HW2

Gabriel ROMON

Problem 4.15

(a) Since $\delta_2 = \delta_1 + \beta(\delta_3 - P(X > \mu))$ and $P(X > \mu)$ is a constant,
$$\begin{aligned}\operatorname{var}(\delta_2) &= \operatorname{var}(\delta_1) + \operatorname{var}(\beta\delta_3) + 2\operatorname{cov}(\delta_1, \beta\delta_3)\,. \\ &= \operatorname{var}(\delta_1) + \beta\operatorname{var}(\delta_3) + 2\beta\operatorname{cov}(\delta_1, \delta_3)\end{aligned}$$
Since the X_i are i.i.d, $\operatorname{var}(\delta_3) = \frac{1}{n^2} \cdot n \cdot \operatorname{var}(\mathbb{1}_{X > \mu}) = \frac{1}{n} P(X > \mu)(1 - P(X > \mu))$. Besides,

$$\begin{aligned}
\operatorname{cov}(\delta_1, \delta_3) &= \operatorname{cov}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i > \alpha}, \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{X_j > \mu}\right) = \frac{1}{n^2} \sum_{1 \le i,j \le n} \operatorname{cov}(\mathbb{1}_{X_i > \alpha}, \mathbb{1}_{X_j > \mu}) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{cov}(\mathbb{1}_{X_i > \alpha}, \mathbb{1}_{X_i > \mu}) \quad \text{by independence} \\
&= \frac{1}{n^2} \cdot n \cdot [E(\mathbb{1}_{X > \alpha}\mathbb{1}_{X > \mu}) - E(\mathbb{1}_{X > \alpha})E(\mathbb{1}_{X > \mu})] \\
&= \frac{1}{n} P(X > a)(1 - P(X > \mu))
\end{aligned}$$

(b) Since $\operatorname{var}(\delta_2) = \operatorname{var}(\delta_1) + \beta\operatorname{var}(\delta_3) + 2\beta\operatorname{cov}(\delta_1, \delta_3)$, δ_2 improves over δ_1 in terms of variance iff $\beta^2 \operatorname{var}(\delta_3) + 2\beta\operatorname{cov}(\delta_1, \delta_3) < 0$. Since $\frac{\operatorname{cov}(\delta_1, \delta_3)}{\operatorname{var}(\delta_3)} = \frac{P(X > a)}{P(X > \mu)} > 0$, this can only happen iff $\beta < 0$ and $\beta > -2\frac{\operatorname{cov}(\delta_1, \delta_3)}{\operatorname{var}(\delta_3)} = -2\frac{P(X > a)}{P(X > \mu)}$.

In order to find a suitable β, one has to know a lower bound on $P(X > a)$. Note that the optimal β remains unknown.

(c) In the normal case, deriving a lower bound on $P(X > a)$ can be done through repeated integration by parts:

$$\begin{aligned}
P(X > a) &= \int_a^\infty \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = \int_a^\infty \frac{1}{t} \cdot \frac{te^{-t^2/2}}{\sqrt{2\pi}} dt = \frac{e^{-a^2/2}}{\sqrt{2\pi}a} - \int_a^\infty \frac{1}{t^2} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt \\
&= \frac{e^{-a^2/2}}{\sqrt{2\pi}a} - \left(\frac{e^{-a^2/2}}{\sqrt{2\pi}a^3} - \int_a^\infty \frac{3}{t^4} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt\right) \\
&\ge \frac{e^{-a^2/2}}{\sqrt{2\pi}} \left(\frac{1}{a} - \frac{1}{a^3}\right)
\end{aligned}$$

Since the normal distribution is symmetric, we set $\mu = 0$ and our value for β is therefore $-4\frac{e^{-a^2/2}}{\sqrt{2\pi}}\left(\frac{1}{a} - \frac{1}{a^3}\right)$.
Python code for simulations is in the Appendix. For $a \in \{3, 5, 7\}$, we sample 10^5 points from the normal distribution, compute δ_1 and δ_2 and collect the values of $(\delta_1 - P(X > a))^2$ and $(\delta_2 - P(X > a))^2$. We repeat this experiment 10^4 times and average the values we found to get an estimate of the variance of each estimator. The results in the table below show that δ_2 provides indeed a slight improvement.

a	$P(X > a)$	Lower bound on $P(X > a)$	Mean of $(\delta_1 - P(X > a))^2$	Mean of $(\delta_2 - P(X > a))^2$
3	$1.3499 \cdot 10^{-3}$	$1.3131 \cdot 10^{-3}$	$1.5195 \cdot 10^{-8}$	$1.5183 \cdot 10^{-8}$
5	$2.8665 \cdot 10^{-7}$	$2.8545 \cdot 10^{-7}$	$2.8726 \cdot 10^{-12}$	$2.8725 \cdot 10^{-12}$
7	$1.2799 \cdot 10^{-12}$	$1.2783 \cdot 10^{-12}$	$1.6381 \cdot 10^{-24}$	$1.6380 \cdot 10^{-24}$

