Probabilistic models, Spring 2013
Exercise 2: Solutions

---

5. The numbers of heads and tails are $N_H = 9$, $N_T = 6$.

a)

$$\begin{aligned} P(H|\theta) &= \theta \\ P(T|\theta) &= 1-\theta \\ P(D|\theta) &= \theta^{N_H}(1-\theta)^{N_T} \end{aligned}$$

$$\frac{\partial}{\partial\theta}P(D|\theta) = \theta^{(N_H-1)}(1-\theta)^{(N_T-1)}\left[N_H - \theta(N_H+N_T)\right]$$

$$\begin{aligned} \frac{\partial}{\partial\theta}P(D|\theta) = 0 \quad &\rightarrow \quad \theta^{(N_H-1)}(1-\theta)^{(N_T-1)}\left[N_H - \theta(N_H+N_T)\right] = 0 \\ &\rightarrow \quad N_H - \theta(N_H+N_T) = 0 \\ &\rightarrow \quad \theta = \frac{N_H}{N_H+N_T} \end{aligned}$$

$$\hat{\theta} = \arg\max_{\theta} P(D \mid \theta) = \frac{N_H}{N_H+N_T} = \frac{9}{15} = 0.6$$

b)

$$P(D \mid \theta = \hat{\theta}) = \hat{\theta}^{N_H}(1-\hat{\theta})^{N_H} = \left(\frac{9}{15}\right)^9\left(\frac{6}{15}\right)^6 \approx 0.000041$$

c)

$$P(D \mid \theta = \frac{1}{2}) = \left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^6 \approx 0.000031$$

d)

Both uniform and Jeffreys priors are special cases of the Beta distribution, so the posterior with a Beta prior is calculated. The Beta prior is

$$P(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}c$$

where α and β are its parameters and c is a normalization constant. The posterior distribution with the Beta prior is thus

$$\begin{aligned} P(\theta|D) = P(D|\theta)P(\theta) &= \theta^{N_H}(1-\theta)^{N_T} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}c \\ &= \theta^{N_H+\alpha-1}(1-\theta)^{N_T+\beta-1} \end{aligned}$$

which itself is a Beta distribution: $\text{Beta}(\alpha+N_H, \beta+N_T)$.

The posteriors with the uniform and Jeffreys priors can now be easily calculated:

Uniform prior: $\alpha = 1, \beta = 1 \quad \Rightarrow \quad P(\theta \mid D) \sim \text{Beta}(10,7)$

Jeffreys prior: $\alpha = 0.5, \beta = 0.5 \quad \Rightarrow \quad P(\theta \mid D) \sim \text{Beta}(9.5,6.5)$

e)

ML parameters: $P(X = H \mid D) = \frac{N_H}{N_H+N_T}$

Uniform prior: $P(X = H \mid D) = \frac{N_H+1}{N_H+N_T+2}$

	ML parameters	uniform prior
$P(X=H)=$	–	$1/2$
$P(X=H \mid H)=$	$1/1$	$2/3$
$P(X=T \mid HH)=$	$0/2$	$1/4$
$P(X=H \mid HHT)=$	$2/3$	$3/5$
$P(X=T \mid HHTH)=$	$1/4$	$2/6$
$P(X=T \mid HHTHT)=$	$2/5$	$3/7$
$P(X=H \mid HHTHTT)=$	$3/6$	$4/8$
$P(X=H \mid HHTHTTH)=$	$4/7$	$5/9$
$P(X=T \mid HHTHTTHH)=$	$3/8$	$4/10$
$P(X=H \mid HHTHTTHHT)=$	$5/9$	$6/11$
$P(X=T \mid HHTHTTHHTH)=$	$4/10$	$5/12$
$P(X=T \mid HHTHTTHHTHT)=$	$5/11$	$6/13$
$P(X=H \mid HHTHTTHHTHTT)=$	$6/12$	$7/14$
$P(X=H \mid HHTHTTHHTHTTH)=$	$7/13$	$8/15$
$P(X=H \mid HHTHTTHHTHTTHH)=$	$8/14$	$9/16$
Product $\approx$	–	0.0000125

