此點在不少理論/實驗研究與綜述中被提及（見引用）。arXiv+1
 5. 並非絕對法則／存在例外：若生成假設（如 GDA 的高斯）近似成立（或你能正確指定生成族），則生成式在有限樣本上會明顯勝出；實務上常用的策略是：若 n 小或你確信模型近似成立，用生成式；否則（或你想要更魯棒），用判別式或用折衷/正則化方法（**regularized / hybrid / RDA / semi-supervised generative+discriminative** 等）。Aman+1
-

B — 關鍵已查證文獻（立即可打開閱讀）

下面列的是直接支撐上面主要結論的經典或近年工作（我已檢查每篇都是真實、可取用的資源並附上連結）：

1. **Ng, A. Y. & Jordan, M. I. (2001). *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes.***
 - 主張（實驗+分析）：generative（Naive Bayes）在小樣本時常優於 discriminative（logistic），但 logistic 在大樣本時極限誤差較小（更好）。這篇是整個討論的經典來源。
 - PDF（NIPS 2001）：<https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf>. ai.stanford.edu
2. **Zheng, C., et al. (2023). *Revisiting Discriminative vs. Generative Classifiers: Theory and Implications* (arXiv).**
 - 近期工作，回顧並拓展 Ng & Jordan 的觀察，討論在更寬廣假設下（包含 misspecification、different regimes）兩類模型的行為，並提供現代實驗。對於理解最新觀察（尤其深度模型與大數據情境）很有用。
 - arXiv: <https://arxiv.org/pdf/2302.02334.pdf>. [arXiv](https://arxiv.org)
3. **Tong Zhang (2004). *Statistical behavior and consistency of classification methods based on convex risk minimization.*** (Annals of Statistics)
 - 提供判別式（基於凸風險最小化，如 logistic）的一般一致性與收斂分析，說明在較弱假設下判別式方法的理論性質（重要於理解 discriminative 的漸近魯棒性）。
 - PDF: https://tongzhang-ml.org/papers/aos04_consistency.pdf. [通章的首頁](https://tongzhang-ml.org)
4. **Efron, B. (1975). *The efficiency of logistic regression compared to normal discriminant analysis.*** Journal of the American Statistical Association.

- 早期研究比較 logistic 與「正態鑑別」(類似 LDA) 的效率 (asymptotic relative efficiency)，提供關於效率差異的理論計算 (在某些 canonical 設定下)。有助理解不同方法在統計效率上的比較。
- PDF (mirror):
https://www.uio.no/studier/emner/matnat/math/STK9200/h21/efron_jasa1975.pdf. [挪威大學](#)

5. 綜述 / 實務比較 (現代實驗與實務建議)

- Amazon Science / repo 和多篇教科書、課程講義也整理了這個 trade-off (sample efficiency vs asymptotic accuracy) 與實務指引；可作為補充閱讀 (例如 CS229 / Hastie 等教材的 GDA vs logistic 條目)。(教材與現代回顧幫助實務選擇。)
- 例如教材說明 (線上筆記) : <https://aman.ai/cs229/gda/> 及 Hastie et al., *The Elements of Statistical Learning* (第 4 章相關部分)。[Aman+1](#)

C — 更細節的解釋 (為什麼會出現這種行為？數學直覺)

- 為什麼生成式學得快 (sample efficient) ?
因為生成式在學參數時對 $p(x | y)$ 做了結構化假設 (例如 GDA 假設條件分佈是高斯)。這把資料的統計資訊「壓縮」進較少的參數空間，因而在樣本不多時估計方差較小 (bias 可能較低)，導致早期樣本下表現好。Ng & Jordan 的定量實驗清楚顯示這點。[ai.stanford.edu](#)
- 為什麼判別式 asymptotically 更好 / 更魯棒於 misspecification ?
判別式直接最小化與分類相關的目標 (例如 log-loss for $p(y | x)$)，不需要對 $p(x | y)$ 做具體假設；當生成假設錯誤 (misspecification) 時，生成式可能得到系統性偏差，而判別式不受錯誤生成假設的直接約束，因此通常有較小的 asymptotic excess risk (在大樣本下)。Tong Zhang 與其他理論工作討論了此類一致性與風險界。[通章的首頁](#)
- finite-sample 與 asymptotic 的 trade-off 可被定量描述 (Ng & Jordan 給了數值示例；後續 work 提供更嚴謹界)。近期文獻 (例如

2023 的 revisiting) 在更廣泛 setting 下也驗證該結論。 [arXiv+1](#)

D — 還有哪些 open / 活躍研究問題？

- 在強烈 **misspecification** (例如多模態、**heavy-tailed**、非橢圓) 下的嚴格有限樣本界 (**minimax**): 雖然有大量結果在特定假設 (sub-Gaussian、稀疏、ES 分布) 下可以得到界, 但在完全一般 misspecified setting 下的最小樣本複雜度 (以及分類器之間的最優劃分) 仍是活躍研究領域 — 意味著: 很多情形已有答案, 但沒有一個「全能」的、對任意 misspecification 都成立的單一法則。Zheng (2023) 與其他近年工作就是在推進這類理解。 [arXiv](#)
- 深度生成 vs 深度判別 (**modern deep nets**) 之間同樣 **trade-off**? : Ng & Jordan 的結論是在比較 parametric linear-logistic vs naive Bayes。把類比推到大型深度生成模型 vs discriminative deep nets, 理論上與實作上都有新挑戰; 這也是目前的活躍領域 (部分 recent papers 在 revisiting 中討論)。 [arXiv](#)

E — 實務建議 (簡單可執行)

1. 小樣本 (n 小) 且你有理由相信生成假設近似成立 (例如檢查過特徵分布近似高斯) → 優先考慮生成式 (**GDA / Naive Bayes**) 或混合方法 (生成 + 判別微調)。 [ai.stanford.edu](#)
2. 大樣本或資料可能非高斯 / **model misspecification** 明顯 → 優先考慮判別式 (**logistic / regularized**) (因為 asymptotic 魯棒性); 同時用 cross-validation 比較實際泛化誤差。 [通章的首頁](#)
3. 若不確定: 做交叉驗證 (LDA/GDA, logistic, RDA, regularized alternatives) 比較, 或採用折衷方法 (RDA: Friedman 1989; 或 semi-generative + discriminative fine-tune)。交叉驗證在有限樣本下通常是最實用的決策依據。 [Aman](#)

F — 你要的快速參考清單 (再次列出已確認可讀連結)

- Ng, A. Y. & Jordan, M. I. (2001). *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes*. NIPS 2001 PDF: <https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf>. ai.stanford.edu
- Zheng, C. et al. (2023). *Revisiting Discriminative vs. Generative Classifiers: Theory and Implications*. arXiv: <https://arxiv.org/pdf/2302.02334.pdf>. [arXiv](https://arxiv.org)
- Zhang, T. (2004). *Statistical behavior and consistency of classification methods ...* Annals of Statistics. PDF: https://tongzhang-ml.org/papers/aos04_consistency.pdf. [通章的首頁](#)
- Efron, B. (1975). *The efficiency of logistic regression compared to normal discriminant analysis*. JASA (1975). Mirror PDF: https://www.uio.no/studier/emner/matnat/math/STK9200/h21/efron_jasa_1975.pdf. [挪威大學](#)
- 教材/入門（實務指引）：CS229/GDA 註記 <https://aman.ai/cs229/gda/> ； Hastie, Tibshirani & Friedman 教科書（Elements of Statistical Learning）◦ [Aman+1](#)

● Σ 的估計在 **high-dim** (**d large**) 或 **n < d** 時會如何失敗？

Ans:

A. 當 **d** 大或 **n < d** 時，樣本協方差會如何失敗（具體現象）

1. 奇異（**singularity**）／不可逆：

- 樣本協方差矩陣 $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$ 的秩最多為 $\min(d, n-1)$ 。因此若 $n \leq d$ （常寫成 $p > d$ 或 $d \geq n$ ）， $\hat{\Sigma}$ 通常是奇異的，無法直接求逆。這是線性代數的直接結論。 [Cross Validated](#)

2. 數值病態（**ill-conditioning**）——極大條件數：

- 即便 $n > d$ 但 n 與 d 規模接近， $\hat{\Sigma}$ 的最小／最大特徵值可能很小／

很大，導致條件數（condition number）非常大，求逆時會把估計誤差放大，數值與統計上都不穩定。Ledoit & Wolf 強調 sample covariance 在大維度下「通常不 well-conditioned」，需要收縮。[ENS de Lyon](#)

3. 樣本特徵值分佈被嚴重扭曲（Marčenko–Pastur 現象）：

- 在高維隨機矩陣理論中，樣本協方差的特徵值分佈會偏離真實協方差的譜，形成 Marčenko–Pastur 分佈與離散的前緣（spike）行為。這代表大部分特徵值會散佈，並非集中在真實譜上，導致 PCA / 特徵向量不可靠。這由隨機矩陣理論（MP 法則）給出精確描述。[加州大學爾灣分校數學系+1](#)

4. 估計誤差量級（operator norm）隨維度膨脹：

- 在常見 sub-Gaussian 假設下，樣本協方差在 operator 範數（spectral norm）上的誤差大致量級是 $O_p(\sqrt{d/n})$ 。因此要 $\|\hat{\Sigma} - \Sigma\| \rightarrow 0$ 通常需要 $d/n \rightarrow 0$ （即 $n \gg d$ ）。否則誤差不會收斂。這類速率可由矩陣濃縮不等式推導（見 Vershynin、Tropp 的教材/survey）。[Numdam+1](#)

5. 特徵向量（eigenvector）與逆矩陣估計不可靠：

- 當特徵值被噪聲淹沒或樣本噪聲大時，對應的主成分與逆矩陣方向可能與真實方向偏離，導致用 $\hat{\Sigma}^{-1}$ 的 downstream 方法（LDA、Mahalanobis 距離、投影等）效果惡化。隨機矩陣理論與諸多模擬都顯示此現象。[加州大學爾灣分校數學系](#)

B. 這些失敗對下游任務（例如 LDA/QDA／回歸／PCA）的具體影響

- **LDA/QDA**：LDA 需要 $\hat{\Sigma}^{-1}$ 與 pooled covariance；QDA 需為每類 $\hat{\Sigma}_k^{-1}$ 。若 $\hat{\Sigma}$ 奇異或病態，線性分界權重 $w = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)$ 會很不穩定，導致高方差與過擬合（尤其 QDA 因更高參數量更脆弱）。（實務上常見 QDA 在 n 小或 d 大時表現差）[ENS de Lyon](#)
- **PCA / 特徵提取**：主成分會被樣本噪聲支配（spurious components），導致降維後的表現與真實結構脫節（false discoveries / false PCs）。MP

法則描述了在隨機噪聲下特徵值如何分散，並給出何時 **spike** 可被可靠檢出。[加州大學爾灣分校數學系](#)

- **逆矩陣相關量 (Mahalanobis, whitening)**: 如果 $\hat{\Sigma}^{-1}$ 不穩定，任何基於其的距離或投影都具有高方差，會惡化分類/回歸表現。Ledoit-Wolf 提出收縮以穩定逆。[ENS de Lyon](#)
-

C. 已有修正 / 解法 (每項附代表性文獻與可點開網址)

我把每種常見修正方法列出，並給代表性、已知度高的 reference (可打開閱讀)。

1. **Shrinkage / Ledoit-Wolf shrinkage** — 對 sample covariance 做線性收縮到簡單目標 (如 $\alpha I + (1 - \alpha)\hat{\Sigma}$)，使估計可逆且 well-conditioned；在高維下可顯著改善逆的穩定性與下游性能。
 - Ledoit, O. & Wolf, M. (2004). *A well-conditioned estimator for large-dimensional covariance matrices*. Journal of Multivariate Analysis. PDF: https://perso.ens-lyon.fr/patrick.flandrin/LedoitWolf_JMA2004.pdf. [ENS de Lyon](#)
2. **Thresholding / Tapering (sparsity exploiting)** — 假如真實 Σ 稀疏或近似稀疏，對 $\hat{\Sigma}$ 做 thresholding (置零小元素) 可以在 operator norm 上得到一致性，要求通常是 $(\log d)/n \rightarrow 0$ 。Bickel & Levina 提出並分析了 thresholding 方法的性質。
 - Bickel, P. J., & Levina, E. (2008). *Covariance regularization by thresholding*. Annals of Statistics. PDF: <https://www.stat.berkeley.edu/~bickel/BL2008-thresholding.pdf>. [加州大學伯克利分校統計系+1](#)
3. **Graphical Lasso / sparse precision estimation** — 若 precision matrix $\Omega = \Sigma^{-1}$ 稀疏，可用 graph-lasso (L1 penalty on precision) 估計稀疏逆矩陣，適合高維下結構化估計。
 - Friedman, Hastie & Tibshirani (2008). *Sparse inverse covariance estimation with the graphical lasso*. Biostatistics. PDF:

<https://jerryfriedman.su.domains/ftp/glasso-bio.pdf> (或官方期刊頁). [Jerry Friedman+1](#)

4. **Minimax / optimal rates 理論與 tapering estimator** — Cai, Zhang & Zhou 等給出高維下 covariance estimation 的最小速率與達到速率的估計量 (tapering, thresholding)。這說明在不同結構假設下需要的樣本複雜度差別。

- Cai, T. T., Zhang, C.-H., & Zhou, H. H. (2010/2012). *Optimal rates of convergence for covariance matrix estimation*. arXiv / Ann. Statist. PDF: <https://arxiv.org/abs/1010.3866> ；期刊版本可見 <https://repository.upenn.edu/bitstreams/c92f8fe1-1a65-48e1-b17a-dd2aeabc318d/download>. [arXiv+1](#)

5. **Robust / Elliptical estimators (Tyler's M-estimator 等)** — 當資料 heavy-tailed 或有離群點，Tyler 的 M-estimator 與其正則化版本對「散度 (scatter)」的估計更魯棒，比 sample covariance 在這些情況下更穩定。近年也有把 Tyler 與稀疏 / threshold 方法結合的工作。

- 參考 (survey / 實作) : *Robust sparse covariance estimation by thresholding Tyler's M-estimator*, 相關文獻與 review (見 arXiv / ProjectEuclid)。例如 arXiv: <https://arxiv.org/pdf/1009.5331.pdf> (Chen et al., regularized Tyler) 及 ProjectEuclid 頁。 [Project Euclid+1](#)

6. **Regularized Discriminant Analysis (RDA) 與其他折衷法** — 在 LDA/QDA 的情況下，若 Σ 估計不穩，可用 RDA (Friedman 1989) 或 Ledoit-Wolf 之類的收縮替代，或用降維 (PCA、reduced-rank LDA) 減少參數。

- Friedman, J. H. (1989). *Regularized Discriminant Analysis*. Journal of the American Statistical Association. (可在 JSTOR / 作者頁找到) leddoit.net

7. **Other modern approaches** : tapering estimators, spectral corrections (eigenvalue shrinkage / MP inversion)、cross-validation for tuning、bootstrapping for threshold selection 等都是常見實務工具。近期也有用

Marchenko–Pastur 理論做 eigenvalue correction 的方法（見當代隨機矩陣應用文獻）。[arXiv+1](#)

● 為什麼 **score function** 比直接學密度函數更容易學習？

Ans:

要點 1 — 消除 **normalization constant**（避免 **partition function**）

直觀：很多好用的能量模型只會給出 unnormalized form $p(x) \propto \exp(q(x))$ ，但是常數 $Z = \int \exp(q(x))dx$ 很難算。score $\nabla_x \log p(x)$ 裡面 Z 會消掉（ $\nabla_x \log p(x) = \nabla_x q(x)$ ），因此不需估計或近似 Z 。

為什麼重要：最大似然或直接估密度通常要處理 Z （或用 MCMC）——計算量或收斂問題都很嚴重。Score-matching 直接避開這道難題。

代表性參考（理論基礎）：Hyvärinen, *Estimation of non-normalized statistical models by score matching* (JMLR, 2005).

- PDF／官方頁面（可點）：

<https://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf>
[jmlr.org+1](#)

要點 2 — 局部性（**score** 是局部資訊，比全域 **density simpler**）

直觀：密度 $p(x)$ 是一個全域函數（需要決定整個空間上的相對高度）；score 是局部向量場（在每個點告訴你哪個方向 **density** 變大），學習局部方向往往比要預測整個標度更簡單。尤其在高維或資料位於低維流形時，直接估 $p(x)$ 的 global normalisation/shape 很難，而學梯度場可以聚焦在「方向」信息。

研究 / 證據：這一觀點在 score-based literature 與後續理論工作中多有討論，且是 score-matching 與 denoising-based 方法被採用的核心直覺（見 Song & Ermon 2019/2021 的引言與動機說明）。

- Song & Ermon (2019), *Generative Modeling by Estimating Gradients of the Data Distribution* (arXiv) 說明了學 score 在高維/流形情況下的優勢與多尺度處理策略。<https://arxiv.org/abs/1907.05600> [arXiv](#)
-

要點 3 — 與去噪（denoising）有直接的可解目標（實作方便且數值穩定）

直觀：把原始資料加上已知型態的雜訊（例如 Gaussian），那麼 conditional 的 score $\nabla_x \log p(x | x_0)$ 有解析式（對 Gaussian： $-\frac{x-x_0}{\sigma^2}$ ），可以把學 score 變成一個普通回歸問題（把 noisy x 對應到解析標籤）。這種做法（denoising score matching, DSM）避開了要計算二階導數或 divergence 的困難，數值上非常穩定且容易在深網路上實作。

代表性參考：Vincent (2010/2011) 證明了 denoising autoencoder 的去噪目標等價於一種 score-matching；Song 等（2019/2021）把它用在多尺度/時間依賴的情況，得到強大的生成模型。

- Vincent tech report / PDF (DAE \leftrightarrow score connection) :
https://www.iro.umontreal.ca/~vincentp/Publications/smdae_techreport.pdf [iro.umontreal.ca](https://www.iro.umontreal.ca)
 - Song & Ermon (2019) 與 Song et al. (2021 SDE paper) 說明 DSM 在生成與 sampling 的實務價值與方法：<https://arxiv.org/abs/1907.05600>，<https://arxiv.org/abs/2011.13456>。arXiv+1
-

要點 4 — 在生成（sampling）上可以直接用 Langevin / reverse SDE

直觀：一旦你學到 $\nabla_x \log p_t(x)$ （可能是對不同 noise level 的 time-dependent score），就可以用 Langevin dynamics 或反向 SDE 把簡單噪聲「倒回」成資料分布。也就是說，score 不只是好學，還直接成為生成器的控制訊號。這使得學 score 在實務上特別有用（與 GAN、VAE 不同的工作流程）。

代表性參考：Song et al. *Score-based Generative Modeling through SDEs*，以及 Ho et al. *DDPM* 都把這個想法做成強力的實作。

- Song et al. SDE paper: <https://arxiv.org/abs/2011.13456>。
 - Ho et al. DDPM:
<https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf>。arXiv+1
-

要點 5 — 計算負擔與數值穩定性

直觀：直接最小化 density-based objective（如直接最大化 likelihood）在很多能量模型下需要近似（MCMC、partition function 估計、importance sampling 等），這些東西在高維時往往很難收斂與調參。相較之下，把訓練問題轉成回歸（DSM）或只需要一階導數（score）的式子，訓練更穩定、計算更簡單。

Hyvärinen 的原始工作與 Vincent、Song、Ho 等後續實作都展示了這點。參見上面連結。jmlr.org+2iro.umontreal.ca+2

還有沒有 open questions？（哪些面向仍在研究）

短答：**有**。雖然上面幾個理由都有 strong 理論或實驗支撐，但一些細節仍屬活躍研究，例如：

- 在極高維或資料極端稀疏時 score 的估計誤差如何影響最終生成的置信度 / likelihood（理論保證仍在推進）。
- 最佳的 noise schedule / weighting 策略（如何在多個 σ 之間加權）有越來越多分析，且仍有改進空間。
- 非 Gaussian 或結構化 noise 的 DSM（是否能更有效地 capture 結構）是近年研究方向之一。

這些都是有文獻正在討論但尚未完美解決的問題（非完全 open-problem，但仍是 active research）。可參考近年針對 weighting、理論收斂與非 Gaussian noise 的論文（例如 2022–2025 年的 follow-up work；若你要我可以把這些 newer papers 一併列出）。

● 在 DSM 中，雜訊程度 σ 如何影響模型的泛化能力？

Ans:

1. 目標分布（target score）的變異性

- 當你用噪音 σ 加在 clean 樣本上時，你實際學的是 noisy 分布 $p_\sigma(x)$ 的 score，而這個 score 跟 clean 分布 $p_0(x)$ 的 score 不完全一樣。也就是說，你的訓練目標是帶噪下的一個「模糊版

gradient」。

- 對於某些 σ 值，這個模糊版本可能更平滑、更容易學；但同時它也可能偏離真實乾淨資料的 **score**。這種偏差 (**bias**) + 變異性 (**variance**) 之間的權衡會影響學出來的模型在真實資料上的泛化。

2. 誘導偏差 (Inductive bias)

- 很新的工作指出： σ 的變異結構本身會在訓練中引入一種 **inductive bias** (模型偏好)：「泛化 through variance」現象。換句話說，不只是模型容量或架構在決定泛化能力，訓練時噪音的協方差結構 (**noise covariance**) 也很關鍵。
- 具體來說，有研究發現，當訓練目標是帶噪 **score** 時，模型學到的分布有可能比原始訓練資料分布「補上缺失區」(**training data** 中低密度或稀疏部分)，這部分是由於 **noise target** 的統計特性。

3. 噪音 **schedule** (多尺度噪音)

- 如果只用一個固定 σ ，模型可能只能學到那一層次 (尺度) 的特性；但如果訓練時用多個 σ (或隨時間改變 σ)，模型可以從粗到細學習：大 σ 捕捉全局結構，小 σ 學細節。這樣有助於泛化。
- 同時，不同的噪音 **schedule** (如何選 σ 階段、每個階段損失加權) 對最終生成品質和泛化效果有重大影響。

4. 訓練與樣本複雜度

- 噪音太大：樣本變得非常模糊，模型可能容易學「很平滑但不精確」的方向。
- 噪音太小：學習目標非常接近原始 **score**，但訓練樣本變化較小，可能導致 **overfitting**、欠覆蓋低密度區。
- 最好的泛化往往是在中間策略或多尺度噪音設計中取得。

相關研究 (有文獻支撐)

以下是一些和這個問題密切相關的**最近研究**，它們從理論或實驗角度探討 σ

(或噪音調度) 對泛化的影響：

1. **Vastola, John J. (2025).** *Generalization through variance: how noise shapes inductive biases in diffusion models.*

- 這篇論文直接探討了 DSM 目標中帶噪分數 (noisy target) 的協方差結構如何影響模型的泛化。作者建立理論模型 (使用 path integral 方法) 來分析在不同參數化和噪音條件下，模型學到的是什麼樣分布，並指出透過變異 (variance) 噪音而產生的偏差 (bias) 可能正是泛化能力的一部分。
- 論文連結 (arXiv): <https://arxiv.org/abs/2504.12532> arXiv

2. **Chen, Ting (2023).** *On the Importance of Noise Scheduling for Diffusion Models.*

- 這篇實驗性工作 (empirical study) 表示 noise scheduling (如何選擇噪音強度隨時間變化) 對最終生成模型性能非常重要。作者發現最佳噪音 schedule 取決於任務 (例如圖片尺寸)，而且隨著圖像尺寸變大，較強的噪音更有利。這暗示了 σ 的選擇與泛化 / 樣本多樣性有密切關係。
- 論文連結 (arXiv): <https://arxiv.org/abs/2301.10972> arXiv

3. **Han, Yinbin; Razaviyayn, Meisam; Xu, Renyuan (2024).** *Neural Network-Based Score Estimation in Diffusion Models: Optimization and Generalization.*

- 這篇論文從更理論的角度出發，把 DSM 看作是一種有雜訊標籤 (noisy labels) 的迴歸問題 (regression with noise)，分析神經網路通過梯度下降訓練 score function 時的泛化誤差 (generalization error)。作者給出了 sample complexity (樣本數量) 與泛化界 (generalization bounds)，針對具有不同噪音強度 (不同 σ) 的訓練情境進行分析。
- 論文連結 (arXiv): <https://arxiv.org/abs/2401.15604> arXiv

- **結論**：不是「 σ 越大越好也不是越小越好」，而是存在一個 **trade-off**：噪音越強，目標越平滑但偏差大；噪音越弱，更貼近真實但可能過擬合。最佳策略往往是使用多尺度噪音（或 **schedule**）來平衡。
- **與泛化**：噪音結構本身（其方差、分布等）會導出一種 **inductive bias**，影響模型學到的分布，這可能是 **diffusion** 模型能夠「填補訓練資料中稀疏區域」的一種原因（如 Vastola 的理論分析所示）。
- **實務建議**：設計 **noise schedule**（ σ 階段、加權策略）是訓練 DSM / **diffusion** 模型時一個非常關鍵的超參數。

● DSM、ISM、ESM 三者在實作時有何差異？何者更穩定？

Ans:

快速結論（1-句版）

- **ESM (Explicit Score Matching)**：理論直接、但通常需要二階導數或明確 p ；難實作，在大型神經網路上不常用。jmlr.org
- **ISM (Implicit Score Matching / Hyvärinen identity)**：把未知的 p 用積分分部換掉，但會產生 **divergence** (trace/Jacobian) 項，需要用隨機估計 (Hutchinson) 或近似技巧；可理論化但計算上有挑戰。jmlr.org+1
- **DSM (Denoising Score Matching)**：以加入雜訊 + 去噪為監督信號，把目標轉成解析式（對 Gaussian 有 closed-form），在深網路上最容易實作、數值也最穩定，因此是目前實務上（diffusion / score-based generation）最常用的方法。蒙特利爾大學資訊研究所+1

下面分段把每個方法的 實作要點、計算成本、穩定性與代表性研究 一一整理。

1) ESM (Explicit Score Matching) —— 定義與實作要點

核心想法：直接最小化

$$\mathbb{E}_{x \sim p} \| S_{\theta}(x) - \nabla_x \log p(x) \|^2。$$

實作問題：

- 需要知道或估計真實 $\nabla_x \log p(x)$ ，若你在做 parametric energy model $p_\theta \propto \exp(q_\theta)$ ，把 loss 展開會出現 q 的二階導數 (Hessian trace) 項。

- 二階導數對深度網路很昂貴，且數值不穩，不利大尺度訓練。

何時用：分析性/理論研究、低維或可解析模型時。

關鍵參考：Hyvärinen (2005) — 原始 ESM/ISM 理論推導與公式。

jmlr.org

2) ISM (Implicit Score Matching / Hyvärinen identity) ——實作要點與挑戰

核心想法：用積分分部把 ESM 中含 p 的項轉成只含 S_θ 與其 divergence ($\nabla_x \cdot S_\theta$) 的表達式：

$$\mathbb{E}_{x \sim p} [\|S\|^2 + 2\nabla \cdot S] = 0.$$

優點：

- 不需要直接使用 $\nabla_x \log p$ 或 partition function。

主要實作難點：

- divergence 是對每個輸出維度對輸入求偏導再相加——相當於 trace(Jacobian)，直接計算在高維非常昂貴。
- 常用技巧：用 Hutchinson 隨機 trace estimator (用隨機向量 v 評估 $v^T J v$ 的期望) 來近似 divergence，或用 network 結構限制 (makes divergence cheap)。但 Hutchinson 的估計方差會影響訓練穩定性。

jmlr.org+1

何時用：如果你想直接用 score-matching objective 做 energy-based model 的訓練，或你能有效估計 divergence (例如用 Hutchinson 或 sliced-score 技術)，ISM 有理論優勢。

相關進展：Sliced Score Matching (Song et al., 2019) 提出可擴展的替代式以降低二階計算負擔。[arXiv](https://arxiv.org/abs/1906.00442)

3) DSM (Denoising Score Matching) ——實作細節與為何最常用

核心想法：把 x_0 加已知噪聲得到 x ，利用 $\nabla_x \log p(x | x_0)$ (對 Gaussian 有

解析式) 作為監督目標去回歸 score :

$$\text{loss} \approx \mathbb{E}_{x_0} \mathbb{E}_{x|x_0} \| S_\theta(x) - \nabla_x \log p(x | x_0) \|^2 .$$

為何實作上容易/穩定 :

- 對於 Gaussian corruption , $\nabla_x \log p(x | x_0) = -(x - x_0)/\sigma^2$ 有 closed-form , 讓訓練成為標準回歸問題 (不需二階導數或 divergence trace) 。
- 可以做為多尺度 (noise-conditional) learning : 把不同 σ 合併進一個 conditional 網路 , 從粗到細學習 , 數值上更穩健 。

實務經驗 :

- DSM 是現行最成功的做法 (NCSN、DDPM、Score-SDE 等都基於此訓練思想) , 在高維影像上可穩定訓練並且產生高品質樣本 。 Song & Ermon (2019) 與 Ho et al. (2020) 等展示了許多實驗支撐 。 [蒙特利爾大學資訊研究所+2arXiv+2](#)

4) 比較表 (簡短、實作導向)

面向	ESM	ISM	DSM
需要真實 $\nabla \log p$?	是 (或需 analytic p)	否 (透過 identity 轉換)	否 (用 $p(x$
需不需要二階導數 / divergence ?	通常需要二階	需要 divergence (trace of Jacobian)	不需要 (對 Gaussian)
計算成本 (高維)	很高	高 (可用 Hutchinson 估計)	中等 (標準回歸, 依網路大小)
數值穩定性 (大型 NN)	差	中 (取決於 trace estimator 變異)	好 (最穩)
在影像 / 高維上常用 ?	否	少 (除非特別技巧)	是 (主流)
代表論文	Hyvärinen	Hyvärinen (2005),	Vincent (2010), Song &

面向	ESM	ISM	DSM
	(2005)	Sliced SM (2019)	Ermon (2019), Ho et al. (2020)

（上表結論基於 Hyvärinen、Vincent、Song、Ho 等核心論文與後續實驗性報告。）proceedings.neurips.cc+3jmlr.org+3 蒙特利爾大學資訊研究所+3

5) 穩定性來源與常見改良（實務提示）

- **DSM 最穩**的原因：把目標視為「帶噪資料 → 已知解析標籤」的回歸任務；標準的回歸損失在深網路上表現穩定，且多尺度訓練可避免局部誤差累積。[蒙特利爾大學資訊研究所+1](#)
- **ISM/ESM 的不穩點**：二階項或 Hutchinson 隨機估計的高方差可能導致訓練震盪或需要非常多的樣本與步數去收斂。研究者提出 *sliced*、*stochastic trace*、*Hutch++* 等方法以降低 variance（參見 Sliced Score Matching (2019) 和 Hutch++ / trace estimator 的文獻）。[arXiv+1](#)
- **額外技巧（來自實務論文）**：Song & Ermon (2020) 提到的穩定化技巧（EMA、proper weighting of noise levels、predictor-corrector samplers）對於在高維資料上穩定訓練非常重要。[arXiv](#)

6) 每一個問題是否有對應研究？（總結式回答）

- **ESM 的基礎理論**：有 — Hyvärinen (2005) 詳盡推導並討論 ESM / ISM 的數學性質。[jmlr.org](#)
- **ISM 的實作/trace 估計問題與改進**：有 — Hutchinson estimator（古典）與後續改進（Hutch++、現代方差分析等）；Sliced Score Matching (Song et al., 2019) 提供可擴展替代方法。[arXiv+1](#)
- **DSM 在大型神經網路與 diffusion model 的實務成功與穩定化技巧**：有 — Vincent (DAE ↔ DSM) 及 Song/Ho 等關鍵工作，並有 NeurIPS/ICLR 相關實作改良（noise schedule、EMA、predictor-corrector）。[蒙特利爾大學資訊研究所+2arXiv+2](#)

7) 重要原文 (可點的參考 / 閱讀順序)

1. **Hyvärinen A. (2005)** — *Estimation of non-normalized statistical models by score matching*. JMLR. (ESM/ISM 理論基礎) . PDF:
<https://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf>. jmlr.org
2. **Vincent P. (2010/2011)** — *A connection between score matching and denoising autoencoders*. Tech report. (DAE \leftrightarrow DSM) . PDF:
https://www.iro.umontreal.ca/~vincentp/Publications/smdae_techreport.pdf. 蒙特利爾大學資訊研究所
3. **Song & Ermon (2019)** — *Generative Modeling by Estimating Gradients of the Data Distribution* (NCSN). arXiv/PDF:
<https://arxiv.org/abs/1907.05600>. [arXiv](https://arxiv.org)
4. **Song et al. (2020/2021)** — *Score-Based Generative Modeling through SDEs* (ICLR 2021). arXiv/PDF: <https://arxiv.org/abs/2011.13456>. (multi-scale / SDE view) . [arXiv](https://arxiv.org)
5. **Ho, Jain, Abbeel (2020)** — *Denoising Diffusion Probabilistic Models (DDPM)*. NeurIPS 2020. PDF:
<https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf>. (diffusion/DSM 實作) .
proceedings.neurips.cc
6. **Song, Garg, Shi, Ermon (2019)** — *Sliced Score Matching* (可擴展的 score matching 近似) . arXiv/PDF: <https://arxiv.org/abs/1905.07088>.
[arXiv](https://arxiv.org)
7. **Skorski (2020)** — *A modern analysis of Hutchinson's trace estimator* (Hutchinson variance/性質分析) . arXiv:
<https://arxiv.org/pdf/2012.12895>. [arXiv](https://arxiv.org)

- 擴散模型的「反向過程」為何需要準確的 **score** 函數？

Ans:

要點一 — 反向 SDE/ODE 的 drift 明確包含 score

數學上，若 forward process 是一個 SDE（把資料加入噪聲走向簡單先驗），Anderson（1982）與後續 Score-based work 證明：**reverse-time SDE 的 drift 需要 time-dependent marginal 的 score $\nabla_x \log p_t(x)$** 。換句話說，反向動力學直接用到 score 作為「方向場」，沒有正確的 score 就無法構造正確的反向 drift。

- 來源：Anderson (1982)；Song et al. (2020/21) 的 SDE 框架詳細推導了 reverse SDE 依賴 score 的形式。[IDEAS/RePEc+1](#)
-

要點二 — score 誤差會透過時間積累，改變最終分佈（誤差放大）

反向過程是時間串列／迭代的過程：每一步若使用有偏或有 variance 的 score 估計，都會把誤差帶到下一步，隨著步數誤差可能被放大或以系統方式累積，導致生成樣本偏離目標分佈。多篇理論／實驗工作證明 score 的估計精度直接相關於最終採樣的分佈誤差（例如 Wasserstein 或 KL bounds）。

- 來源：最近對「score 誤差累積」做理論與實驗分析的工作（例如 “The Accumulation of Score Estimation Error” 與 Chen et al. 2023 等）給出上界並展示了誤差如何影響生成品質。[OpenReview+1](#)
-

要點三 — Score 精準度影響兩件最直接的東西：樣本質量（perceptual quality）與 likelihood（數值近似/概率準確度）

- 在實務上（影像生成），score 的不準確常表現為模糊、失真或模式崩潰（mode collapse）。Ho et al.（DDPM）和 Nichol & Dhariwal 等實驗工作也指出改善 score 訓練與 noise schedule 能提升樣本質量與 likelihood。[NeurIPS 會議論文集+1](#)
 - 在理論上，若想證明 reverse process 近似目標分佈（或給出 likelihood 下界），需要對 score 的估計誤差有量化界（這是多篇近年工作推進的方向）。[Proceedings of Machine Learning Research+1](#)
-

要點四 — 為什麼部分方法（**predictor-corrector**、**stochasticity**）可以減緩 **score** 誤差影響

- 實作上常用 **predictor-corrector (P-C)** 框架：predictor 用一次步進近似 reverse SDE / ODE，corrector（例如 Langevin 提振步）用估計的 score 做多次小步修正。這在數值上能部分抵消 score 的估計誤差，提升穩定性與樣本質量。Song et al. 就提出並實驗驗證了這套方法。[arXiv](#)
- 另外，在某些情況下隨機性（**stochastic sampling**）能起到「糾錯」作用，使系統不那麼容易陷入由估計誤差造成的錯誤吸引子；但 **stochasticity** 也會放大方差，最佳做法依 problem 與 score 精度而定。近期工作在分析 **stochasticity** 的正負效應。[OpenReview](#)

要點五 — 實務上的證據：**score** 越準，**sampling** 越好（實驗與 **likelihood**）

大量實驗性研究（DDPM、NCSN / Score-SDE、Improved DDPM 等）表明：更好的 score estimator（例如更深的網路、better conditioning、更合適的 noise schedule、更多 training steps）通常導致更好的 FID / Inception score 與更高的 likelihood；反之，欠佳的 score 會使生成品質下降。這支持了「反向過程需要準確 score」的直觀。[NeurIPS 會議論文集+2arXiv+2](#)

研究與參考文獻（已查證、重要條目，點開可看原文）

1. **Song, Y., et al. (2020/2021)** — *Score-Based Generative Modeling through Stochastic Differential Equations*. arXiv (ICLR).
 - 內容：提出 SDE 框架、導出 reverse-time SDE（顯示 reverse drift 包含 score）、提出 predictor-corrector 等採樣方法並展示實驗。
 - 連結（PDF）：<https://arxiv.org/pdf/2011.13456.pdf>. [arXiv](#)
2. **Ho, J., Jain, A., & Abbeel, P. (2020)** — *Denoising Diffusion Probabilistic Models (DDPM)*. NeurIPS 2020.

- 內容：實作與實驗基準，建立 DSM/score-based 與變分推理的聯繫，並給出演算法細節，證明高品質生成需要良好的反向（即良好 score）。
 - 連結（PDF）：
<https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf>. [NeurIPS 會議論文集](#)
3. **Anderson, B. D. O. (1982)** — *Reverse-time diffusion equation models*. Stochastic Processes and their Applications.
- 內容：reverse-time SDE 的數學基礎（早期嚴謹推導），是後來 score-based reverse SDE 理論的基礎。
 - 引文頁：<https://ideas.repec.org/a/eee/spapps/v12y1982i3p313-326.html>. [IDEAS/RePEc](#)
4. **Chen, M., Huang, K., Zhao, T., Wang, M. (2023)** — *Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data*. (ICML/MLR proceedings)
- 內容：分析 score 估計與分佈恢復之間的關係，在低維流行情境下給出誤差界與 sample complexity。對於「score 精度如何影響 recovery」提供理論依據。
 - 連結（PDF）：
<https://proceedings.mlr.press/v202/chen23o/chen23o.pdf>.
[Proceedings of Machine Learning Research](#)
5. **“The Accumulation of Score Estimation Error” (recent work, OpenReview / arXiv)** — 分析 score estimation error 在 diffusion sampling 中如何累積、給出界與實證。
- 連結（OpenReview / PDF）：
<https://openreview.net/pdf?id=end8EBwFOU>. [OpenReview](#)

總結（一句話）

反向過程的 drift 明確依賴 time-dependent score；若 score 被估錯，反向 SDE/ODE 會使用錯誤的方向場，誤差會隨時間迭代累積，最終導致樣本品質 / likelihood 下降 —— 因此準確的 **score** 對於正確還原資料分布是核心且必要（或至少非常重要）的。多篇理論與實驗研究支持並量化了這一點。

[OpenReview+3arXiv+3NeurIPS 會議論文集+3](#)

● 是否能將 **DSM** 理解為對 **score** 的一種自監督學習？

Ans:

是的。DSM 本質上就是一種自監督學習方法 —— 訓練標籤是從資料自動生成（把 clean sample 加已知噪音，然後用那個條件分佈的解析式當作監督目標），網路學習從 noisy x 預測這個標籤（也就是 noisy 分布的 score）。這一點在理論與實作上都有大量研究支持。

為什麼可以這樣理解（重點解釋，簡潔）

1. **自監督的特徵**：自監督就是用資料本身構造標籤（例如遮住一塊影像要模型預測被遮住的部分）。DSM 做的就是將 x_0 加噪得到 x ，而標籤 $\nabla_x \log p(x | x_0)$ （對常見的 Gaussian corruption 有解析式 $-(x - x_0)/\sigma^2$ ）是由資料生成程序自帶的 —— 沒有人為外部標註。這完全符合自監督的定義。
 2. **數學連結 (DAE \leftrightarrow score)**：Vincent 等人的工作嚴謹地把去噪自編碼器（DAE）的訓練目標和 score matching（學分數場）連起來，證明了「去噪等價於學 noisy distribution 的 score」——這就是把去噪視為學分數的一種自監督機制。
 3. **實務上就是回歸任務**：在 Gaussian DSM 裡，訓練變成「輸入 noisy x 、標籤 $-(x - x_0)/\sigma^2$ 」的回歸問題；這種把合成標籤當監督信號的做法正是自監督學習的經典模式。
-

重要且已被查證的研究（核心參考，點連結可看原文）

- **Vincent, P. (2010/2011)** — *A Connection Between Score Matching and Denoising Autoencoders*. (技術報告 / 論證 DAE 與 score matching 的等價) . PDF:
https://www.iro.umontreal.ca/~vincentp/Publications/smdae_techreport.pdf. 蒙特利爾大學資訊研究所+1
- **Alain, G. & Bengio, Y. (2014)** — *What Regularized Auto-Encoders Learn from the Data-Generating Distribution*. (證明某些自編碼器會學到 score 的近似；與 DAE/DSM 觀點一致) . PDF:
<https://jmlr.org/papers/volume15/alain14a/alain14a.pdf>. jmlr.org+1
- **Hyvärinen, A. (2005)** — *Estimation of Non-Normalized Statistical Models by Score Matching*. (score matching 的理論基礎) . PDF:
<https://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf>. jmlr.org
- **Song, Y. & Ermon, S. (2019)** — *Generative Modeling by Estimating Gradients of the Data Distribution*. (把 DSM/score-matching 用於生成模型，採用多尺度 Gaussian corruption 與 Langevin sampling) . arXiv:
<https://arxiv.org/abs/1907.05600>. arXiv

(以上文獻我已檢索並確認為公開、可下載的來源。)

補充說明（實務/直覺）

- DSM 是自監督的一個特定實例：它用「加噪 → 去噪的已知規則」產生標籤，然後做回歸去預測 score。這跟現代自監督視覺／語言中的「遮蔽預測」「拼圖還原」「對比學習造假標籤」等方法在形式上很類似。
- 在 diffusion / score-based generation 的現代框架中，DSM 被進一步擴展為 **time-conditional** 的自監督任務（網路學 $s_\theta(x, t)$ ），仍屬自監督範疇：標籤由 x_0 和選定的 noise level t 自動產生。參見 Song & Ermon (2019)。

● DAE 與 Diffusion Model 的能量函數有何相似與不同？

Ans:

核心相似處 (Why they look related)

1. 兩者都與 **energy-based view** 有關 ($\text{score} = \nabla \log p = -\nabla \text{energy}$)
 - DAE 的重建向量場 $r(x) - x$ 或其變體在適當條件下逼近資料分布的 score (即 log-density 的梯度)，從而隱含一個能量/未正規化 log-density 的結構。Vincent 與後續 Alain & Bengio 的分析詳述了這個連結 (DAE \leftrightarrow score matching)。 iro.umontreal.ca+1
2. 兩者都能透過 **score** (或等價場) 來做生成
 - 若你有某個點的 score，你可以用 Langevin dynamics 或 reverse SDE 把噪聲變成樣本——DAE/DSM 與 diffusion models 都利用「去噪」/score 來做生成或重建。Song & Ermon 在 score-based generative modeling 中把 DSM 作為核心訓練手段並展示了採樣流程。 arXiv

主要差異 (What makes them different in practice and in theory)

1. 目標分布的尺度與時間性 (DAE 通常是單尺度的 **smoothing** ; **Diffusion** 是 **time-dependent / multiscale**)
 - DAE 一般把資料用某個固定噪聲 (或小集合噪聲) 平滑，學到的是被平滑後分布的 score (local, single/noise-level)。而 diffusion/score-based 模型學的是隨時間 t 變化的一系列 marginal p_t 的 score (multi-scale)，能以連續或離散時間的反向流程逐步恢復資料。這使 diffusion models 在復原細節與全域結構上更靈活。 iro.umontreal.ca+1
2. 是否明確對應到 **全局** 能量函數 (**conservative field**)
 - **DAE / RCAE 類推導** 在許多理想化條件 (例如無限容量、適當正則化) 下能近似一個 conservative 場 (即存在潛在能量使其為

gradient)。Vincent 與 Alain/Bengio 的結果就是這種情形。

iro.umontreal.ca+1

- **Diffusion models 的 score networks** 通常直接 parameterize score $s_\theta(x, t)$ 作為任意向量場 (unconstrained)。若這個向量場不是 conservative (不是某個全局 scalar 能量的梯度)，則它不對應任何單一全域能量函數。因此，除非特別設計 (constrained parameterization 或同時學 energy)，否則 diffusion 的 score \neq 某個能量的 exact gradient。Salimans 等與其他工作討論了「應該直接模型化能量還是 score」的取捨。[OpenReview](#)

3. 訓練與估計目標的差別 (直接 vs 隱式 vs time-conditional)

- DAE/DSM 通常把訓練轉成可求解的去噪回歸 (在 Gaussian case 有 closed form label)，訓練穩定且易實作。Diffusion frameworks 則把 DSM 擴展為 time-conditional loss，並把這些 time-dependent score 用於 reverse SDE/ODE 採樣，從而可以得到更好的生成效果與 likelihood 計算 (在某些形式下)。[arXiv](#)

4. 生成機制與可解釋性

- DAE 的能量解釋多半是局部且基於 Parzen-window style 的平滑，產生樣本或評分需要額外步驟 (例如 MCMC)。Diffusion models 有明確的 forward / reverse dynamics (SDE/ODE)，因此在理論上能直接從 learned scores 建構 reverse process 並進行高品質生成。iro.umontreal.ca+1

研究與討論：哪些論文探討這些關聯？

以下為已查證且可取得的核心文獻 (每條都附有效連結) —— 這些工作直接或間接討論了 DAE、score、energy、以及 diffusion models 的關係：

1. **Vincent, Pascal (2010/2011)** — *A Connection Between Score Matching and Denoising Autoencoders*. (DAE \leftrightarrow score matching 的主要技術報告). PDF:

https://www.iro.umontreal.ca/~vincentp/Publications/smdae_techreport.pdf. iro.umontreal.ca

2. **Alain, Guillaume & Bengio, Yoshua (2014)** — *What Regularized Auto-Encoders Learn from the Data-Generating Distribution*. (分析正則化 autoencoder/DAE 學到的 score 與能量) . PDF: <https://jmlr.org/papers/volume15/alain14a/alain14a.pdf>. jmlr.org
 3. **Song, Yang & Ermon, Stefano (2020/2021)** — *Score-Based Generative Modeling through Stochastic Differential Equations*. (把 DSM/score learning 放到 SDE / time-dependent 框架；詳細討論 reverse SDE 與 sampling) . arXiv / PDF: <https://arxiv.org/pdf/2011.13456.pdf>. [arXiv](https://arxiv.org)
 4. **Du, Y., et al. (2023)** — *Compositional Generation with Energy-Based Diffusion Models*. (把 diffusion 與 EBM 的視角結合，討論如何把能量式 parameterization 用於 diffusion 步) . PDF: <https://proceedings.mlr.press/v202/du23a/du23a.pdf>. [Proceedings of Machine Learning Research](https://proceedings.mlr.press/v202/du23a/du23a.pdf)
 5. **T. Salimans (2020 talk / paper)** — *Should EBMs model the energy or the score?* (討論直接 parameterize score vs parameterize energy 的取捨，並指出若 score network 未被約束為 conservative，則可能沒有對應能量) . PDF: <https://openreview.net/pdf?id=9AS-TF2jRNb>. [OpenReview](https://openreview.net/pdf?id=9AS-TF2jRNb)
 6. 近年工作 (2023–2025)：有許多文章在嘗試把 diffusion 與 EBM 更緊密結合 (例如 Energy-based Diffusion Language Models, EGC 等)，這顯示學界正積極探索「把 diffusion 的穩定性 + EBM 的靈活性」結合的路徑。例：M. Xu et al. (2024) EDLM (ICLR 2025), Guo et al. (ICCV 2023) 等。 [arXiv+1](#)
-

實務要點 (你可能會關心的結論)

- 如果你想直接學一個能量函數 (可以評分、做任務特定能量引導、或用已知 EBM 工具)，你要麼：
 1. 設計 parameterization 使得你的 score 是 conservative (即可積分得一個 scalar energy)，或
 2. 同時學一個 explicit energy (或用 NCE、contrastive 或 score-induced tricks)。相關文獻討論如何做 (Salimans, Du 等)。

[OpenReview+1](#)

- 如果你只想生成高品質樣本並且要實作方便，直接 parameterize time-dependent score (diffusion / DSM) 是業界最佳實踐；但要注意：這類 score network 未必對應單一全局能量函數。[arXiv](#)

哪些問題仍然是 open / active research ?

- **在實務上，如何把 diffusion 的穩定採樣與 EBM 的 explicit 能量優勢結合（尤其在離散或序列資料）**仍是活躍研究領域（近期多篇論文嘗試結合兩者）。[arXiv+1](#)
- **unconstrained score networks** 是否應該／如何被約束以對應 **conservative fields**（能量函數）——有技術挑戰與理論問題，部分研究正在做保守性檢驗或提出 constrained parameterizations。[chen-hao-chaos.github.io](#)

● 是否存在非高斯雜訊版本的 **DSM**？效果如何？

Ans:

有的：存在多種把 DSM（去噪分數匹配）推廣到 非高斯噪聲 的方法與研究

「DSM 原本常用的是 Gaussian 雜訊（因為條件分布和標籤 $\nabla_x \log p(x | x_0)$ 有簡單解析式），那如果我改用別種雜訊（像是 Gamma、Laplace、或重尾的 α -stable／Lévy 噪聲），可不可以？效果會怎麼變？」

可以 — 研究顯示可以用非高斯或非等方差噪聲來做 DSM / diffusion / score-based 建模；這類做法在某些情形（例如處理重尾分布、提升對離群值的魯棒性、或更靈活的生成動態）上有優勢，但推導會更複雜（條件 score 可能無解析式）、數值與理論細節需要特別處理。文獻包含理論推導、數值實驗與方法擴展。

要點解析（重點 + 實務影響）

1. 為什麼會想用非高斯噪聲？

- 處理非典型資料（heavy-tailed、含離群／不對稱雜訊）時，高斯噪聲有時不是最自然的模型。非高斯噪聲可提供更合適的平滑、不同的泛化偏好，或提供 heavy-tailed prior 的採樣能力。

2. 理論/推導變難：條件 score 不一定有 closed form

- Gaussian 的好處在於 $\nabla_x \log p(x | x_0) = -(x - x_0)/\sigma^2$ 有解析式；非高斯案例常常沒有這種簡單標籤，或需要更複雜的導數/分布積分技巧。研究者會：(a) 對某些家族做解析推導（例如 Gamma、Tweedie、Laplace）；(b) 用數值或學習式估計條件 score；(c) 改用 SDE/Lévy-process 框架推導 reverse 動力學。

3. 實驗結果：在一些任務上表現更好／更穩健

- Heavy-tailed DSM / Lévy noise 在處理不平衡資料、或需要 heavy-tailed generation 的任務上顯示改善（更穩健的 score 估計、可控的收斂行為）。非等方差或空間相關（non-isotropic）Gaussian 也能提升在影像上的表現或數值穩定性。

4. 實務注意

- 當條件 density 沒解析式時你要：自己推導 $\nabla_x \log p(x | x_0)$ （如果可行）、或把目標改為直接學 $s_\theta(x)$ （而非用 closed-form label）、或使用變分近似/EM 類方法。
- 在採樣端（reverse process）若用 Lévy / jump processes，採樣器需處理跳躍（jump）事件，數值上比純 Gaussian SDE 更複雜。

代表性論文（已查證、含連結）

1. “Non Gaussian Denoising Diffusion Models” — E. Nachmani et al., 2021 (arXiv)

- 探討用 Gamma 等非高斯噪聲於 diffusion，給出理論與實驗。這篇在 community 被引用較多（示範非高斯噪聲可行）。

- arXiv: <https://arxiv.org/abs/2106.07582>. [arXiv](#)

2. “Heavy-tailed denoising score matching” — J. Deasy et al., 2021 (arXiv preprint)

- 研究用 heavy-tailed 噪聲（非 Gaussian）做 DSM 的理論與實驗，報告在某些設定下能改善 score estimation 與 sampling convergence，對不均衡資料有正面效果。
- arXiv: <https://arxiv.org/abs/2112.09788>. [arXiv](#)

3. “Score-based Generative Models with Lévy Processes” — Eunbi Yoon et al., 2023 (NeurIPS)

- 把 Lévy (α -stable, 重尾) 過程引入 score-based 模型，推導 reverse dynamics，並用於生成任務；這是把 jump / heavy-tailed noise 用於 score models 的代表作。
- PDF (NeurIPS):
https://papers.nips.cc/paper_files/paper/2023/file/8011b23e1dc3f57e1b6211ccad498919-Paper-Conference.pdf. [NIPS 會議論文](#)

4. “Denoising Lévy Probabilistic Models” — (arXiv, 2024)

- 近期 work 把 Lévy 型噪聲與 denoising/diffusion 框架系統化，討論 heavy-tailed prior 的好處與數值挑戰。
- arXiv: <https://arxiv.org/abs/2407.18609> (或 HTML). [arXiv](#)

5. “Score-based Denoising Diffusion with Non-Isotropic Gaussian Noise” — V. Voleti et al., 2022 / NeurIPS 2023 related

- 探討非等方差 / 非同向 (non-isotropic) Gaussian noise (仍屬 Gaussian 類，但放寬 isotropy 假設)，並展示在 CIFAR-10 等上與標準噪聲相比的表現與數學推導。
- arXiv / PDF: <https://arxiv.org/pdf/2210.12254.pdf>. [arXiv](#)

6. Tweedie / Generalized-noise denoising work (e.g. Kim et al., CVPR 2022)

- 顯示用 Tweedie 類（包含指數族、某些非高斯）可建立 closed-

form denoising 推估 (Tweedie formula)，擴展去噪的適用雜訊範圍。

- CVPR paper:

https://openaccess.thecvf.com/content/CVPR2022/papers/Kim_Noise_Distribution_Adaptive_Self-Supervised_Image_Denoising_Using_Tweedie_Distribution_and_CVPR_2022_paper.pdf. [CVF 開放存取](#)

● DSM 損失中的常數項 C 是否對訓練完全沒有影響？

Ans:

1. 理論上：如果推導成立且你實作時沒有近似，那這個常數 C 不含參數 θ ，對參數的梯度為零，因此對優化（參數更新）完全沒有影響^{**}。換句話說，忽略這個常數不會改變最終的最優 θ 或參數更新方向。這是 score-matching / DSM 的標準結論。 [Journal of Machine Learning Research+1](#)
2. 實務上的 **caveats**（為什麼你還是要注意）：雖然理論上 C 不影響梯度，但在真實訓練 / 實作中有幾種情況會讓它間接「看起來」重要或造成影響：
 - 加權 / 多尺度 **loss** 的情況：當你對不同 noise level（不同 σ 或不同時間 t ）使用不同比重（weighting）時，每個尺度可能有自己的常數項。雖然每個常數仍然不含 θ ，但權重會改變不同尺度上的相對貢獻，從而實際改變優化行為（這是因為你改變了真正由 network 最小化的有用項的相對權重，而不是因為 C 本身）。因此在 multi- σ 設計、或採用時間條件化 loss 時要慎選 weighting。 [arXiv](#)
 - 近似 / 數值實作會破壞理論假設：若你在評估期望時用有限樣本估計、或用近似的 conditional score（例如用蒙特卡羅近似、離散化 SDE、或用不精確的 trace estimator），則這些近似可能引入 θ -依賴的偏差，讓原本被視為常數的項實際上與 θ 有弱耦合，

進而影響訓練。文獻有比較 ISM/DSM 在有限樣本下差異的實驗與討論。 [Semantic Scholar+1](#)

- **監控 / 比較不同模型時的誤導**：即便 C 不影響梯度，常數會移動 loss 的絕對值。若你用 raw DSM loss 值做早停、learning-curve 比較或跨實驗報表，忽略常數可能導致誤解（不同方法或不同 noise-level 的 loss baseline 不可直接比較）。應使用相同的基準或移除公用常數來比較。（這是工程實務建議，相關教材與 review 有說明。） [Ernest K. Ryu](#)

3. 總結判斷：

- **理論層面**： C 對優化方向與最終 θ 沒有直接影響（可安全忽略）。 [Journal of Machine Learning Research+1](#)
- **實務層面**：在多尺度加權、有限樣本估計或其他近似情況下，間接效應會出現 — 這些情況需要仔細設計 loss weighting、估計器以及監控手段。若不注意，雖不是 C 本身的因果，但整體訓練行為會受影響。 [Semantic Scholar+1](#)

重要參考（已查證、可點開閱讀）

- **Hyvärinen, A. (2005)** — *Estimation of Non-Normalized Statistical Models by Score Matching*. (score matching 的數學基礎；其中指出常數項不依賴參數 θ) . PDF: <https://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf>. [Journal of Machine Learning Research](#)
- **Vincent, P. (2010/2011)** — *A Connection Between Score Matching and Denoising Autoencoders*. (DSM 的推導與常數項出現、等價關係的討論) . PDF: https://www.iro.umontreal.ca/~vincentp/Publications/smdae_techreport.pdf. [IRO 蒙特利尔大学](#)
- **Song, Y. & Ermon, S. (2019/2021)** — *Generative Modeling by Estimating Gradients of the Data Distribution / Score-based Generative Modeling through SDEs*. (在多尺度/時間條件化 DSM 中對 loss 結構與

weighting 的實作討論) . arXiv / PDF: <https://arxiv.org/abs/1907.05600>
and <https://arxiv.org/pdf/2011.13456.pdf>. [arXiv+1](#)

- **Artemev, Das, et al. (Comparing Implicit and Denoising Score-Matching)** — (比較 ISM 與 DSM, 指出在有限樣本或實作近似下, 二者會出現差異, 並討論常數/偏差的實務影響) . 摘要/來源頁面: (比較分析可參考 summary)
<https://www.semanticscholar.org/paper/Comparing-Implicit-and-Denoising-Score-Matching-Artemev-Das/3ee1589a0473c2314f00e29bccbb1fe09b0ae5b>. Semantic Scholar
-

實務小建議 (工程上你可以立刻採取的做法)

1. 在訓練上: 直接忽略 C , 因為它不會改變梯度。把注意力放在 noise-schedule、loss weighting、網路容量與數值穩定技巧 (EMA、梯度裁剪) 上。 [arXiv](#)
 2. 在比較/報告上: 若要比較不同方法的 loss 值或報表, 移除/對齊常數基準或比較同一 loss 的相對變化 (例如 validation metric 如 FID / log-likelihood), 不要直接比較 raw DSM loss 的絕對值。 [Ernest K. Ryu](#)
 3. 若使用近似估計 (有限樣本、Hutchinson 等): 檢查估計器是否引入與 θ 相關的偏差; 在必要時用更多樣本、降低估計器方差或使用控制變量 (variance reduction) 技巧。相關比較研究指出這些近似會使 DSM / ISM 在有限樣本下表現不同。 [Semantic Scholar](#)
- 對任意 **SDE**, 反向時間 **SDE** 存在且唯一解的充分條件是什麼? 有哪些技術細節 (例如正則性、邊界條件) 需要注意?

Ans:

一、結論要點 (快速摘要)

1. 不是所有 **SDE** 的時間反轉都能自動得到另一個（良態的）**SDE**；要能得到「反向 **SDE** 並且解存在且唯一」通常需要：
 - 正向 **SDE** 的係數（drift、diffusion）具有適當的正則性（例如連續、Lipschitz 或更強： C^1 、 C^2 ），且正向 **SDE** 有唯一強解；
 - 每個中間時刻 t 的邊際密度 $p_t(x)$ 存在且夠光滑（至少要能求出 $\nabla_x \log p_t(x)$ ）；
 - 擴散矩陣非退化（或滿足相應的 Hörmander/ellipticity 條件）以保證 transition density 的存在與正則性；
 - 若擴散係數依賴狀態（state-dependent $G(x, t)$ ），反向漂移會包含額外發散與矩陣項（需能定義且可積分）。
在這些條件下，反向過程可寫成一個 **SDE**（Anderson, Haussmann–Pardoux 等給出嚴格定理）。[科學直接+1](#)
2. 技術要點（實務上最重要的）：
 - 需要密度 p_t ：反向漂移常含 $\nabla_x \log p_t(x)$ （或類似條項），若 p_t 不存在或不夠光滑，反向 **SDE** 形式就不成立或沒有意義。
[projecteuclid.org+1](#)
 - 非退化性（或 Hörmander 條件）用以保證 transition density 為 C^∞ ；沒有非退化性者，密度可能不存在或奇異，時間反轉困難。
[ma.imperial.ac.uk](#)
 - 邊界/吸收/反射條件 會改變反向律（邊界可能在時間反轉時變成不同類型條件），需專門分析（見 Cattiaux 等關於有邊界的處理）。
[科學直接](#)
3. 研究與文獻： 這一題目有大量嚴謹理論工作（不是 open problem）。關鍵論文與教科書包括 Anderson (1982)、Haussmann & Pardoux (1986)、以及多本 **SDE** 教科書／綜述（例如 Karatzas & Shreve、Pavliotis）。下方列出與此最相關且可查證的引用與連結。
[科學直接+2projecteuclid.org+2](#)

二、較嚴格的 充分條件（可作為 theorem 假設的集合）

以下條件通常是能保證「time-reversed process is again a diffusion with a well-posed SDE」的一組常用假設（不同文獻會有變形、或更弱/更強的版本）。

1. Forward SDE 已有唯一強解（well-posed forward SDE）

- $dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t$ 在每個初始值下有唯一強解（常見條件： b, σ 對 x 全域 Lipschitz 且線性增長）。（這保證 forward process 是良態的 Markov process。）

ma.imperial.ac.uk

2. Transition density 存在且夠光滑

- 對每個 $0 < t < T$ ， X_t 有密度 $p_t(x)$ （相對於 Lebesgue），且 $p_t(x)$ 在 x 上至少可微（理想情況下 $C^{1,2}$ 類），使得 $\nabla_x \log p_t(x)$ （score）能定義且為合理函數（可積/受界或局部平方可積）。
- 這通常需要擴散矩陣 $\sigma\sigma^T$ 非退化（uniform ellipticity）或滿足 Hörmander 條件以取得平滑密度。Haussmann & Pardoux 的結果以「存在適當的解於 Kolmogorov / Fokker-Planck 方程」為假設展開。projecteuclid.org+1

3. 正則性條件（係數 smoothness）

- $b(t, x)$ 與 $\sigma(t, x)$ 至少連續，且常取 C^1 或 C^2 以便在反向漂移項出現 $\nabla \cdot (\sigma\sigma^T)$ 等導數項時能合法化。若 σ 是 state dependent，反向式子包含額外的散度項（divergence，或 $\nabla[\cdot] \cdot (\sigma\sigma^T)$ ），所以需要這些偏導存在。Haussmann-Pardoux 提出的技術假設即在這方向。projecteuclid.org

4. 邊界與停止時間處理

- 若過程在有界空間內或與邊界互動（吸收/反射），反向過程可能需要另外的“邊界條件”或被改寫成帶邊界條件的擴散（Cattiaux 等人的工作討論這類情況）。若沒有邊界（全空間 \mathbb{R}^d ）通常情況簡單許多。[科學直接](http://science24.com)

在上述假設下（或其變形）可得到反向 SDE 的形式（Anderson, Haussmann-Pardoux）：大致上（簡化表示），若 forward SDE 是

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t,$$

則其 time-reversal $\tilde{X}_s = X_{T-s}$ 在 $s \in [0, T]$ 下若條件成立可寫成

$$d\tilde{X}_s = \tilde{b}(s, \tilde{X}_s) ds + \sigma(T-s, \tilde{X}_s) d\tilde{W}_s,$$

其中反向漂移 \tilde{b} 大致等於（符號/位置依不同慣例）：

$$\tilde{b}(s, x) = -b(T-s, x) + \text{div}(\sigma\sigma^\top)(T-s, x) - (\sigma\sigma^\top)(T-s, x) \nabla_x \log p_{T-s}(x),$$

（實際精確公式視 σ 的形態而定；這裡只是示意性列出重要項：原有 drift、散度項、與含 $\nabla \log p$ 的校正項）。Haussmann & Pardoux 與 Anderson 在嚴格假設下給出精確形式與證明。[科學直接+1](#)

三、技術細節與實務注意（較完整說明）

1. 為何需要 $\nabla \log p_t$ ：反向漂移中最關鍵的額外項是 $-(\sigma\sigma^\top)\nabla_x \log p_t(x)$ （或其同類），因此除非能估計／界定這個 score，否則無法寫出並模擬反向 SDE。這是 score-based generative modeling（Song et al., Song & Ermon, Song et al. ICLR 2021）中必須估計 score 的理由。[arXiv+1](#)
2. 密度不存在或退化時發生什麼？
 - 若在某些 t 上 p_t 不存在或不連續， $\nabla \log p_t$ 可能不存在／發散；反向過程即可能不是（或不能被表示成）傳統形式的 SDE。某些工作以弱解、martingale 問題或廣義分佈做處理，但理論更複雜（見 Millet–Nualart–Sanz 等）。[projecteuclid.org](#)
3. 邊界條件與反射/吸收：若 forward process 與邊界互動，反向過程邊界行為不一定是簡單反射，Cattiaux 等人給出有邊界情形的充分條件與構造方法。[科學直接](#)
4. 弱解 vs 強解：有時候只能證明反向過程存在為某種弱解（martingale problem 的解），而不是強解（以同一 Brownian motion 驅動）。Haussmann & Pardoux 的處理是以 martingale/生成元與 Kolmogorov 方程分析時間反轉屬性。[projecteuclid.org](#)

5. **數值與估計層面**：在應用（例如生成模型）中，我們通常沒有 p_t 的解析式，所以會用神經網路估計 $\nabla \log p_t$ 。理論上估計誤差會影響反向 SDE 的準確性與穩定性；這是實務研究的重要議題（Song et al., ICLR 2021 等提供實作與實驗分析）。（你上傳的那篇 ICLR 2021 的論文也討論 SDE 時間反轉在生成模型上的應用 — 本地檔案路徑：
`/mnt/data/2011.13456v2.pdf`）。[arXiv+1](#)
-

四、代表性重要參考（可直接打開／檢視）

我挑了幾篇核心、且你可直接查到 PDF / 期刊頁面的參考（每項都附上可以點開的連結）

1. **B. D. O. Anderson, “Reverse-time diffusion equation models”, Stochastic Processes and their Applications, 1982.**（反向 SDE 設定與主要形式）
 - 期刊頁面（Elsevier）：
<https://www.sciencedirect.com/science/article/pii/0304414982900515>. [科學直接](#)
2. **U. G. Haussmann & É. Pardoux, “Time-Reversal of Diffusions”, Annals of Probability, 1986.**（嚴謹理論化處理：何時時間反轉仍為擴散、技術假設）
 - Project Euclid: <https://projecteuclid.org/journals/annals-of-probability/volume-14/issue-4/Time-Reversal-of-Diffusions/10.1214/aop/1176992362.full>. [projecteuclid.org](#)
3. **P. Cattiaux et al., “Time reversal of diffusion processes with a boundary condition” (或相關工作)**（有邊界情形）、以及後續延伸文章討論在不同假設下的時間反轉。
 - 範例（可取）：
<https://www.sciencedirect.com/science/article/pii/0304414988901019>. [科學直接](#)
4. **教科書與綜述（背景與充分條件）：**

- Karatzas & Shreve, *Brownian Motion and Stochastic Calculus* (教科書, SDE 基礎與馬氏過程觀點)。DOI/書籍頁面：
<https://doi.org/10.1007/978-1-4684-0302-2>. [SpringerLink](#)
- G. A. Pavliotis, *Stochastic Processes and Applications* (章節中有 time-reversal 與對 Kolmogorov / Fokker-Planck 的討論)。PDF 教學筆記：
<https://www.ma.imperial.ac.uk/~pavl/PavliotisBook.pdf>.
[ma.imperial.ac.uk](https://www.ma.imperial.ac.uk)

5. 應用於生成模型 (實作面): Song et al., *Score-based generative modeling through SDEs*, ICLR 2021 — 把前述時間反轉定理套到生成模型, 並討論如何用 neural net 估計 score. [arXiv+1](#)

五、若無嚴格條件：哪些情形是 open / 難題？

- 當 **transition density** 不存在 (例如高度退化的擴散、或帶強吸收/跳躍的過程), 時間反轉要不要仍為擴散、或是否能以 **SDE** 表示, 則需要以更弱的框架 (martingale problem、廣義生成元、或以分佈/測度流描述) 處理。這些情形有活躍研究 (例如對退化擴散、infinite-dimensional SDE、或帶奇異邊界的情況), 但不是「完全 open」——只是理論複雜且各種條件很技術化。Millet–Nualart–Sanz、Cattiaux 等人在某些方向提供結果。 [projecteuclid.org+1](#)
- 推導反向 **SDE** 的嚴格假設有哪些？在擴散係數依賴狀態 (非齊次矩陣) 時, 公式如何改寫？

Ans:

一、結論要點 (快速摘要)

1. 不是所有 **SDE** 的時間反轉都能自動得到另一個 (良態的) **SDE**; 要能得到「反向 **SDE** 並且解存在且唯一」通常需要：
 - 正向 **SDE** 的係數 (drift、diffusion) 具有適當的正則性 (例如連

續、Lipschitz 或更強： C^1 、 C^2 ），且正向 SDE 有唯一強解；

- 每個中間時刻 t 的邊際密度 $p_t(x)$ 存在且夠光滑（至少要能求出 $\nabla_x \log p_t(x)$ ）；
- 擴散矩陣非退化（或滿足相應的 Hörmander/ellipticity 條件）以保證 transition density 的存在與正則性；
- 若擴散係數依賴狀態（state-dependent $G(x, t)$ ），反向漂移會包含額外發散與矩陣項（需能定義且可積分）。

在這些條件下，反向過程可寫成一個 SDE（Anderson, Haussmann–Pardoux 等給出嚴格定理）。[科學直接+1](#)

2. 技術要點（實務上最重要的）：

- 需要密度 p_t ：反向漂移常含 $\nabla_x \log p_t(x)$ （或類似條項），若 p_t 不存在或不夠光滑，反向 SDE 形式就不成立或沒有意義。
[projecteuclid.org+1](#)
- 非退化性（或 Hörmander 條件）用以保證 transition density 為 C^∞ ；沒有非退化性者，密度可能不存在或奇異，時間反轉困難。
[ma.imperial.ac.uk](#)
- 邊界/吸收/反射條件 會改變反向律（邊界可能在時間反轉時變成不同類型條件），需專門分析（見 Cattiaux 等關於有邊界的處理）。
[科學直接](#)

3. 研究與文獻：這一題目有大量嚴謹理論工作（不是 open problem）。關鍵論文與教科書包括 Anderson (1982)、Haussmann & Pardoux (1986)、以及多本 SDE 教科書／綜述（例如 Karatzas & Shreve、Pavliotis）。下方列出與此最相關且可查證的引用與連結。
[科學直接+2projecteuclid.org+2](#)

二、較嚴格的 充分條件（可作為 theorem 假設的集合）

以下條件通常是能保證「time-reversed process is again a diffusion with a well-posed SDE」的一組常用假設（不同文獻會有變形、或更弱/更強的版本）。

1. Forward SDE 已有唯一強解 (well-posed forward SDE)

- $dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t$ 在每個初始值下有唯一強解 (常見條件: b, σ 對 x 全域 Lipschitz 且線性增長)。(這保證 forward process 是良態的 Markov process。)

ma.imperial.ac.uk

2. Transition density 存在且夠光滑

- 對每個 $0 < t < T$, X_t 有密度 $p_t(x)$ (相對於 Lebesgue), 且 $p_t(x)$ 在 x 上至少可微 (理想情況下 $C^{1,2}$ 類), 使得 $\nabla_x \log p_t(x)$ (score) 能定義且為合理函數 (可積/受界或局部平方可積)。
- 這通常需要擴散矩陣 $\sigma\sigma^\top$ 非退化 (uniform ellipticity) 或滿足 Hörmander 條件以取得平滑密度。Haussmann & Pardoux 的結果以「存在適當的解於 Kolmogorov / Fokker-Planck 方程」為假設展開。 projecteuclid.org+1

3. 正則性條件 (係數 smoothness)

- $b(t, x)$ 與 $\sigma(t, x)$ 至少連續, 且常取 C^1 或 C^2 以便在反向漂移項出現 $\nabla \cdot (\sigma\sigma^\top)$ 等導數項時能合法化。若 σ 是 state dependent, 反向式子包含額外的散度項 (divergence, 或 $\nabla[\cdot] \cdot (\sigma\sigma^\top)$), 所以需要這些偏導存在。Haussmann-Pardoux 提出的技術假設即在這方向。 projecteuclid.org

4. 邊界與停止時間處理

- 若過程在有界空間內或與邊界互動 (吸收/反射), 反向過程可能需要另外的“邊界條件”或被改寫成帶邊界條件的擴散 (Cattiaux 等人的工作討論這類情況)。若沒有邊界 (全空間 \mathbb{R}^d) 通常情況簡單許多。 [科學直接](#)

在上述假設下 (或其變形) 可得到反向 SDE 的形式 (Anderson, Haussmann-Pardoux): 大致上 (簡化表示), 若 forward SDE 是

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t,$$

則其 time-reversal $\tilde{X}_s = X_{T-s}$ 在 $s \in [0, T]$ 下若條件成立可寫成

$$d\tilde{X}_s = \tilde{b}(s, \tilde{X}_s) ds + \sigma(T-s, \tilde{X}_s) d\tilde{W}_s,$$

其中反向漂移 \tilde{b} 大致等於（符號/位置依不同慣例）：

$$\tilde{b}(s, x) = -b(T-s, x) + \operatorname{div}(\sigma\sigma^\top)(T-s, x) - (\sigma\sigma^\top)(T-s, x) \nabla_x \log p_{T-s}(x),$$

（實際精確公式視 σ 的形態而定；這裡只是示意性列出重要項：原有 drift、散度項、與含 $\nabla \log p$ 的校正項）。Haussmann & Pardoux 與 Anderson 在嚴格假設下給出精確形式與證明。[科學直接+1](#)

三、技術細節與實務注意（較完整說明）

1. **為何需要 $\nabla \log p_t$ ：**反向漂移中最關鍵的額外項是 $-(\sigma\sigma^\top)\nabla_x \log p_t(x)$ （或其同類），因此除非能估計／界定這個 score，否則無法寫出並模擬反向 SDE。這是 score-based generative modeling（Song et al., Song & Ermon, Song et al. ICLR 2021）中必須估計 score 的理由。[arXiv+1](#)
2. **密度不存在或退化時發生什麼？**
 - 若在某些 t 上 p_t 不存在或不連續， $\nabla \log p_t$ 可能不存在／發散；反向過程即可能不是（或不能被表示成）傳統形式的 SDE。某些工作以弱解、martingale 問題或廣義分佈做處理，但理論更複雜（見 Millet–Nualart–Sanz 等）。[projecteuclid.org](#)
3. **邊界條件與反射/吸收：**若 forward process 與邊界互動，反向過程邊行為不一定是簡單反射，Cattiaux 等人給出有邊界情形的充分條件與構造方法。[科學直接](#)
4. **弱解 vs 強解：**有時候只能證明反向過程存在為某種弱解（martingale problem 的解），而不是強解（以同一 Brownian motion 驅動）。Haussmann & Pardoux 的處理是以 martingale/生成元與 Kolmogorov 方程分析時間反轉屬性。[projecteuclid.org](#)
5. **數值與估計層面：**在應用（例如生成模型）中，我們通常沒有 p_t 的解析式，所以會用神經網路估計 $\nabla \log p_t$ 。理論上估計誤差會影響反向 SDE 的準確性與穩定性；這是實務研究的重要議題（Song et al., ICLR 2021）

等提供實作與實驗分析)。(你上傳的那篇 ICLR 2021 的論文也討論 SDE 時間反轉在生成模型上的應用 — 本地檔案路徑：
[/mnt/data/2011.13456v2.pdf](#))。 [arXiv+1](#)

四、代表性重要參考(可直接打開／檢視)

我挑了幾篇核心、且你可直接查到 PDF / 期刊頁面的參考(每項都附上可以點開的連結)

1. **B. D. O. Anderson, “Reverse-time diffusion equation models”, *Stochastic Processes and their Applications*, 1982.** (反向 SDE 設定與主要形式)
 - 期刊頁面 (Elsevier):
<https://www.sciencedirect.com/science/article/pii/0304414982900515>. [科學直接](#)
2. **U. G. Haussmann & É. Pardoux, “Time-Reversal of Diffusions”, *Annals of Probability*, 1986.** (嚴謹理論化處理：何時時間反轉仍為擴散、技術假設)
 - Project Euclid: <https://projecteuclid.org/journals/annals-of-probability/volume-14/issue-4/Time-Reversal-of-Diffusions/10.1214/aop/1176992362.full>. projecteuclid.org
3. **P. Cattiaux et al., “Time reversal of diffusion processes with a boundary condition” (或相關工作)** (有邊界情形)、以及後續延伸文章討論在不同假設下的時間反轉。
 - 範例 (可取):
<https://www.sciencedirect.com/science/article/pii/0304414988901019>. [科學直接](#)
4. **教科書與綜述 (背景與充分條件):**
 - Karatzas & Shreve, *Brownian Motion and Stochastic Calculus* (教科書, SDE 基礎與馬氏過程觀點)。DOI/書籍頁面:
<https://doi.org/10.1007/978-1-4684-0302-2>. [SpringerLink](#)

- G. A. Pavliotis, *Stochastic Processes and Applications* (章節中有 time-reversal 與對 Kolmogorov / Fokker–Planck 的討論)。PDF 教學筆記：

<https://www.ma.imperial.ac.uk/~pavl/PavliotisBook.pdf>.
[ma.imperial.ac.uk](https://www.ma.imperial.ac.uk)

5. 應用於生成模型（實作面）：Song et al., *Score-based generative modeling through SDEs*, ICLR 2021 — 把前述時間反轉定理套到生成模型，並討論如何用 neural net 估計 score. [arXiv+1](#)

五、若無嚴格條件：哪些情形是 open / 難題？

- 當 **transition density** 不存在（例如高度退化的擴散、或帶強吸收/跳躍的過程），時間反轉要不要仍為擴散、或是否能以 SDE 表示，則需要以更弱的框架（martingale problem、廣義生成元、或以分佈/測度流描述）處理。這些情形有活躍研究（例如對退化擴散、infinite-dimensional SDE、或帶奇異邊界的情況），但不是「完全 open」——只是理論複雜且各種條件很技術化。Millet–Nualart–Sanz、Cattiaux 等人在某些方向提供結果。 [projecteuclid.org+1](https://projecteuclid.org)

● OU 過程的 score function

$$\nabla_x \log p_t(x | x_0) = -\frac{x - e^{-\beta t} x_0}{\sigma_t^2}$$

為何其形式與標準高斯分布的 **score** 相同？是否暗示 OU 流是高斯族封閉的？

Ans:

重點答案

1. 為什麼

條件轉移密度 $p_t(x | x_0)$ 是一個一維常態分布（Gaussian）：

$$p_t(x | x_0) = \mathcal{N}(x; e^{-\beta t} x_0, \sigma_t^2),$$

因為 OU 方程是線性齊次（affine）SDE：解可以寫成初值的線性函數加上一個高斯積分（標準布朗運動的線性變換），因此條件分布必為高斯。

而對於 $\mathcal{N}(\mu, \sigma^2)$ 的分布，score（對 x 的梯度）恰為

$$\nabla_x \log \mathcal{N}(x; \mu, \sigma^2) = -\frac{x - \mu}{\sigma^2}.$$

把 $\mu = e^{-\beta t} x_0$ 代回就得到你寫的那個形式。[維基百科+1](#)

2. 是否暗示「OU 流是高斯族封閉」？

是的 —— 在這裡有兩層含義：

- （條件）轉移密度為高斯：對於任意固定 x_0 ，在時間 $t > 0$ 的條件分布 $p_t(\cdot | x_0)$ 是高斯（這是線性 SDE 的一般事實）。

ma.imperial.ac.uk

- 若初始分布本身為高斯，那麼全體邊際分布 $p_t(x)$ （不條件於 x_0 ）仍為高斯；也就是說 OU 在「高斯分布的類（family）」內不會走出來 —— 這通常稱為線性 SDE 對高斯族的閉合性（Gaussian closure）。[維基百科](#)

直觀

- OU SDE（簡化形式）

$$dx_t = -\beta x_t dt + \sigma dW_t.$$

- 解的結構是「初始值乘上衰減因子」加上一個由布朗運動積分構成的高斯隨機項：

$$x_t = e^{-\beta t} x_0 + \int_0^t e^{-\beta(t-s)} \sigma dW_s.$$

那個積分本身是均值 0、已知變異數的高斯變量，所以條件在 x_0 下整體是高

斯。 ma.imperial.ac.uk

- 因此條件 `score` 直接變成 $-(x - \text{mean})/\text{var}$ ，和任意高斯分布的 `score` 公式一致 —— 沒有奇怪的額外項。
-

哪些研究/參考可查證

- 綜述 / 教科書（對 OU 性質與線性 SDE 的嚴格敘述）
Pavliotis, G. A., *Stochastic Processes and Applications* (lecture notes / book). （有關 OU 為 Gauss–Markov process、轉移密度公式等）。
ma.imperial.ac.uk
（連結 PDF）：<https://www.ma.imperial.ac.uk/~pavl/PavliotisBook.pdf>
- 綜覽條目（快速參考 OU 的轉移密度、平均與變異數公式）
Wikipedia: *Ornstein–Uhlenbeck process*. [維基百科](#)
（連結）：
https://en.wikipedia.org/wiki/Ornstein%E2%80%93Uhlenbeck_process
- 與 score-based generative modeling（SDE/反向 SDE）關聯的現代文獻（說明為何知道解析 `score` 在實務上很有用，且 OU/線性案例常作為解析例子或 bridge）
Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., Poole, B., *Score-Based Generative Modeling through Stochastic Differential Equations*, 2021 (arXiv). （提供 reverse-time SDE 與 `score` 的整體框架）。[arXiv](#)
（連結）：<https://arxiv.org/abs/2011.13456>
- 應用：用廣義 OU 或 OU-bridge 作為可解析的 forward-noising model，並以此比較／訓練（示例性近代論文）
Yue et al., *Image Restoration Through Generalized Ornstein–Uhlenbeck Bridge* (ICML 2024 類別/報告)，其中直接利用解析的 OU 條件密度與 `score` 作為基準或訓練目標的一部分。pengxuwei.github.io
（連結 PDF）：https://pengxuwei.github.io/assets/files/2024ICML-Bridge/Image_Restoration_Through_Generalized_Ornstein-Uhlenbeck_Bridge.pdf

- 當 $x_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$ 時，

合併初始隨機性後的 $p_t(x)$ 仍為高斯分布，這樣的結構是否對所有線性 SDE 成立？

Ans:

是的，這樣的結構不僅對 OU (Ornstein–Uhlenbeck) 過程成立，也對非常一般的線性 SDE (Linear SDE) 成立 — 只要漂移和擴散都是線性形式（或更普遍是加性、高斯噪聲），那麼如果初始分布是高斯 (Gaussian)，整條過程在每個時間點的邊際分布仍然是高斯。

為什麼對所有線性 SDE 都成立 — 理論說明

1. 線性 SDE 的解是高斯過程

- 對於一般線性 SDE (例如 $dX_t = A(t)X_t dt + B(t) dW_t$)，其解可以寫成初始值經過線性演化 + 一個由布朗運動 (Wiener 過程) 積分而來的隨機項。由於 Wiener 過程本身是高斯過程，且線性轉換下高斯過程仍是高斯，因此 X_t 為高斯過程。這在控制理論 / 濾波理論 (Kalman–Bucy filter) 的文獻裡非常常見。

maxim.ece.illinois.edu+1

- 更明確地，在「Statistics of Linear Stochastic Differential (SDE)」中就提到：解是線性映射 (linear transformation) of Brownian motion，因此是高斯過程 (Gaussian process)；邊際分布 $p(x, t) = \mathcal{N}(m(t), P(t))$ ，其中 mean 和 covariance 分別通過線性微分方程 (transition matrix) 求出。 csml-cam.github.io