Table 1: Simulation results for $f \sim \mathcal{N}(0, 1)$

(d) It is known that Student's distribution $\mathcal{T}_\nu$ converges in distribution to $\mathcal{N}(0,1)$ as $\nu \to \infty$. For large ν, the cdf of $\mathcal{T}_\nu$ can therefore be approximated by that of the gaussian. However, $\nu = 5$ in the problem. Unlike the normal case, deriving a lower bound for $P(X > a)$ isn't trivial. Consequently, the lower bounds we use are obtained by truncation of the true value. They could also be found by running simulations with δ_1 only, but for the sake of simplicity we chose to truncate the true value instead. The simulation method is identical to the normal case.

a	$P(X > a)$	Lower bound on $P(X > a)$	Mean of $(\delta_1 - P(X > a))^2$	Mean of $(\delta_2 - P(X > a))^2$
3	$1.5049 \cdot 10^{-2}$	$1.5000 \cdot 10^{-2}$	$1.5019 \cdot 10^{-7}$	$1.4968 \cdot 10^{-7}$
5	$2.0523 \cdot 10^{-3}$	$2.0000 \cdot 10^{-3}$	$2.0805 \cdot 10^{-8}$	$2.0760 \cdot 10^{-8}$
7	$4.5837 \cdot 10^{-4}$	$4.5000 \cdot 10^{-4}$	$4.6487 \cdot 10^{-9}$	$4.6454 \cdot 10^{-9}$

Table 2: Simulation results for $f \sim \mathcal{T}_5$

Problem 5.9

(a) Let $(x_1, z_1), \dots, (x_n, z_n)$ be an i.i.d sample from (X, Z). Since Z follows a Bernoulli with parameter θ, Bayes' theorem yields $f_{(X,Z)}(x_i, z_i) = f_{X|Z=z_i}(x_i) f_Z(z_i)$, hence

$$\begin{aligned} L^c(\theta|\mathbf{x}, \mathbf{z}) &= \prod_{i=1}^n f_{(X,Z)}(x_i, z_i) = \prod_{i=1}^n f_{X|Z=z_i}(x_i) f_Z(z_i) \\ &= \prod_{i=1}^n (z_i g(x_i) + (1 - z_i) h(x_i)) \theta^{z_i} (1-\theta)^{1-z_i} \end{aligned}$$

(b) Applying Bayes' theorem again,

$$f_{Z|X=x_i}(z_i) = \frac{f_{X|Z=z_i}(x_i) f_Z(z_i)}{f_X(x_i)} = \frac{(z_i g(x_i) + (1 - z_i) h(x_i)) \theta^{z_i} (1-\theta)^{1-z_i}}{\theta g(x_i) + (1-\theta) h(x_i)}$$

Since $z_i \in \{0, 1\}$, the numerator can be rewritten as $[\theta g(x_i)]^{z_i} [(1-\theta) h(x_i)]^{1-z_i}$, and we identify the density of a Bernoulli with parameter $\dfrac{\theta g(x_i)}{\theta g(x_i) + (1-\theta) h(x_i)}$. The conditional distribution of Z given θ and $X = x_i$ is thus $\mathcal{B}(\frac{\theta g(x_i)}{\theta g(x_i) + (1-\theta) h(x_i)})$, hence

$$E(Z|\theta, X = x_i) = \frac{\theta g(x_i)}{\theta g(x_i) + (1-\theta) h(x_i)}$$

Let

$$\begin{aligned} Q(\theta|\hat{\theta}_j, \mathbf{x}) &= E_{\mathbf{z}|\hat{\theta}_j, \mathbf{x}}(\log L^c(\theta|\mathbf{x}, \mathbf{z})) \\ &= E_{\mathbf{z}|\hat{\theta}_j, \mathbf{x}} \left(\log \prod_{i=1}^n [\theta g(x_i)]^{z_i} [(1-\theta) h(x_i)]^{1-z_i} \right) \\ &= \sum_{i=1}^n E_{\mathbf{z}|\hat{\theta}_j, \mathbf{x}} \left(z_i \log(\theta g(x_i)) + (1 - z_i) \log((1-\theta) h(x_i)) \right) \\ &= \sum_{i=1}^n \log(\theta g(x_i)) \frac{\hat{\theta}_j g(x_i)}{\hat{\theta}_j g(x_i) + (1-\hat{\theta}_j) h(x_i)} + \log((1-\theta) h(x_i)) \frac{(1-\hat{\theta}_j) h(x_i)}{\hat{\theta}_j g(x_i) + (1-\hat{\theta}_j) h(x_i)} \end{aligned}$$

Note that $\theta \mapsto Q(\theta|\hat{\theta}_j, \mathbf{x})$ is differentiable and concave as the sum of concave functions. Critical points are therefore global maxima. Since

$$\frac{dQ(\theta|\hat{\theta}_j, \mathbf{x})}{d\theta} = \sum_{i=1}^n \frac{1}{\theta} \frac{\hat{\theta}_j g(x_i)}{\hat{\theta}_j g(x_i) + (1-\hat{\theta}_j) h(x_i)} - \frac{1}{1-\theta} \frac{(1-\hat{\theta}_j) h(x_i)}{\hat{\theta}_j g(x_i) + (1-\hat{\theta}_j) h(x_i)}$$

we have $\dfrac{dQ(\theta|\hat{\theta}_j, \mathbf{x})}{d\theta} = 0 \iff (1-\theta)\displaystyle\sum_{i=1}^{n} \frac{\hat{\theta}_j g(x_i)}{\hat{\theta}_j g(x_i) + (1-\hat{\theta}_j)h(x_i)} = \theta \sum_{i=1}^{n} \frac{(1-\hat{\theta}_j)h(x_i)}{\hat{\theta}_j g(x_i) + (1-\hat{\theta}_j)h(x_i)}$ which yields

$$\hat{\theta}_{j+1} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\theta}_j g(x_i)}{\hat{\theta}_j g(x_i) + (1-\hat{\theta}_j)h(x_i)}$$

(c) Let $\phi : \theta \mapsto \frac{1}{n}\sum_{i=1}^{n} \frac{\theta g(x_i)}{\theta g(x_i)+(1-\theta)h(x_i)}$. We have proved that $\forall j, \hat{\theta}_{j+1} = \phi(\hat{\theta}_j)$, so the EM sequence is a discrete dynamical system. Since $\phi'(\theta) = \sum_{i=1}^{n} \frac{g(x_i)h(x_i)}{(\theta g(x_i)+(1-\theta)h(x_i))^2} \geq 0$, ϕ increases. Besides, it is bounded above by 1. As a result, $\hat{\theta}_j$ is increasing and bounded, thus convergent.

$\hat{\theta}_j$ converges to some $\hat{\theta}$ such that $\phi(\hat{\theta}) = \hat{\theta}$, which rewrites as

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\theta} g(x_i)}{\hat{\theta} g(x_i) + (1-\hat{\theta})h(x_i)} = \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}$$

$$\iff \hat{\theta}(1-\hat{\theta})\sum_{i=1}^{n} \frac{g(x_i) - h(x_i)}{\hat{\theta} g(x_i) + (1-\hat{\theta})h(x_i)} = 0$$

We expect that $\hat{\theta} \notin \{0,1\}$ hence $\displaystyle\sum_{i=1}^{n} \frac{g(x_i) - h(x_i)}{\hat{\theta} g(x_i) + (1-\hat{\theta})h(x_i)} = 0$

Note that the log-likelihood writes as

$$\log L(\theta, \mathbf{x}) = \sum_{i=1}^{n} \log(\theta g(x_i)) + (1-\theta)h(x_i))$$

which is concave and differentiable. The first-order condition is $\displaystyle\sum_{i=1}^{n} \frac{g(x_i) - h(x_i)}{\theta g(x_i) + (1-\theta)h(x_i)} = 0$, hence $\hat{\theta}$ is a maximum likelihood estimator, as well as being the EM limiting estimator.

Problem 6.9

Let P be the transition matrix of an aperiodic and irreducible chain on some finite state space $E = \{1, \dots, p\}$. We will first prove the following lemma:
Lemma: Let A be a subset of $\mathbb{N}$ closed under addition and with $\gcd A = 1$. Then there exists n_0 such that $\forall n \geq n_0, n \in A$.
Proof: By Bézout's identity, there exists $a_1, \dots, a_n \in A$ and $b_1, \dots, b_n \in \mathbb{Z}$ such that $\sum_{k=1}^{n} a_k b_k = 1$. Note that

$$\underbrace{\sum_{k, b_k > 0} a_k b_k}_{:= x \in A} = 1 + \underbrace{\sum_{k, b_k \leq 0} a_k(-b_k)}_{:= y \in A}$$