---

6. The numbers of heads and tails are $N_H = 7$, $N_T = 3$.

i)

$$\hat{\theta} = \frac{7}{10} = 0.7$$

ii)

$$P(D \mid \hat{\theta}) \approx 0.00222$$

iii)

$$P(D \mid \theta = \frac{1}{2}) \approx 0.00098$$

iv)

Uniform prior: $P(\theta \mid D) \sim \text{Beta}(8,4)$

Jeffreys prior: $P(\theta \mid D) \sim \text{Beta}(7.5,3.5)$

v)

	ML parameters	uniform prior
$P(X=T)=$	–	$1/2$
$P(X=T \mid T)=$	1	$2/3$
$P(X=T \mid TT)=$	1	$3/4$
$P(X=H \mid TTT)=$	0	$1/5$
$P(X=H \mid TTTH)=$	$1/4$	$2/6$
$P(X=H \mid TTTHH)=$	$2/5$	$3/7$
$P(X=H \mid TTTHHH)=$	$3/6$	$4/8$
$P(X=H \mid TTTHHHH)=$	$4/7$	$5/9$
$P(X=H \mid TTTHHHHH)=$	$5/8$	$6/10$
$P(X=H \mid TTTHHHHHH)=$	$6/9$	$7/11$
Product $\approx$	–	0.000758

---

7. a) The maximum likelihood parameters $\theta_1, \theta_2, \ldots, \theta_6$ are known$^{(*)}$ to be of form $\hat{\theta}_i = \frac{N_i}{N_1+N_2+\ldots+N_6}$. That is:

$$\hat{\theta}_1 = \frac{8}{50}$$

$$\hat{\theta}_2 = \frac{4}{50}$$

$$\hat{\theta}_3 = \frac{9}{50}$$
$$\hat{\theta}_4 = \frac{7}{50}$$
$$\hat{\theta}_5 = \frac{12}{50}$$
$$\hat{\theta}_6 = \frac{10}{50}$$

(*) Proof (not required): Let's assume a multinomial model with k classes and observed counts $N_1, N_2, \ldots, N_k$. We want to maximise the likelihood function

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \theta_1^{N_1} \theta_2^{N_2} \cdots \theta_k^{N_k}$$

subject to restrictions $g(\theta_1, \theta_2, \ldots, \theta_k) = \theta_1 + \theta_2 + \ldots + \theta_k = 1$ and $\theta_i \geq 0$ for $i = 1, \ldots, k$. If $\theta_i = 0$ for any i, then $L = 0$ (unless $N_i = 0$). On the other hand, if $\theta_i > 0$ for all i, then $L > 0$. Therefore, L is maximized in the open set where $\theta_i > 0$ for all i. Since the logarithm is a strictly increasing function, we can instead maximize the logarithm of L, that is:

$$l(\theta_1, \theta_2, \ldots, \theta_k) = \ln L(\theta_1, \theta_2, \ldots, \theta_k) = N_1 \ln \theta_1 + N_2 \ln \theta_2 + \ldots + N_k \ln \theta_k$$

So we want to find $\arg\max_{\theta_1, \theta_2, \ldots, \theta_k} l(\theta_1, \theta_2, \ldots, \theta_k)$. Using Lagrange multipliers we get the following equations that need to be satisfied for some $\lambda \in \mathbb{R}$:

$$\begin{cases} \nabla l = \lambda \nabla g \\ g = 1 \end{cases} \Leftrightarrow \begin{cases} \frac{\partial l}{\partial \theta_1} = \lambda \frac{\partial g}{\partial \theta_1} \\ \frac{\partial l}{\partial \theta_2} = \lambda \frac{\partial g}{\partial \theta_2} \\ \vdots \\ \frac{\partial l}{\partial \theta_k} = \lambda \frac{\partial g}{\partial \theta_k} \\ g = 1 \end{cases} \Leftrightarrow \begin{cases} \frac{N_1}{\theta_1} = \lambda \\ \frac{N_2}{\theta_2} = \lambda \\ \vdots \\ \frac{N_k}{\theta_k} = \lambda \\ \theta_1 + \theta_2 + \ldots + \theta_k = 1 \end{cases} \Leftrightarrow \begin{cases} \theta_1 = \frac{N_1}{\lambda} \\ \theta_2 = \frac{N_2}{\lambda} \\ \vdots \\ \theta_k = \frac{N_k}{\lambda} \\ \theta_1 + \theta_2 + \ldots + \theta_k = 1 \end{cases}$$

By inserting the k first equations to the last equation we can solve $\lambda = N_1 + N_2 + \ldots + N_k$. Thus the likelihood L is maximized when

$$\theta_i = \frac{N_i}{N_1 + N_2 + \ldots + N_k} \quad , i = 1, 2, \ldots, k.$$

b) (i) uniform prior:

$$P(\theta_1, \theta_2, \ldots, \theta_6 \mid D) \sim \text{Dir}(N_1 + 1, N_2 + 1, \ldots, N_6 + 1) = \text{Dir}(9, 5, 10, 8, 13, 11)$$

(ii) Jeffreys prior:

$$P(\theta_1, \theta_2, \ldots, \theta_6 \mid D) \sim \text{Dir}(N_1 + 0.5, N_2 + 0.5, \ldots, N_6 + 0.5) = \text{Dir}(8.5, 4.5, 9.5, 7.5, 12.5, 10.5)$$

c) In uniform prior we have $\alpha_1 = \alpha_2 = \ldots = \alpha_6 = 1$. Thus we get

$$
\begin{aligned}
P(X=1 \mid D) &= \frac{1+N_1}{\sum_{j=1}^{6}(1+N_j)} = \frac{9}{56} \\
P(X=2 \mid D) &= \frac{1+N_2}{\sum_{j=1}^{6}(1+N_j)} = \frac{5}{56} \\
P(X=3 \mid D) &= \frac{1+N_3}{\sum_{j=1}^{6}(1+N_j)} = \frac{10}{56} \\
P(X=4 \mid D) &= \frac{1+N_4}{\sum_{j=1}^{6}(1+N_j)} = \frac{8}{56} \\
P(X=5 \mid D) &= \frac{1+N_5}{\sum_{j=1}^{6}(1+N_j)} = \frac{13}{56} \\
P(X=6 \mid D) &= \frac{1+N_6}{\sum_{j=1}^{6}(1+N_j)} = \frac{11}{56}
\end{aligned}
$$

d) Since have two equations and 6 free variables $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$, we can fix four of them arbitrarily: for example let $\alpha_1 = \alpha_2 = \alpha_4 = \alpha_6 = 1$. Now by solving

$$
\begin{cases} P(X=3 \mid D) = \frac{1}{4} \\ P(X=5 \mid D) = \frac{1}{4} \end{cases}
\Leftrightarrow
\begin{cases} \frac{\alpha_3+9}{50+\sum_{i=1}^{6}\alpha_i} = \frac{1}{4} \\ \frac{\alpha_5+12}{50+\sum_{i=1}^{6}\alpha_i} = \frac{1}{4} \end{cases}
\Leftrightarrow
\begin{cases} \frac{\alpha_3+9}{\alpha_3+\alpha_5+54} = \frac{1}{4} \\ \frac{\alpha_5+12}{\alpha_3+\alpha_5+54} = \frac{1}{4} \end{cases}
\Leftrightarrow
$$

$$
\begin{cases} 4\alpha_3+36 = \alpha_3+\alpha_5+54 \\ 4\alpha_5+48 = \alpha_3+\alpha_5+54 \end{cases}
\Leftrightarrow
\begin{cases} 3\alpha_3 = \alpha_5+18 \\ 3\alpha_5 = \alpha_3+6 \end{cases}
\Leftrightarrow
\begin{cases} \alpha_5 = 3\alpha_3-18 \\ \alpha_3 = 3\alpha_5-6 \end{cases}
$$

