作业二答案

8.1

解： 对数似然函数为

$$\begin{aligned}\ell\ell(\lambda) &= \ln \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \sum_{i=1}^{n} (\ln \lambda - \lambda x_i) \\ &= n \ln \lambda - n\lambda \bar{x},\end{aligned}$$

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 是样本均值，由于 x_i 是来自指数分布的样本，我们隐式地假设 $[\![x_i \geq 0]\!] = 1 \quad (1 \leq i \leq n)$

由于

$$\frac{\partial \ell\ell}{\partial \lambda} = \frac{n}{\lambda} - n\bar{x} \quad \text{以及} \quad \frac{\partial^2 \ell\ell}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0,$$

$\ell\ell(\theta)$ 是一个凹函数，其局部最大值也是全局最大值。令 $\frac{\partial \ell\ell}{\partial \lambda} = 0$，我们可以通过 $\lambda = \frac{1}{\bar{x}}$ 找到其局部最大值。

因此，最大似然估计是样本均值的倒数。

8.6

解： 设判别函数分别为 g_1 和 g_2，对应于类别 1 和类别 2。我们有

$$g_i(\boldsymbol{x}) = -\frac{D}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln 0.5.$$

现在我们计算 $g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$，这等于

$$\begin{aligned}&-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \\ =& \left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right) \\ &- \left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) \\ =& (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \frac{1}{2}\left(\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right).\end{aligned}$$

如果 $g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) > 0$，则该分类规则会预测为类别 1，否则为类别 2。

因此，我们令 $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$，以及 $b = \frac{1}{2}\left(\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right)$，

则该分类规则为

$$y^* = \begin{cases} 1 & \text{if } \boldsymbol{w}^T\boldsymbol{x} + b > 0 \\ 2 & \text{if } \boldsymbol{w}^T\boldsymbol{x} + b \leq 0 \end{cases}$$

9.1

解:

$$
\begin{aligned}
d_A^2(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2^2 \\
&= (\boldsymbol{y}_1 - \boldsymbol{y}_2)^T (\boldsymbol{y}_1 - \boldsymbol{y}_2) \\
&= \left(\boldsymbol{E}_d^T (\boldsymbol{x}_1 - \boldsymbol{x}_2)\right)^T \left(\boldsymbol{E}_d^T (\boldsymbol{x}_1 - \boldsymbol{x}_2)\right) \\
&= (\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \boldsymbol{E}_d \boldsymbol{E}_d^T (\boldsymbol{x}_1 - \boldsymbol{x}_2).
\end{aligned}
$$

因此，这是一个有效的距离度量。我们令 $\boldsymbol{A} = \boldsymbol{E}_d \boldsymbol{E}_d^T$，其显然是一个对称半正定矩阵。

9.3

(a). 这是平凡的，因为当 $x_i \geq 0$ 时 $|x_i| = x_i$。实际上假设 $x_i > 0$ 也可以。

(b). 取 "$\geq$" 两端的 q 次幂，并注意到 $x_i^p = y_i, x_i^q = (x_i^p)^{q/p} = y_i^r$。

(c). 不失一般性，我们假设 $y_1 \geq y_2 \geq 0$，以及记 $f(x) = x^r$ $(r > 1)$。

那么，

$$
f(y_1 + y_2) = f(y_1) + f'(y_1) y_2 + \frac{f''(\theta)}{2} y_2^2,
$$

其中 $\theta \in (y_1, y_1 + y_2)$。

由于

$$
\frac{f''(\theta)}{2} y_2^2 = \frac{r(r-1)}{2} \theta^{r-2} y_2^2 \geq 0 \quad \text{及} \quad f'(y_1) y_2 = r y_1^{r-1} y_2 \geq r y_2^r \geq y_2^r,
$$

故 $(y_1 + y_2)^r \geq y_1^r + y_2^r$。使用数学归纳法向 $d > 2$ 扩展是平凡的。

9.4

解: 我们需要验证向量范数的三个性质。

(a) 显然，对任意 $c \in \mathbb{R}$，$\|cG\boldsymbol{x}\| = |c| \|G\boldsymbol{x}\|$。

(b) $\|G\boldsymbol{x}\| = 0$ 意味着 $\|G\boldsymbol{x}\|^2 = \boldsymbol{x}^T G^T G \boldsymbol{x} = 0$。我们记 $A = G^T G$。由于 $G \succ 0$，我们知道 $A \succ 0$，即当 $\boldsymbol{x} \neq 0$ 时 $\boldsymbol{x}^T A \boldsymbol{x} > 0$。因此，当 $\|G\boldsymbol{x}\| = 0$，我们必须有 $\boldsymbol{x} = 0$。

(c) 我们需要证明 $\|G(\boldsymbol{x} + \boldsymbol{y})\| \leq \|G\boldsymbol{x}\| + \|G\boldsymbol{y}\|$，这等价于

$$
\|G(\boldsymbol{x} + \boldsymbol{y})\|^2 \leq (\|G\boldsymbol{x}\| + \|G\boldsymbol{y}\|)^2.
$$

展开该式，我们得到：

$$
(\boldsymbol{x} + \boldsymbol{y})^T G^T G (\boldsymbol{x} + \boldsymbol{y}) \leq \boldsymbol{x}^T G^T G \boldsymbol{x} + \boldsymbol{y}^T G^T G \boldsymbol{y} + 2\sqrt{\boldsymbol{x}^T G^T G \boldsymbol{x} \boldsymbol{y}^T G^T G \boldsymbol{y}}
$$