If $y = 0$, then $1 = x \in A$ and we're done. Otherwise, $y \geq 1$. Let $m \geq y^2$ and write the Euclidean division of m by y as $m = ky + r$ with $0 \leq r < y$. Note that $y^2 \leq m < (k+1)y$, hence $k > y - 1 \implies k \geq y \implies k > r$. Besides, $m = ky + r = ky + r(x-y) = (k-r)y + rx$, where $k - r > 0$, $x, y \in A$. Hence $m \in A$, and this holds for all $m \geq y^2$. $\square$

To apply the lemma, we prove that for any state i, the set $A_i := \{n, P_{ii}^{(n)} > 0\}$ is closed under addition. Indeed, if $P_{ii}^{(n)} > 0$ and $P_{ii}^{(m)} > 0$, by Chapman-Kolmogorov equation we have $P_{ii}^{(n+m)} \geq P_{ii}^{(n)} P_{ii}^{(m)} > 0$. By the lemma, for every state i, there exists some n_i such that $n \geq n_i \implies n \in A_i$. For fixed i and any j, there exists by irreducibility some m_{ij} such that $P_{ij}^{(m_{ij})} > 0$. Thus, for $q \geq n_i + m_{ij}$, $P_{ij}^{(q)} \geq P_{ii}^{(q-m_{ij})} P_{ij}^{(m_{ij})} > 0$, hence for $q \geq n_i + \max_{1 \leq j \leq p} m_{ij}$ and any state j, $P_{ij}^{(q)} > 0$. Finally, it suffices to set $n_0 := \max_{1 \leq i \leq p}(n_i + \max_{1 \leq j \leq p} m_{ij})$ to get

$$\forall n \geq n_0, \forall i, j, P_{ij}^{(n)} > 0$$

Note that we proved a stronger statement than what was asked in the problem.

For the converse, if $P^{(n)}$ has positive coefficients, then all states communicate with one another, hence the chain is irreducible.

Problem 6.54

1. A two-state chain has transition matrix $P = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}$. If the chain is ergodic, is has a stationary distribution $\pi = (\pi_1, \pi_2)$ such that $\pi P = \pi$. After solving the resulting linear system of equations, one gets $\pi = (\frac{1-b}{2-a-b}, \frac{1-a}{2-a-b})$ and it's easy to check next that $\tilde{P} = P$, hence the chain is reversible.

2. A chain with symmetric transition matrix has $(\frac{1}{n}, \dots, \frac{1}{n})$ as invariant distribution. Indeed,

$$(\pi P)_j = \sum_{i=1}^{n} \pi_i P_{ij} = \frac{1}{n} \sum_{i=1}^{n} P_{ij} = \frac{1}{n} \sum_{i=1}^{n} P_{ji} = \frac{1}{n}$$

   Therefore, $\tilde{P}_{ij} = \frac{\pi_j P_{ji}}{\pi_i} = P_{ji} = P_{ij}$ and P is reversible.

3. For the given matrix, solving the linear system $\pi P = \pi, \mathbf{1}^T \pi = 1$ yields $\pi = (0.1, 0.2, 0.4, 0.2, 0.1)$. Note that $\tilde{P}_{12} = \frac{0.2 \cdot 0.5}{0.1} = 1 \neq P_{12} = 0$, hence the chain is not reversible.

Appendix

```
import numpy as np
from scipy.stats import norm, t

#Simulations for the normal case
for a in [3, 5, 7]:
    beta = -4*1/np.sqrt(2*np.pi)*np.exp(-a*a/2)*(1/a-1/a**3)
    param_true = 1-norm.cdf(a)
    var1, var2 = [], []
    for _ in range(10000):
        sample = np.random.normal(size=100000)
        param_esti1 = np.mean(sample>a)
        param_esti2 = np.mean(sample>a) + beta*(np.mean(sample>0)-0.5)
        var1.append((param_true-param_esti1)**2)
        var2.append((param_true-param_esti2)**2)
    print('a='+str(a)+':', np.mean(var1), np.mean(var2))

#Simulations for the Student distribution
lower = [0.015, 0.002, 0.00045]
for (a, lower) in zip([3, 5, 7], lower):
    beta = -4*lower
    param_true = 1-t.cdf(a, df =5)
    var1, var2 = [], []
    for _ in range(10000):
        sample = np.random.standard_t(5, size=100000)
        param_esti1 = np.mean(sample>a)
        param_esti2 = np.mean(sample>a) + beta*(np.mean(sample>0)-0.5)
        var1.append((param_true-param_esti1)**2)
        var2.append((param_true-param_esti2)**2)
    print('a='+str(a)+':', np.mean(var1), np.mean(var2))
```