2. 初始隨機性 (高斯) 被保留

- 若 $X_0 \sim \mathcal{N}(m_0, P_0)$ ，那麼透過線性的演化 (mean) 加上 Gaussian 隨機項， X_t 的 mean 依然是線性變換 $m(t) = \Psi(t, 0)m_0 + \dots$ ，covariance 也由類似 Riccati 型積分來計算 (見上面式子)。因此 X_t 邊際仍是高斯。 csml-cam.github.io
- 這在控制 / 隨機系統理論 (systems-theoretic SDE) 中是標準假

設：很多教材就假設初值高斯，並證明高斯性是保持 (Gaussian closure)。例如在 *Stochastic Differential Equations: A Systems-Theoretic Approach* 就明確用了這個假設。

maxim.ece.illinois.edu

3. Gauss–Markov 過程觀點

- “Gauss–Markov process” 就是指同時具有高斯性 (Gaussian) + Markov (馬可夫) 性質的過程。線性 SDE (與恆係數或 time-varying 係數) 是典型會生成 Gauss–Markov 過程的模型。事實上，Kalman 濾波理論本身就是基於這樣的假設。
- 在機器學習 / 高斯過程 (GP) 文獻中，也經常把這種線性 SDE 作為生成時序 Gauss–Markov prior (如 Krämer & Hennig, JMLR) 的基礎。舉例來說，Krämer 和 Hennig 在 JMLR 論文裡就考慮 time-invariant 線性 SDE ($dx = Fxdt + LdW$)，並假設初始 $x(0) \sim \mathcal{N}(m_0, C_0)$ ，他們證明整條過程是高斯 (Gauss–Markov) 的。 jmlr.org

參考文獻 (References)

- “Statistics of Linear Stochastic Differential (SDE)” — 提出一般線性 SDE 解為高斯過程，並給出 mean 和 covariance 的表達式。 csml-cam.github.io
 - *Stochastic Differential Equations: A Systems-Theoretic Approach* — 控制理論教材，解釋線性 SDE 中若初值高斯，則過程為高斯。 maxim.ece.illinois.edu
 - Krämer, M., & Hennig, P. (2024). “Gauss–Markov priors defined by linear, time-invariant SDEs” (JMLR) — 以線性 SDE 定義高斯–馬可夫先驗。 jmlr.org
- **score matching** 損失中的加權函數 $\lambda(t)$ 理論上應如何選取？不同選擇會如何影響訓練的穩定性與樣本品質？

Ans:

重要結論

- 如果你的目標是**最大化模型的對數似然 (MLE)** 或靠近 **likelihood-based objective**，理論上有一個特定的時間加權（通常稱為 *likelihood weighting*）能把加權的 **score-matching loss** 與 **negative log-likelihood** 建立起上界／等價關係（需搭配一些變形與重要採樣來控制變異）。這在 Song et al. (2021) 與後續最大化似然的工作中有嚴格說明。 [arXiv+1](#)
- 實務上常見、且效果穩定的 heuristic 有： $\lambda(t) = 1$ （均勻）、 $\lambda(t) \propto g(t)^2$ （或 $\lambda(t) = g(t)^2/2$ 等跟 SDE 擴散強度有關的選法），或依照 **log-SNR (signal-to-noise ratio)** 重排／採樣 timesteps。不同選擇會改變各時間尺度上的梯度方差與訓練焦點，進而影響收斂穩定性與生成質量。 [NeurIPS 會議論文+1](#)

1) 理論基礎 — 為什麼要加 $\lambda(t)$

Score-matching 在連續時間（或 denoising score matching 的積分形式）通常寫成

$$\mathcal{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{x_t \sim p_t} [\| s_\theta(x_t, t) - \nabla \log p_t(x_t) \|^2] dt.$$

加權函數 $\lambda(t)$ 的存在，一方面是表示我們可以在不同時間尺度（不同噪聲強度）上賦予不同重要度；另一方面某些特定選擇會令損失與對數似然（或其上界）產生解析關係。Song et al. (2021) 證明：採用某種特定的「**likelihood weighting**」時，這個時間加權的 **score-matching** 損失可以上界 **negative log-likelihood**（並在某些情況下趨近 **tight**）——因此在希望最大似然估計時這個選擇有理論支持；但這也會導致更高的估計方差，需要搭配 **variance-reduction / importance sampling**。 [arXiv+1](#)

重要引文（理論支撐）：

- Song et al., *Score-Based Generative Modeling through SDEs* (ICLR 2021 / arXiv). (likelihood weighting 與 loss-to-NLL 的關係與討論) [arXiv+1](#)

2) 常見 $\lambda(t)$ 的選擇與直覺

下面列出被廣泛採用或討論的幾種做法，並解釋其直覺與優缺點。

1. 均勻加權： $\lambda(t) = 1$

- 直覺：每個時間點一視同仁。實作簡單、梯度方差通常較小（在某些設定下）。常作為 **baseline**。
- 缺點：若某些時段（如極小或極大噪聲）對樣本品質特別重要，均勻會浪費資源或忽略重要尺度。[NeurIPS 會議論文](#)

2. 與擴散強度成比例： $\lambda(t) \propto g(t)^2$ （或 $\lambda(t) = g(t)^2/2$ 等）

- 直覺：把時間尺度的權重與 SDE 中噪聲強度綁在一起。Song et al. 在某些推導中使用與 $g(t)^2$ 有關的 **weighting**（**likelihood weighting** 的具體形式與常數可能依推導不同而略有差異）。此類權重常見於把 **score-matching** 損失連接到模型 **likelihood** 的推導。
- 缺點：可能使某些時間點（例如大 $g(t)$ 的區段）在梯度上佔優勢，導致梯度方差增加，需要 **importance sampling** 或 **variance reduction**。[arXiv+1](#)

3. 基於 **SNR / log-SNR** 的非均勻分佈或採樣策略（等同於改變 $\lambda(t)$ 或替換為非均勻抽樣 **timesteps**）

- 近年多篇工作指出對「中等噪聲」區間多給資源（或採樣密度更高）效果最好；把注意力放在太小（很少噪聲）或太大（純噪聲）處通常回報較差。這可以透過修改 $\lambda(t)$ 或直接用 **importance-sampling / nonuniform timestep sampling** 實現。[ResearchGate+1](#)

4. **likelihood weighting**（理論上的特殊選擇）

- Song et al. 與後續最大似然訓練文獻指出：存在一種 **weighting**，使得加權的 **score-matching loss** 與負對數似然（**ELBO** 或其連續時間對應）有明確關係。採用這個 **weighting**

有助於直接優化或上界 likelihood，但會顯著提高訓練方差，需要額外技巧（如 timesteps 的 importance sampling）來減少梯度噪聲。[NeurIPS 會議論文+1](#)

3) 不同 $\lambda(t)$ 對訓練穩定性與樣本品質的影響（具體化）

- **梯度方差 (train variance)**：某些 weighting（特別是把大量權重放在少數 t ）會把梯度貢獻集中在少數 time-steps，造成高方差、訓練不穩定。這是為什麼實務上會搭配 importance sampling（重抽樣 t 的分佈）或 variance reduction 技術。Nichol & Dhariwal、Improved Denoising Diffusion Probabilistic Models 等工作均討論了重要採樣來降低訓練方差。[Proceedings of Machine Learning Research+1](#)
 - **偏差 vs 變異的 trade-off**：把更多權重放在能夠改進樣本質量的 time range（通常是 SNR 中段）會提升最終樣本視覺質量，但如果太過極端則可能導致模型在其他噪聲尺度上學得很差（生成多樣性或某些細節受損）。Recent noise-schedule / timestep-sampling works 建議把資源集中在中間範圍較有效。[arXiv+1](#)
 - **對 likelihood 的影響**：若你的目標是 log-likelihood（例如用於壓縮、密度估計比對），採用 Song 等人提出的 likelihood-weighting 並結合 variance-reduction，能理論上得到更好或更可解釋的 likelihood 表現；但在視覺質量（FID、IS）上不一定總是最優。[NeurIPS 會議論文+1](#)
-

4) 已有研究（主要、已查證、有用的參考）

（我把最重要的/最相關的列出 — 每個都已核對可存取連結）

1. Yang Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations” (ICLR 2021 / arXiv).

- 關鍵點：介紹連續時間 score-based 模型與 reverse-SDE，並討論「likelihood weighting」使加權 score loss 與 NLL 之間的關係，還提出 variance reduction 的做法。
- 連結（arXiv / PDF）：<https://arxiv.org/abs/2011.13456>。（補充：

該文章的 supplement 有關公式推導)。 [arXiv+1](#)

2. Yang Song et al., “Improved Techniques for Training Score-Based Generative Models” (NeurIPS 2020).

- 關鍵點：實務技巧（包括時間加權、架構與訓練技巧），提供了許多實作面的建議。
- 連結（PDF）：
<https://proceedings.neurips.cc/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf>. [NeurIPS 會議論文](#)

3. Nichol & Dhariwal, “Improved Denoising Diffusion Probabilistic Models” (ICLR 2021 / 2022 pub).

- 關鍵點：提出 hybrid objectives、重要採樣 timesteps 的思想，討論不同訓練目標對方差與 likelihood 的影響。
- 連結：<https://arxiv.org/pdf/2102.09672.pdf>. [Proceedings of Machine Learning Research](#)

4. C. Lu et al., “Maximum Likelihood Training of Score-Based Diffusion Models” (MLR 2022).

- 關鍵點：從最大似然角度研究 score-based 模型，討論時間加權與 likelihood 的對應，以及實務上緩解方差的方法。
- 連結（PDF）：
<https://proceedings.mlr.press/v162/lu22f/lu22f.pdf>. [Proceedings of Machine Learning Research](#)

5. Valentin De Bortoli et al., “Target Score Matching” (2024) — 與 denoising score matching 的限制、改善方法有關。

- 關鍵點：從統計角度分析 DSM，討論不同目標的性質與偏差，對理解 weighting 的效應有啟發。
- 連結（arXiv / PDF）：<https://arxiv.org/pdf/2402.08667.pdf>. [arXiv](#)

6. Noise-schedule / timestep sampling 相關近年工作（2023–2025）：多篇論文（例如 “Improved Noise Schedule for Diffusion Training”

2024/2025，以及 Non-uniform Timestep Sampling 工作）討論以非均勻時間分佈或 $\log\text{-SNR}$ 為基礎重新安排訓練重心，等價於調整 $\lambda(t)$ 或改變抽樣策略。這些實驗性研究支持「把資源放在中間 SNR／中間 t 區段通常更有效」。 [OpenReview+1](#)

5) 實務建議（如果你要訓練一個 **score model**，該怎麼選 $\lambda(t)$ ？）

1. 明確你的目標：

- 若主目標是 **最大化 likelihood / 密度估計** → 優先考慮 Song/Lu 提到的「likelihood weighting」並搭配重要採樣或 variance-reduction（但要準備面對高方差）。 [NeurIPS 會議論文+1](#)
- 若主目標是 **生成質量（視覺）** → 可以採用經驗上穩定的 heuristic（如 $\lambda(t) = 1$ 或 $\lambda(t) \propto g(t)^2$ ），並搭配 noise-schedule 調整或非均勻 timestep sampling。 [NeurIPS 會議論文+1](#)

2. 若採 **likelihood weighting**，務必同時使用 **variance-reduction**

（importance sampling over t 、historical loss-based sampling、或其他）— Song & Lu 等建議如此以控制梯度噪聲。 [NeurIPS 會議論文+1](#)

3. 實驗驗證：在你自己的數據集上做幾個候選 $\lambda(t)$ （例如 1 、 $g(t)^2$ 、 $\log\text{-SNR-centered schedule}$ ），比較（a）train loss 曲線、（b）gradient variance（可估）、（c）sample-quality（FID/IS）與（d）likelihood（若可估）。近期 noise-schedule 論文也顯示在計算資源有限時，把抽樣密度集中在中間噪聲尺度比直接修改 λ 更有利。 [arXiv](#)