这等价于：

$$
\boldsymbol{x}^T G^T G \boldsymbol{y} \leq \sqrt{\boldsymbol{x}^T G^T G \boldsymbol{x} \boldsymbol{y}^T G^T G \boldsymbol{y}}.
$$

记 $\boldsymbol{x}_G = G\boldsymbol{x}$ 和 $\boldsymbol{y}_G = G\boldsymbol{y}$，该式重写作：

$$
\boldsymbol{x}_G^T \boldsymbol{y}_G \leq \sqrt{\|\boldsymbol{x}_G\|^2 \|\boldsymbol{y}_G\|^2} = \|\boldsymbol{x}_G\| \|\boldsymbol{y}_G\|,
$$

这由柯西-施瓦茨不等式保证。

11.2

(a). 由于 $f(\boldsymbol{\alpha}) = \|\boldsymbol{x} - D\boldsymbol{\alpha}\|^2 = (\boldsymbol{x} - D\boldsymbol{\alpha})^T(\boldsymbol{x} - D\boldsymbol{\alpha})$，我们有

$$\nabla f = 2D^T(D\boldsymbol{\alpha} - \boldsymbol{x}).$$

因此，

$$\nabla f(\boldsymbol{\alpha}_1) - \nabla f(\boldsymbol{\alpha}_2) = 2D^T D(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2).$$

令 $D^T D = E\Lambda E^T$ 是 $D^T D$ 的谱分解，且 $[\Lambda]_{11} = \sigma_{\max}$ 是 $D^T D$ 的最大特征值。由于 $D^T D$ 是对称非负半定的，$\sigma_{\max} \geq 0$。由于 E 是正交的，对任意 $\boldsymbol{\alpha}$，我们有 $\|E\boldsymbol{\alpha}\| = \|E^T\boldsymbol{\alpha}\| = \|\boldsymbol{\alpha}\|$。那么，

$$\|\nabla f(\boldsymbol{\alpha}_1) - \nabla f(\boldsymbol{\alpha}_2)\| = 2\|D^T D(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)\| = 2\|E\Lambda E^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)\|.$$

$$= 2\|\Lambda E^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)\| \leq 2\sigma_{\max}\|E^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)\| = 2\sigma_{\max}\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|.$$

因此，$L = 2\sigma_{\max}$。李普希茨常数等于 $D^T D$ 最大特征值的两倍。

(b). 我们想要求解

$$\arg\min_{\boldsymbol{\alpha}} \lambda\|\boldsymbol{\alpha}\|_1 + \frac{L}{2}\left\|\boldsymbol{\alpha} - \left(\boldsymbol{\beta} - \frac{2}{L}D^T(D\boldsymbol{\beta} - \boldsymbol{x})\right)\right\|^2,$$

或等价地

$$\arg\min_{\boldsymbol{\alpha}} \left\|\boldsymbol{\alpha} - \left(\boldsymbol{\beta} - \frac{2}{L}D^T(D\boldsymbol{\beta} - \boldsymbol{x})\right)\right\|^2 + \frac{2\lambda}{L}\|\boldsymbol{\alpha}\|_1.$$

使用第一题的结果，我们有

$$p_L(\boldsymbol{\beta}) = \mathcal{T}_{\frac{\lambda}{L}}\left(\boldsymbol{\beta} - \frac{2}{L}D^T(D\boldsymbol{\beta} - \boldsymbol{x})\right).$$

由于使用软阈值得到 $\boldsymbol{\alpha}_{t+1} = p_L(\boldsymbol{\alpha}_t)$，就像我们在本章介绍的那样，每次迭代会导致稀疏性。

12.1

解：两个参数 $\boldsymbol{\pi}$ 和 A 完全确定一个 DTMC。$\boldsymbol{\pi}$ 的概率质量函数有 N 个数。但是，由于它们求和为 1，对 $\boldsymbol{\pi}$，我们只需要 $N-1$ 个数字。类似地，A 是一个 $N \times N$ 矩阵，但是对 A 的每一行，我们只需要 $N-1$ 个数。因此，我们需要

$$(N-1) + N \times (N-1) = N^2 - 1$$

个数字确定 DTMC。

12.2

解：我们需要证明 A^k 的每个元素是非负的，以及 A^k 的每行相加为 1。为了证明此结论，我们首先证明另一个结论：如果 A 和 B 是两个 $d \times d$ 的右随机矩阵，那么 AB 是一个右随机矩阵。由于对任意有效的 $1 \leq i, j, k \leq d$,

$$[AB]_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}, \quad a_{ik} \geq 0 \text{ 和} b_{kj} \geq 0,$$

我们知道 $[AB]_{ij} \geq 0$。

对任意 $1 \leq i \leq d$，我们有：

$$\begin{aligned}
\sum_{j=1}^{n} [AB]_{ij} &= \sum_{j=1}^{d} \sum_{k=1}^{d} a_{ik} b_{kj} \\
&= \sum_{k=1}^{d} a_{ik} \sum_{j=1}^{d} b_{kj} \\
&= \sum_{k=1}^{d} a_{ik} \\
&= 1.
\end{aligned}$$

因此，AB 的每行相加为 1。

现在令 $B = A$，我们知道 $A^2 = AA$ 是一个右随机矩阵。简单应用一下数学归纳法即可证明：对任意正整数 k，A^k 是一个右随机矩阵。