8.1 令 $D=\{x_1,x_2,\ldots,x_n\}$ 为从指数分布中 i.i.d. 采样得到的样本, 其 p.d.f. 为

$$p(x)=\lambda\exp(-\lambda x)[\![x\geq 0]\!]=\begin{cases}\lambda\exp(-\lambda x) & \text{if } x\geq 0\\ 0 & \text{if } x<0\end{cases}, \tag{8.42}$$

其中 $\lambda>0$ 是一个参数，$[\![\cdot]\!]$ 是指示函数。找出 λ 的最大似然估计。

8.6 在一个二分类问题中，令两个类条件分布为 $p(x|y=i)=N(\mu_i,\Sigma), i\in\{1,2\}$。也就是说，这两个类都是高斯，并且共享相同的协方差矩阵。令 $Pr(y=1)=Pr(y=2)=0.5$，并使用 0-1 损失函数。那么预测结果由公式 $y^\star=\underset{1\leq i\leq m}{\operatorname{argmax}}\, p(y=i\mid x;\theta)$ 给出。证明该预测规则可被重写为如下的等价形式：

$$y^*=\begin{cases}1 & \text{if } w^Tx+b>0\\ 2 & \text{if } w^Tx+b\leq 0\end{cases} \tag{8.45}$$

基于 μ_1,μ_2 和 Σ 给出 w 与 b 的表达式。

9.1 主成分分析 (PCA) 使用

$$y=E_d^T(x-\bar{x}),$$

将一个向量 $x\in\mathbb{R}^D$ 变换成一个低维向量 $y\in\mathbb{R}^d(d<D)$，其中 $\bar{x}$ 是 x 的样本均值，E_d 是一个由样本 x 的协方差矩阵的前 d 个特征向量构成的 $D\times d$ 矩阵（见第 5 章）。令 x_1 和 x_2 为 x 的任意两个样本，y_1 和 y_2 是与它们对应的 PCA 变换之后的结果。证明

$$d_A^2(x_1,x_2)=\|y_1-y_2\|_2^2$$

是由公式 $d_A^2(x,y)=(x-y)^TA(x-y)$ 定义的距离度量家族中的合法成员。为此我们需要将什么值赋给矩阵 A?

9.3 在本题中，我们将证明 $\|x\|_p(p>0)$ 是关于 p 的非递增函数。换言之，若 $0<p<q$，证明

$$(|x_1|^p+|x_2|^p+\cdots+|x_d|^p)^{\frac{1}{p}}\geq(|x_1|^q+|x_2|^q+\cdots+|x_d|^q)^{\frac{1}{q}}. \tag{9.56}$$

(a) 证明公式 (9.56) 在额外的约束条件 $x_i \geq 0, 1 \leq i \leq d$ 下等价于

$$(x_1^p + x_2^p + \cdots + x_d^p)^{\frac{1}{p}} \geq (x_1^q + x_2^q + \cdots + x_d^q)^{\frac{1}{q}}. \tag{9.57}$$

(b) 记 $r = \frac{q}{p}(0 < p < q)$ 并假设对任意 $1 \leq i \leq d, x_i \geq 0$。证明公式 (9.57) 等价于

$$(y_1 + y_2 + \cdots + y_d)^r \geq (y_1^r + y_2^r + \cdots + y_d^r), \tag{9.58}$$

其中 $y_i = x_i^p$。
(c) 证明当 $r > 1$ 和 $y_i \geq 0 (i = 1, 2, \ldots, d)$ 时，公式 (9.58) 成立。(提示：当 $d = 2$ 时用泰勒展开，当 $d > 2$ 时用数学归纳法。)

9.4 证明当 $G \in \mathbb{R}^d \times \mathbb{R}^d$ 是正定矩阵时，$\|Gx\|$ $(x \in \mathbb{R}^d)$ 是一个有效的向量范数。

11.2 (ISTA) ISTA, 或迭代的收缩-阈值算法 (Iterative Shrinkage-Thresholding Algorithms) 是一系列方法，当字典 D 和 x 已知时，可以解如下问题：

$$\arg\min_{\alpha} \|x - D\alpha\|^2 + \lambda\|\alpha\|_1.$$

令 $f(\alpha)$ 和 $g(\alpha)$ 是 α 的两个函数，其中 f 是一个光滑的凸函数，g 是一个连续的凸函数。然而，g 不一定是光滑的，这使得优化 $\min_\alpha f(\alpha) + g(\alpha)$ 很困难。
ISTA 迭代地求解 $\min_\alpha f(\alpha) + g(\alpha)$，每步迭代是一个收缩-阈值步骤。因此，它被命名为迭代的收缩-阈值算法 (ISTA)。在本题中，我们只考虑 $f(\alpha) = \|x - D\alpha\|^2$ 和 $g(\alpha) = \lambda\|\alpha\|_1 (\lambda > 0)$ 的简单情况。
(a) ISTA 的一个额外约束是 f 是连续可微的，并且其梯度满足常数为 $L(f)$ 的李普希茨连续条件 (Lipschitz continuous)，即存在一个依赖于 f 的常数 $L(f)$，使得对任意 α_1 和 α_2 有

$$\|\nabla f(\alpha_1) - \nabla f(\alpha_2)\| \leq L(f)\|\alpha_1 - \alpha_2\|, \tag{11.27}$$

其中 ∇f 是 f 的梯度。
对我们选择的 f，证明其对应的 $L(f)$ (或者简写为 L) 是 $D^T D$ 最大特征值的两倍，$L(f)$ 被称为 f 的李普希茨常数 (Lipschitz constant)。
(b) ISTA 首先初始化 α (例如通过忽略 $g(\alpha)$ 并求解普通最小二乘回归问

题)。然后在每步迭代中求解如下的问题:

$$p_L(\beta) \stackrel{\text{def}}{=} \arg\min_{\alpha} g(\alpha) + \frac{L}{2} \left\| \alpha - \left(\beta - \frac{1}{L} \nabla f(\beta) \right) \right\|^2, \tag{11.28}$$

其中 L 是李普希茨常数，β 是一个参数。在第 t 次迭代时，ISTA 通过如下公式更新解

$$\alpha_{t+1} = p_L(\alpha_t).$$

对我们选择的 f 和 g，解这个优化问题。解释为什么 ISTA 迭代时每步会导致稀疏性。

12.1 假设一个 DTMC 对一个离散并且有 N 个可能取值（状态）的随机变量 X 的演化进行建模。证明我们需要 $N^2 - 1$ 个数来完全描述这个 DTMC。

12.2 令 A 为一个 HMM 模型的转移矩阵。证明对任意正整数 $k, A^k = \underbrace{A \dots A}_{kA's}$ 是一个右随机矩阵。