统计计算算法题

Ran Liu

2026

部分摘抄于 Givens, G. H., & Hoeting, J. A. (2012). Computational statistics. John Wiley & Sons，以下简称为参考书。

第 1 章　求解非线性方程（Solving Nonlinear Equations）

1. 设下列数据为来自 Cauchy$(\theta, 1)$ 分布的独立同分布样本（i.i.d.）：

   1.77, -0.23, 2.76, 3.80, 3.47, 56.75, -1.34, 4.24, -2.44, 3.29, 3.71, -2.40, 4.53, -0.07, -1.05, -13.87, -2.53, -1.75, 0.27, 43.21

   (a) 绘制对数似然函数图像，并使用牛顿-拉夫森（Newton-Raphson）方法估计 θ 的极大似然估计（MLE）。尝试以下初始值：

   $$-11, -1, 0, 1.5, 4, 4.7, 7, 8, 38$$

   讨论你的结果。数据的均值是否是一个好的初始点？

   (b) 使用二分法（bisection method）并以 -1 和 1 为起始区间，估计 θ。尝试其他初始区间以说明二分法在某些情况下可能无法找到全局极大值。

   (c) 根据公式（2.29）使用不动点迭代（fixed-point iterations），初始值取 -1，比例因子（scaling factor）α 分别设为 1、0.64、0.25。探讨其他初始值和比例因子的选择结果。

   (d) 使用割线法（secant method）从初始点 $(\theta^{(0)}, \theta^{(1)}) = (-2, -1)$ 开始估计 θ。另外，尝试 $(\theta^{(0)}, \theta^{(1)}) = (-3, 3)$ 以及其他初始点组合，观察结果。

   (e) 比较上述牛顿-拉夫森方法、二分法、不动点迭代法和割线法的收敛速度和稳定性。如果将这些方法应用于从 $N(\theta, 1)$ 分布中随机抽取的 20 个样本，结论是否会发生变化？

2. 考虑密度函数

$$f(x) = \frac{1 - \cos(x - \theta)}{2\pi}, \quad 0 \le x \le 2\pi,$$

其中参数 θ 满足 $-\pi \le \theta \le \pi$。下列为从该密度中抽取的独立同分布样本：

3.91, 4.85, 2.28, 4.06, 3.70, 4.04, 5.46, 3.53, 2.28, 1.96, 2.53, 3.88, 2.22, 3.47, 4.82, 2.46, 2.99, 2.54, 0.52, 2.50

我们希望估计参数 θ。

(a) 在区间 $[-\pi, \pi]$ 上绘制对数似然函数。

(b) 求 θ 的矩估计（method-of-moments estimator）。

(c) 使用牛顿-拉夫森（Newton-Raphson）方法求 θ 的极大似然估计，以（b）中矩估计为初始值。若初始值改为 -2.7 和 2.7，对应的结果分别是多少？

(d) 重复（c），将 $[-\pi, \pi]$ 区间划分为 200 个等间距的初始值，观察优化过程的结果。将这些初始值划分为不同“吸引域”（attraction sets），即每组初始值最终收敛到同一个局部极大值。讨论你的发现。

(e) 找出两个尽可能接近的初始值，使得牛顿-拉夫森方法分别收敛到两个不同的解。

3. 在 1974-1999 年间，美国水域中共发生了 46 起原油泄漏事件，每起泄漏不少于 1000 桶，均来自油轮运输。课本网站提供了以下年度数据：

- 第 i 年的泄漏次数 N_i；
- 该年通过美国水域进行进出口运输的原油量 b_{i1}（已调整国际/外海溢出风险）；
- 该年美国本土水域中进行国内运输的原油量 b_{i2}。

数据改编自文献 [11]，整理出来的表格为1。所有油运量均以十亿桶（Bbbl）为单位。

假设原油运输的体积是泄漏风险的暴露量。我们采用如下泊松过程模型：

$$N_i \mid b_{i1}, b_{i2} \sim \text{Poisson}(\lambda_i), \quad \text{其中} \quad \lambda_i = \alpha_1 b_{i1} + \alpha_2 b_{i2}$$

模型参数为 α_1 与 α_2，分别表示每 Bbbl 原油运输单位中泄漏事件发生率（进出口运输与国内运输）。

请完成以下任务：

(a) 推导出使用牛顿-拉夫森（Newton-Raphson）方法求 α_1 和 α_2 极大似然估计的更新公式。

表 1: 1974–1999 年美国水域原油泄漏数据（改编自文献 [11]）

年份	泄漏次数 N_i	进出口运输量 b_{i1}	国内运输量 b_{i2}
1974	4	3.087	1.391
1975	1	2.455	1.207
1976	3	2.703	1.360
1977	4	2.968	1.430
1978	2	2.852	1.456
1979	2	3.052	1.605
1980	2	2.840	1.527
1981	3	2.253	1.344
1982	1	1.738	1.262
1983	1	1.588	1.265
1984	2	1.691	1.290
1985	1	1.748	1.356
1986	0	1.964	1.360
1987	2	2.059	1.412
1988	2	2.188	1.482
1989	2	2.065	1.529
1990	1	1.922	1.549
1991	0	1.941	1.568
1992	2	1.967	1.563
1993	1	2.052	1.517
1994	1	2.161	1.505
1995	0	2.142	1.428
1996	1	2.338	1.428
1997	0	2.392	1.354
1998	1	2.342	1.307
1999	2	2.547	1.305

(b) 推导出使用 Fisher 评分法（Fisher scoring）求 α_1 和 α_2 极大似然估计的更新公式。

(c) 编程实现上述两种方法，计算极大似然估计值，并比较两种方法在实现难度和性能上的差异。

(d) 估计 α_1 和 α_2 的标准误差。

(e) 使用最速上升法（steepest ascent），在必要时进行步长折半（step-halving）回溯（就是一直减半步长，直到上升）。

(f) 使用拟牛顿法（quasi-Newton optimization），Hessian 近似更新公式如参考书式 (2.49) 所示。比较启用和不启用步长折半的性能差异。

(g) 绘图（仿照参考书图 2.8），展示方法 (a)-(f) 所采用的路径，比较它们的收敛轨迹。选择合适的图像区域和起始点以清晰展示算法性能的主要差异。

4. 上述问题都尝试用 pytorch 中的自动梯度下降求解，优化器可选 adam。

第 2 章 Expectation-Maximization 算法

1. 验证高斯混合模型（Mixture of Gaussians）中 EM 算法的参数更新公式。

(a) 设定混合模型参数。

(b) 使用上述参数从混合模型中生成 n 个观测数据点。

(c) 不使用真实标签，使用 EM 算法对模型参数进行估计，记录每轮迭代的 Q 函数值。

(d) 绘制 Q 函数随迭代次数变化的轨迹图，并说明是否单调增加，是否收敛。

(e) 比较估计出的参数与真实参数之间的差距，并进行分析：哪些参数估计得更准确？哪些不准确？可能原因是什么？

(f) 多次重复实验（不同设定，初始值，seed 或样本量），观察误差随混合分布中各个部分分布的距离的变化。

(g) 尝试 pytorch 计算梯度，使用 Gradient based 的 MCEM 算法更新参数。

2. 使用 EM 算法对 B 站用户等级数据拟合零膨胀泊松混合模型（Zero-Inflated Poisson Mixture Model, ZIPMM）。假设我们观测到 n 个 B 站用户的等级分布，共有 7 个等级，等级为 $i = 0, 1, \ldots, 6$ 的用户人数为 n_i，则总人数为 $n = \sum_{i=0}^{6} n_i$。

设定模型如下：

- 未注册用户以概率 ξ 出现，其等级恒为 0；
- 注册用户以概率 $1 - \xi$ 出现，其等级服从参数为 λ 的泊松分布；

表 2: 等级分布观测频数（记号定义）

用户等级 i	0	1	2	3	4	5	6
人数 n_i	n_0	n_1	n_2	n_3	n_4	n_5	n_6

- 目标是对参数 $\theta = (\xi, \lambda)$ 进行极大似然估计。

(a) 写出该模型的完全数据似然函数，并推导 EM 算法中 E 步和 M 步的更新公式。

(b) 给定下列 B 站用户等级观测数据（来自网络估计，单位：万人），估计参数 ξ 和 λ：

表 3: B 站用户等级观测数据（单位：万人）

用户等级	0	1	2	3	4	5	6
人数（万）	59900	1644	7376	2778	2082	2211	554

(c) 编程实现 EM 算法，记录每次迭代的 Q 函数值，绘制 Q 函数轨迹图，并展示参数 ξ 和 λ 的更新轨迹图。

(d) 汇报最终估计的参数值，并解释这些结果的含义：B 站中大约有多少比例的用户可能是“未注册”状态？注册用户的平均等级是多少？

(e) 尝试不同的初始值，观察算法是否收敛到同一组参数（讨论局部最优问题）。

(f) 将该模型与标准泊松模型（不考虑 ξ）进行对比，比较拟合优度（如 AIC/BIC）并解释为何零膨胀模型更合适。

3. 使用隐马尔可夫模型（Hidden Markov Model, HMM）对 DNA 生物序列进行建模与注释。在 DNA 序列中，由四种碱基 $\{A, C, G, T\}$ 构成的核苷酸链以特定顺序排列，携带着生物遗传信息。其中，G 表示鸟嘌呤（Guanine），C 表示胞嘧啶（Cytosine）。GC 含量是指 DNA 序列中 G 和 C 的比例，它是基因组结构的重要统计特征之一。

   GC-rich 区域（高 GC 含量区域）是指 G 和 C 所占比例显著高于平均水平的 DNA 片段。这些区域通常具有重要的生物功能。例如：

   - GC 配对之间具有三个氢键（相比 AT 配对的两个氢键），使得 GC-rich 区域的 DNA 结构更加稳定；
   - GC-rich 区域常出现在基因的启动子区域（promoters）和 **CpG islands** 中，后者是哺乳动物基因组中常见的调控区域；
   - 这些区域往往参与基因的表达调控，与 DNA 甲基化、转录活性等过程密切相关。

由于 GC-rich 区域和普通区域在碱基组成上存在统计差异（如 GC 含量高 vs 低），我们可以将其建模为由不同“隐藏状态”生成的序列问题。隐马尔可夫模型（Hidden Markov Model, HMM）是一种适合处理这类序列结构的概率图模型。它假设观测序列由一组不可见的隐藏状态驱动生成，不同状态具有不同的碱基发射概率，并通过状态转移矩阵描述状态之间的转换关系。因此，HMM 可用于根据观测的 DNA 碱基序列自动识别哪些区域属于 GC-rich 区域，哪些属于非 GC-rich 区域。

考虑一个简化问题：DNA 序列由四种碱基字母 $\{A, C, G, T\}$ 构成，隐藏状态为 $\{H, L\}$，分别表示高 GC 含量区域和低 GC 含量区域。我们假设序列是由一个两状态 HMM 生成的，状态转移矩阵为 $A = (a_{ij})$，碱基发射概率矩阵为 $B = (b_j(o))$，初始状态分布为 π。

下面给出一个观测到的 DNA 子序列（长度为 60）作为训练样本：

表 4: 观测到的 DNA 碱基序列

GGCACTGAACTGACACGTAGGACGTACGTAGCTAGCTAGACGTAGTCGTAGTACGTAG

(a) 建立一个两状态（H 和 L）的隐马尔可夫模型，对观测碱基序列进行建模。设初始发射概率如下：

表 5: 初始发射概率矩阵 B

状态	A	C	G	T
H	0.2	0.3	0.3	0.2
L	0.3	0.2	0.2	0.3

状态转移矩阵和初始状态分布可设为均匀初始化。

(b) 使用 Baum–Welch 算法估计 HMM 的参数。写出 E 步和 M 步的更新公式，并说明每一步的含义。记录每轮迭代的对数似然值，并绘制收敛轨迹图。

(c) 使用 Viterbi 算法对观测序列进行解码，输出最可能的隐藏状态序列（例如：HHLHHLL...），并与 GC 含量变化进行对比分析。

(d) 将训练后的模型用于生成新的 DNA 序列和状态序列，展示模型生成的结果，并分析其统计特征是否与原序列相似。

4. 阅读序列相互作用模型文献 (图 1)：

Liu, R., Tang, X., & Fan, X. (2025). Sequence interaction model with applications to TCR-peptide binding. The Annals of Applied Statistics, 19(4), 2683-2704.

(a) 如何用 HMM 建模一对序列而不是单个序列？

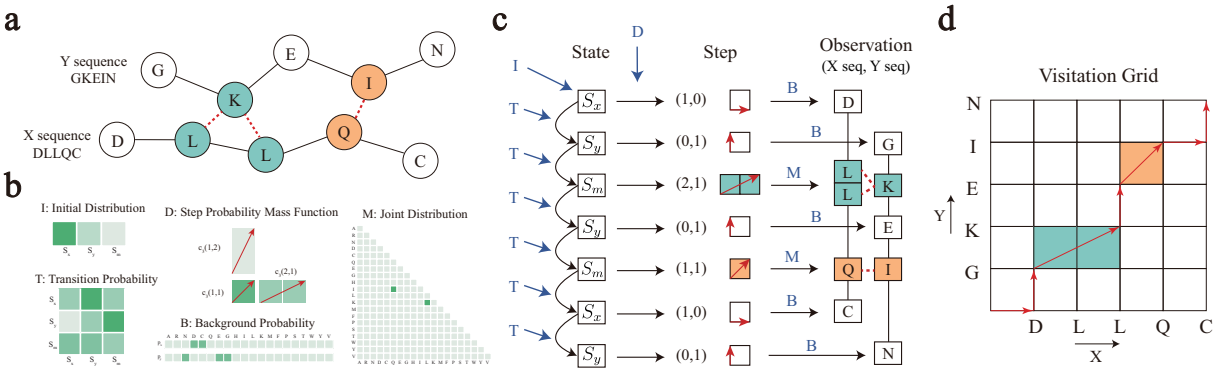

图 1: Sequence interaction model.

(b) 自己设置参数 I, T, D, B, M 产生多条成对的序列。

(c) 试着提出一种参数估计的方法，通过这些生成的成对序列估计参数，与真实的相比较。

(d) 扩展到多个序列的情况。扩展到连续的情况。扩展到别的应用领域中。

第 3 章 Monte Carlo Method

1. 使用 Monte Carlo 方法对一个高维积分进行估计，并通过重要性采样降低估计方差。

   考虑如下在 $[0,1]^5$ 区间内的五维定积分：

$$I = \int_{[0,1]^5} \frac{1}{1 + \left(\sum_{i=1}^{5} x_i\right)^2} dx_1 dx_2 dx_3 dx_4 dx_5$$

   该积分没有简单解析解，但可以使用 Monte Carlo 方法进行近似。

   (a) 使用基本的均匀采样法，在 $[0,1]^5$ 区间内生成 N 个样本，估计该积分的值。分别使用 $N = 10^3, 10^4, 10^5$，记录每次估计值及其方差。

   (b) 尝试改进估计精度。注意到被积函数在 $\sum x_i$ 接近 0 时变化较快，可以考虑使用重要性采样。设重要性分布为联合独立的 Beta 分布：

$$q(x_1, x_2, \ldots, x_5) = \prod_{i=1}^{5} \mathrm{Beta}(x_i; \alpha, \beta)$$

   其中 $\mathrm{Beta}(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$。

   选择 $\alpha = 0.5, \beta = 1.5$，使得样本更多集中于较小的 x_i，尝试使用该分布进行重要性采样估计 I 的值。写出重要性采样估计公式，并分析其方差是否变小。

   (c) 对比两种方法在 $N = 10^4$ 时的结果，包括估计值、估计标准差、收敛速度等，并绘图说明。

(d) 尝试不同的 α, β 组合。

2. 探索逆变换采样跟流形变换的关系。

设 Φ^{-1} 为标准正态分布的一维逆累积分布函数。考虑在 $[0,1]^2$ 中等间距地取点，并分别对每个点的每个分量应用标准正态逆 CDF 变换，得到：

$$\mathbf{z}_i = (\Phi^{-1}(u_{i1}), \Phi^{-1}(u_{i2})), \quad i = 1, 2, 3$$

(a) 写出 Φ^{-1} 的表达式或图像特性，并解释其在 $u = 0.5$ 附近缓变、在 $u \to 0$ 或 $u \to 1$ 时急变的性质。

(b) 选择一条直线上的三个点 $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$，计算 $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ 的坐标，并绘制在 $\mathbb{R}^2$ 中的散点图。连接三点，观察它们是否共线。

(c) 计算向量差 $\mathbf{z}_2 - \mathbf{z}_1$ 与 $\mathbf{z}_3 - \mathbf{z}_2$，判断是否相等，并解释为何不满足“等向量增量”性质。判断三点是否落在某个圆弧上？

3. 使用接受-拒绝采样（Accept-Reject Sampling）从一个复杂分布中生成样本，并比较不同 proposal 分布的接受率。

考虑目标分布 $p(x)$，定义如下：

$$p(x) = \frac{1}{Z} \cdot f(x) = \frac{1}{Z} \cdot \exp\left(-x^4 + 3x^2\right), \quad x \in \mathbb{R}$$

其中 Z 是归一化常数，无法直接计算。我们希望从该分布中采样。

(a) 绘制 $f(x) = \exp(-x^4 + 3x^2)$ 的图像，观察其形状。你会发现该分布具有两个对称的模态（peaks），类似于双峰分布。

(b) 尝试使用以下三种 proposal 分布 $q(x)$ 进行接受-拒绝采样：

(i) $q_1(x) = \mathcal{N}(0, 1)$，标准正态分布；

(ii) $q_2(x) = \mathcal{N}(0, 2^2)$，宽尾正态分布；

(iii) $q_3(x)$ 是双峰 proposal：$q_3(x) = \frac{1}{2}\mathcal{N}(-1.5, 1^2) + \frac{1}{2}\mathcal{N}(1.5, 1^2)$。

对每个 proposal 分布，执行如下步骤：

- 估计最小可行的常数 M，使得 $f(x) \le Mq(x)$ 对所有 x 成立；
- 实现接受-拒绝采样算法，生成 $N = 10^4$ 个样本；
- 记录每种方法的接受率（即采样成功的比例）；
- 绘制每种方法的样本直方图，并与 $f(x)$ 的形状进行比较。

(c) 比较三种 proposal 分布的接受率，并分析哪个 proposal 更适合该目标分布。解释原因。

(d) 如果目标分布是高维的，接受-拒绝采样是否仍然高效？它的主要瓶颈在哪里？

4. 使用 Monte Carlo 方法估算欧式看涨期权（European Call Option）的价格，并通过方差控制方法提高估计精度。假设标的资产价格 S_t 满足 Black–Scholes 模型中的几何布朗运动：

$$dS_t = rS_t\,dt + \sigma S_t\,dW_t, \quad t \in [0, T]$$

其中：

- 初始价格：$S_0 = 100$
- 无风险利率：$r = 5\%$
- 波动率：$\sigma = 20\%$
- 到期时间：$T = 1$ 年
- 执行价：$K = 100$

欧式看涨期权的理论价格为：

$$C = e^{-rT} \cdot \mathbb{E}[\max(S_T - K, 0)]$$

(a) 使用基本的 Monte Carlo 方法估算该期权价格。即，生成 N 个样本 $Z_i \sim \mathcal{N}(0, 1)$，计算：

$$S_T^{(i)} = S_0 \cdot \exp\left((r - \tfrac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z_i\right)$$

然后估计：

$$\hat{C}_N = e^{-rT} \cdot \frac{1}{N}\sum_{i=1}^{N} \max(S_T^{(i)} - K, 0)$$

分别使用 $N = 10^4, 10^5, 10^6$，记录估计价格及其样本标准差。

(b) 使用对偶随机变量（Antithetic Variates）方法进行方差控制：
对每个 $Z_i \sim \mathcal{N}(0, 1)$，同时使用它的相反数 $-Z_i$，生成一对路径：

$$S_T^{(i,+)} = S_0 \cdot \exp\left((r - \tfrac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z_i\right), \quad S_T^{(i,-)} = S_0 \cdot \exp\left((r - \tfrac{1}{2}\sigma^2)T - \sigma\sqrt{T}Z_i\right)$$

对每对样本取平均 payoff：

$$Y_i = \frac{1}{2}\left[\max(S_T^{(i,+)} - K, 0) + \max(S_T^{(i,-)} - K, 0)\right]$$

然后估计期权价格为：

$$\hat{C}_{\text{anti}} = e^{-rT} \cdot \frac{1}{N} \sum_{i=1}^{N} Y_i$$

与普通 Monte Carlo 方法比较估计值与方差，说明是否有方差降低。

(c) 使用控制变量（Control Variate）方法进一步降低方差：

注意到 S_T 的期望在 Black-Scholes 模型下为：

$$\mathbb{E}[S_T] = S_0 e^{rT}$$

构造控制变量估计器如下：

$$\tilde{C}_N = e^{-rT} \cdot \left[\frac{1}{N} \sum_{i=1}^{N} \max(S_T^{(i)} - K, 0) - \beta \left(\frac{1}{N} \sum_{i=1}^{N} S_T^{(i)} - S_0 e^{rT} \right) \right]$$

其中 β 为一个系数，理论上可以通过最小方差计算得出，但此处你可以尝试 $\beta = 1$ 或其他值。比较使用控制变量前后的估计方差，并解释控制变量为何能够降低方差。

(d) 尝试将期权定价公式应用到实际场景。（如上证 50ETF 期权）

第 4 章　MCMC

1. 使用 MCMC 方法从一个复杂的二维分布中采样，分别使用随机游走 Metropolis 算法（Random Walk Metropolis, RWM）、Langevin Monte Carlo（LMC）、Hamiltonian Monte Carlo（HMC），并比较三种采样方式的表现。考虑如下目标分布，其未归一化密度为：

$$\pi(x, y) \propto \exp\left(-\frac{1}{2} \left(\frac{x^2}{100} + (y - 5\sin(x))^2 \right) \right)$$

这是一个具有非线性相关结构的二维分布，形状大致是沿着 $y = 5\sin(x)$ 的长条形曲线。

(a) 绘制该分布的等高线图（contour plot）或热力图（heatmap），帮助理解其结构。

(b) 使用以下三种方法从该分布中采样：

(i) **随机游走 Metropolis（RWM）**：以 $\mathcal{N}(0, \sigma^2 I)$ 为 proposal，尝试不同的 σ（例如 $\sigma = 0.1, 0.5, 1.0$）观察接受率和采样效率；

(ii) **Langevin Monte Carlo（LMC）：**利用梯度信息，使用 Euler–Maruyama 近似进行更新：

$$x_{t+1} = x_t + \frac{\epsilon^2}{2}\nabla \log \pi(x_t) + \epsilon Z_t, \quad Z_t \sim \mathcal{N}(0, I)$$

并使用 Metropolis 修正步骤 (MALA)；

(iii) **Hamiltonian Monte Carlo（HMC）：**引入动量变量，设定步长 ϵ 和步数 L，模拟 Hamilton 动力学轨迹，使用 Metropolis 接受步骤。(NUTS?)

(c) 对每种采样方法，生成 $N = 10^4$ 个样本，记录以下指标：

- 接受率；
- 样本自相关（autocorrelation）；
- 有效样本大小（effective sample size, ESS）；
- 可视化样本轨迹和散点图；
- 估计 $\mathbb{E}[x]$ 和 $\mathbb{E}[y]$。

(d) 比较三种方法在上述指标下的表现，总结它们在处理此类非线性分布时的优劣。

2. 设我们要从以下三维联合分布 $\pi(x, y, z)$ 中采样：

$$\pi(x, y, z) \propto \exp\left(-\frac{(x-y)^2}{2} - \frac{(y-z)^2}{2} - \frac{\lambda}{2}z^4\right), \quad x, y, z \in \mathbb{R}$$

其中 $\lambda > 0$ 是已知常数（例如 $\lambda = 0.1$）。该分布具有链式结构：$x \leftrightarrow y \leftrightarrow z$，但包含一个非高斯的 z^4 项，导致联合分布不可直接采样。

(a) 写出每个变量的条件分布形式：

- $\pi(x|y, z)$；
- $\pi(y|x, z)$；
- $\pi(z|x, y)$；

指出哪些是标准分布（如高斯），哪些需要使用 Metropolis-Hastings 采样。

(b) 构造一个 Gibbs 采样器，其中：

- 对于 x 和 y，使用条件高斯分布直接采样；
- 对于 z，使用 Metropolis-within-Gibbs：给定当前 x, y，从 proposal $z' \sim \mathcal{N}(z^{(t)}, \sigma^2)$ 采样，并进行 Metropolis 接受-拒绝步骤。(不需要接受才往前走)

(c) 实现该采样器，采样 $T = 10^4$ 个样本，绘制三维样本散点图，并估计：

$$\mathbb{E}[x], \quad \mathbb{E}[y], \quad \mathbb{E}[z], \quad \mathrm{Cov}(x, z)$$

(d) 分析接受率随 σ 的变化（例如尝试 $\sigma = 0.2, 1.0, 2.0$），讨论采样效率的变化。

(e) 如果将 z^4 替换为 $|z|$ 或 z^6，采样器是否还能工作？你会如何修改采样方法？

3. 阅读保守串采样的文献 (图 **??**):

Tang, X., & Liu, R. (2026). GAMMA: Gap-aware motif mining under incomplete labeling with applications to MHC motifs. Bioinformatics, 42(2), btag014.

Tang, X., & Liu, R. (2025). De-motif sampling: an approach to decompose hierarchical motifs with applications in T cell recognition. Briefings in Bioinformatics, 26(3), bbaf221.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. Journal of the American Statistical Association, 89(427), 958-966.

(a) 从自己设定的参数中生成一些结合序列和非结合序列。

(b) 推导各参数和隐变量的更新公式，构建算法从基序的后验分布采样 (MCMC)。

(c) 尝试推导出 collapsed Gibbs 版本的 predictive distribution，并用算法验证。

(d) 提出 MH 算法跳出 local mode，并验证它的效果。

(e) 如果结合位置不是连续的会怎么样？如果有一些 label 存在 false positive 或 false negative 怎么办，如何纳入模型中？

第 5 章 Variational Inference

1. 本题中你将构建一个贝叶斯逻辑回归模型，并使用变分推断（Variational Inference）估计后验分布。你首先将手工实现均值场变分推断（Mean-field VI），然后用 Pyro 框架复现并比较效果。

我们考虑如下模型：

$$
\begin{aligned}
&w \sim \mathcal{N}(0, I_d) \quad \text{(prior)} \\
&y_i \sim \text{Bernoulli}(\sigma(x_i^\top w)), \quad i = 1, \ldots, N
\end{aligned}
$$

其中 $\sigma(z) = \frac{1}{1+e^{-z}}$ 是 sigmoid 函数。

- $x_i \in \mathbb{R}^d$ 是输入特征；
- $y_i \in \{0, 1\}$ 是二分类标签；
- $w \in \mathbb{R}^d$ 是权重向量；

(a) 生成一个二分类合成数据集（例如 $d = 2$），使得两类数据可分但有一定重叠。可设真实参数为 $w^* = [2.0, -1.0]$。

(b) 手工实现变分推断：

- 使用均值场高斯分布 $q(w) = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$；
- 构造 ELBO：

$$\mathrm{ELBO}(\mu, \sigma) = \mathbb{E}_{q(w)}[\log p(y|X, w)] - \mathrm{KL}(q(w)\|p(w))$$

- 使用 reparameterization trick（$w = \mu + \sigma \cdot \epsilon, \epsilon \sim \mathcal{N}(0, I)$）；(比较不使用 reparameterization trick，估计的方差变化)
- 使用 PyTorch 优化 $\mu, \log \sigma$ (也可以自己手算梯度)；
- 绘制训练过程中的 ELBO 曲线；

(c) 使用 Pyro 框架重新实现：

- 定义 'model()'：prior + likelihood；
- 定义 'guide()'：均值场变分分布；
- 使用 'SVI' + 'Trace_ELBO' 进行优化；
- 输出最终变分参数 μ, σ；

(d) 对比手工实现与 Pyro 实现的结果：

- 画出后验样本在 (w_1, w_2) 空间的散点图；
- 用后验均值 $w = \mu$ 做预测，画出分类边界；
- 计算测试集的准确率；

2. VAE 和 DDPM 在手写字体上的训练及条件生成。

第 6 章 Bootstrap & Permutation Test

1. 本题中你将使用非参数 Bootstrap 方法估计一个统计量的标准误差与置信区间，并与理论公式进行比较。

   设有一组来自未知分布的独立样本 $X = \{x_1, x_2, \ldots, x_n\}$，我们关心的统计量是样本的中位数 $\hat{\theta} = \mathrm{median}(X)$。中位数的分布在有限样本下很难解析，因此我们使用 Bootstrap 方法近似其分布。

   (a) 从标准正态分布 $N(0, 1)$ 中生成样本 X，大小为 $n = 100$，并计算其中位数 $\hat{\theta}$。

   (b) 使用 Bootstrap 方法重复如下步骤 $B = 1000$ 次：

   - 从 X 中有放回地采样出大小为 n 的样本 X^*；

- 计算 X^* 的中位数 $\hat{\theta}^*$；

得到 Bootstrap 分布 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\}$。

(c) 使用 Bootstrap 分布估计中位数的：

- 标准误差（standard error）；
- 95% 置信区间（使用百分位法：取第 2.5 与 97.5 百分位）；
- 绘制 Bootstrap 分布直方图，并标记 $\hat{\theta}$ 与置信区间边界。

(d) 将上述 Bootstrap 标准误差与理论标准误差 $\mathrm{SE}_{\mathrm{normal}} = \frac{1}{\sqrt{4nf(\theta)^2}}$ 进行比较，其中 $f(\theta)$ 是密度函数在 θ 处的值（对于正态分布 $f(0) = \frac{1}{\sqrt{2\pi}}$）。

2. 本题中你将使用 Permutation Test（置换检验）来判断两个独立样本的均值是否存在显著差异。该方法不依赖于数据的分布假设，适用于非正态或样本量较小的情况。

设有两个独立样本：

$$X = \{x_1, x_2, \ldots, x_m\}, \quad Y = \{y_1, y_2, \ldots, y_n\}$$

我们想检验如下原假设：

$$H_0 : \mu_X = \mu_Y \quad \text{vs.} \quad H_1 : \mu_X \neq \mu_Y$$

(a) 生成两个样本：

- X：从 $\mathcal{N}(0, 1)$ 中采样 $m = 30$ 个样本；
- Y：从 $\mathcal{N}(0.5, 1)$ 中采样 $n = 35$ 个样本；

并计算观察到的统计量：

$$T_{\mathrm{obs}} = \bar{x} - \bar{y}$$

(b) 构造置换检验过程：

- 将 X 和 Y 合并为一个整体样本 Z；
- 重复 $B = 1000$ 次以下操作：
  i. 随机打乱 Z；
  ii. 将前 m 个元素作为 X^*，后 n 个元素作为 Y^*；
  iii. 计算置换统计量 $T^* = \bar{x}^* - \bar{y}^*$；
- 得到置换分布 $\{T_1^*, \ldots, T_B^*\}$。

(c) 计算双尾 p 值：

$$p = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}\left(|T_i^*| \geq |T_{\mathrm{obs}}|\right)$$

(d) 绘制置换分布直方图，标出 T_{obs} 和临界区域；说明你是否拒绝 H_0（显著性水平 $\alpha = 0.05$）。

(e) 与经典 t 检验进行比较：

- 使用 'scipy.stats.ttest_ind(X, Y, equal_var=False)'；
- 比较置换检验与 t 检验的 p 值是否一致；

(f) 若 $H_1 : \mu_X > \mu_Y$ 呢？