That gives us

$$
\alpha_5 = 3(3\alpha_5-6)-18 = 9\alpha_5-36 \quad \Leftrightarrow \quad 36 = 8\alpha_5 \quad \Leftrightarrow \quad \alpha_5 = 4.5
$$
$$
\alpha_3 = 3\times 4.5 - 6 = 7.5
$$

so $\alpha_3 = 7.5$ and $\alpha_5 = 4.5$. Thus an example of such prior distribution is $\mathrm{Dir}(1,1,7.5,1,4.5,1)$.
It is finally verified that the prior distribution gives correct probablities:

$$
\begin{aligned}
P(X=3 \mid D) &= \frac{9+\alpha_3}{50+\sum_{j=1}^{6}\alpha_j} = \frac{9+7.5}{54+7.5+4.5} = \frac{16.5}{66} = \frac{1}{4} \\
P(X=5 \mid D) &= \frac{12+\alpha_5}{50+\sum_{j=1}^{6}\alpha_j} = \frac{12+4.5}{54+7.5+4.5} = \frac{16.5}{66} = \frac{1}{4}
\end{aligned}
$$

---

8. Let

$$
\begin{aligned}
f(\theta) &= P(D \mid \theta)P(\theta) \\
&= \theta^{N_b}(1-\theta)^{N_w} \cdot c\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= c\theta^{N_b+\alpha-1}(1-\theta)^{N_w+\beta-1}
\end{aligned}
$$

where $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is a constant.

We want to find the maximum of $f(\theta)$ on the interval $0 \le \theta \le 1$. If $\theta = 0$ or $\theta = 1$, then $f(\theta) = 0$. Otherwise $f(\theta) > 0$. Thus the maximum can be found from the open interval $\theta \in (0,1)$. Now $f(\theta)$ is maximized when the logarithm of it

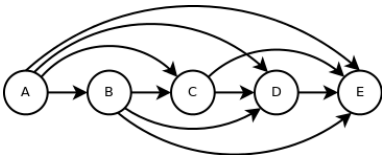
$$
\ln f(\theta) = \ln c + (N_b+\alpha-1)\ln\theta + (N_w+\beta-1)\ln(1-\theta)
$$

is maximized, which in turn happens when its derivative is 0, that is:

$$\begin{aligned}
\frac{N_b+\alpha-1}{\theta}-\frac{N_w+\beta-1}{1-\theta} &= 0 \\
(N_b+\alpha-1)(1-\theta) &= (N_w+\beta-1)\theta \\
(N_b+N_w+\alpha+\beta-2)\theta &= N_b+\alpha-1 \\
\theta &= \frac{N_b+\alpha-1}{N_b+N_w+\alpha+\beta-2}.
\end{aligned}$$

---

9. Let's first construct a full DAG:



This corresponds to the following factorization:

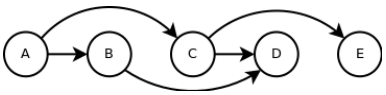
$$P(A,B,C,D,E) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)P(E \mid A,B,C,D)$$

From the description given by the expert we can derive the following independencies:
$P(E \mid A,B,C,D) = P(E \mid C)$
$P(D \mid A,B,C) = P(D \mid B,C)$
$P(C \mid A,B) = P(C \mid A)$

In another form:
$E \perp A \mid C$
$E \perp B \mid C$
$E \perp D \mid C$
$D \perp A \mid \{B,C\}$
$C \perp B \mid A$

Note: The expert did not say anything about possible dependence or independence between the A and B. So we can't say that A and B are (conditionally) independent. In some situations it might make sense to assume independence if not explicitly told otherwise. But here we can't be sure so we are not going to remove the corresponding arc from the DAG.

After removing the arcs corresponding the independencies listed above we get:



This is our final DAG. To make it a bit easier to read, we can still rearrange the nodes as follows: