10-602: Statistical Approaches to Learning and Discovery

Problem Set 4 Solutions

---

1. (a) p_1 and p_2 are independent given the data.

$$p(p_1|e_1,n_1) \sim \mathcal{B}(e_1+1, n_1-e_1+1) \quad (1)$$
$$p(p_2|e_2,n_2) \sim \mathcal{B}(e_2+1, n_2-e_2+1) \quad (2)$$

$$Pr(p_1 < p_2|e_1,n_1,e_2,n_2) = \int_0^1 \int_0^1 \text{step}(p_2-p_1)p(p_1|e_1,n_1)p(p_2|e_2,n_2)dp_1dp_2 \quad (3)$$
$$= \int_0^1 \int_0^{p_2} p(p_1|e_1,n_1)p(p_2|e_2,n_2)dp_1dp_2 \quad (4)$$

Define the Beta function and incomplete Beta function:

$$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad (5)$$
$$\beta(c;a,b) = \int_0^c x^{a-1}(1-x)^{b-1}dx \quad (6)$$

$$Pr(p_1 < p_2|e_1,n_1,e_2,n_2) = \frac{\int_0^1 p_2^{e_2}(1-p_2)^{n_2-e_2}\int_0^{p_2} p_1^{e_1}(1-p_1)^{n_1-e_1}dp_1dp_2}{\beta(e_2,n_2-e_2)\beta(e_1,n_1-e_1)} \quad (7)$$
$$= \frac{\int_0^1 p_2^{e_2}(1-p_2)^{n_2-e_2}\beta(p_2;e_1+1,n_1-e_1+1)dp_2}{\beta(e_2,n_2-e_2)\beta(e_1,n_1-e_1)} \quad (8)$$

(b) The simplest approach is to approximate (3) using a grid of quadrature points in the unit square. However, many of the points will be wasted since the posterior probability that p_1 or $p_2 > 0.2$ is very small. In other words, the integrand is sparse. Instead we can integrate over the square $[0, 0.2]^2$. This will still waste half of the points, i.e. those where $p_1 > p_2$. To avoid this, we can use (7) and apply the midpoint rule to the integral over p_2 and then recursively to the integral over p_1.

Even with these enhancements, a huge number of points is required. As shown in figure 1, at least 50,000 points are required for an error of 10^{-2}. The true value is 0.8217.

A better approach is to exploit a fast routine for the incomplete beta function—such as Matlab's `betainc`. Then you need only apply the midpoint rule to (8). About 15 points are required for an error of 10^{-2}, as shown in figure 1.

(c) If we approximate $p_1 \sim \mathcal{N}(m_1, v_1)$ and $p_2 \sim \mathcal{N}(m_2, v_2)$, then

$$p(p_1 < p_2|D) \approx \phi(\frac{m_2 - m_1}{\sqrt{v_1+v_2}}) \quad (9)$$

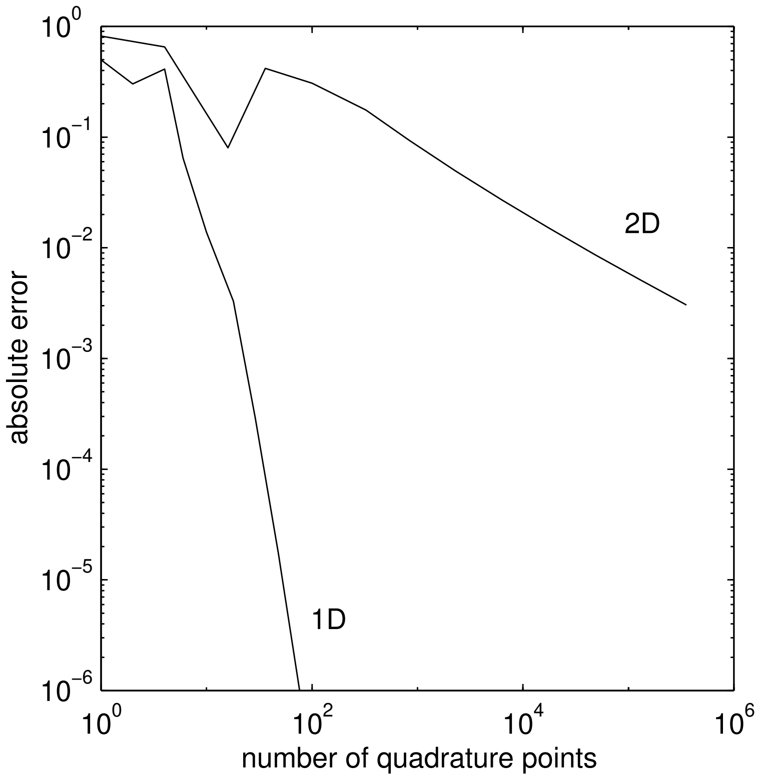

Figure 1: The absolute error versus the number of quadrature points for the midpoint rule applied to (3), a two dimensional integral, and (8), a one dimensional integral.

The exact mean and variance of the posterior are:

$$
\begin{aligned}
E[p_1|D] &= \frac{e_1+1}{n_1+2} = 0.0132 && (10)\\
\mathrm{var}(p_1|D) &= \frac{E[p_1|D](1-E[p_1|D])}{n_1+3} = 8.49\times 10^{-5} && (11)\\
E[p_2|D] &= 0.03 && (12)\\
\mathrm{var}(p_2|D) &= 2.77\times 10^{-4} && (13)\\
p(p_1 < p_2|D) &\approx 0.8035 && (14)
\end{aligned}
$$

The mode of the log-posterior is $\hat{p}_j = e_j/n_j$ and the second derivative about the mode is

$$
\frac{d^2 \log p(p_j|D)}{dp_j^2} = -\frac{e_j}{p_j^2} - \frac{n_j-e_j}{(1-p_j)^2} \qquad (15)
$$

which gives the second approximation

$$
\begin{aligned}
m_1 = \hat{p}_1 &= 0.0067 \qquad m_2 = 0.02 && (16)\\
v_1 = \left(\frac{e_1}{p_1^2} + \frac{n_1-e_1}{(1-p_1)^2}\right)^{-1} &= 4.42\times 10^{-5} \qquad v_2 = 2\times 10^{-4} && (17)\\
p(p_1 < p_2|D) &\approx 0.8052 && (18)
\end{aligned}
$$

(d)

$$
\begin{aligned}
\frac{do_j}{dp_j} &= \frac{1}{p_j(1-p_j)} && (19)\\
m_j &= \log\left(\frac{E[p_j|D]}{1-E[p_j|D]}\right) && (20)\\
v_j &= \left(\left.\frac{do_j}{dp_j}\right|_{p_j=E[p_j|D]}\right)^2 \mathrm{var}(p_j|D) = \frac{1}{E[p_j|D](1-E[p_j|D])(n_j+3)} && (21)\\
p(p_1 < p_2|D) &\approx 0.8143 && (22)
\end{aligned}
$$

(e)

$$
\begin{aligned}
\frac{ds_j}{dp_j} &= \frac{1}{\sqrt{p_j(1-p_j)}} && (23)\\
m_j &= 2\mathrm{arcsin}(\sqrt{E[p_j|D]}) && (24)\\
v_j &= \left(\left.\frac{ds_j}{dp_j}\right|_{p_j=E[p_j|D]}\right)^2 \mathrm{var}(p_j|D) = \frac{1}{n_j+3} && (25)\\
p(p_1 < p_2|D) &\approx 0.8161 && (26)
\end{aligned}
$$

The latter two approximations are more accurate, as expected. Figure 2 plots the posterior distribution of p_1 versus o_1 and s_1. The latter two are more Gaussian.

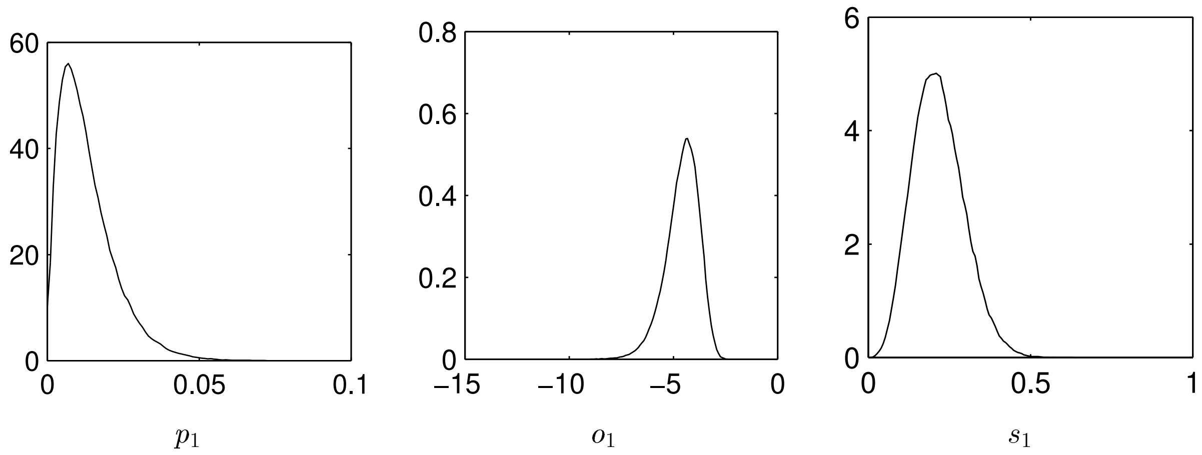

Figure 2: Posterior distributions $p(p_1|D)$, $p(o_1|D)$, and $p(s_1|D)$.

2. (a) The log-likelihood is

$$\begin{aligned} l(\theta) &= y_1 \log(2+\theta) + (y_2+y_3)\log(1-\theta) + y_4 \log\theta && (27) \\ \frac{dl(\theta)}{d\theta} &= \frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta} && (28) \\ \frac{d^2 l(\theta)}{d\theta^2} &= -\frac{y_1}{(2+\theta)^2} - \frac{y_2+y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2} && (29) \end{aligned}$$

A Newton step is $\theta^{new} = \theta - \frac{dl(\theta)/d\theta}{d^2l(\theta)/d\theta^2}$. The second derivative is always negative so it must stop at a maximum. Starting from $\theta = 0.2, 0.4, 0.8$, Newton converges quickly to the MLE $\hat{\theta} = 0.9034$. But starting from $\theta = 0.3, 0.6$, Newton puts the guess out of bounds and the algorithm fails.

(b,c) At the maximum, the Laplace approximation to the log-likelihood is the same as the Newton approximation:

$$l(\theta) \approx l(\hat{\theta}) + \frac{1}{2}\left(\left.\frac{d^2 l(\theta)}{d\theta^2}\right|_{\theta=\hat{\theta}} \right)(\theta - \hat{\theta})^2 \quad (30)$$

The EM bound is

$$\begin{aligned} q &= \frac{\hat{\theta}}{2+\hat{\theta}} && (31) \\ g(\theta) &= y_1 q \log(\theta/q) + y_1(1-q)\log(2/(1-q)) + (y_2+y_3)\log(1-\theta) + y_4\log\theta && (32) \end{aligned}$$

These are exponentiated and plotted along with the likelihood in figure 3. The EM bound is clearly a better approximation, which motivates the use of EM bounds for approximating posterior expectations (i.e. variational methods).

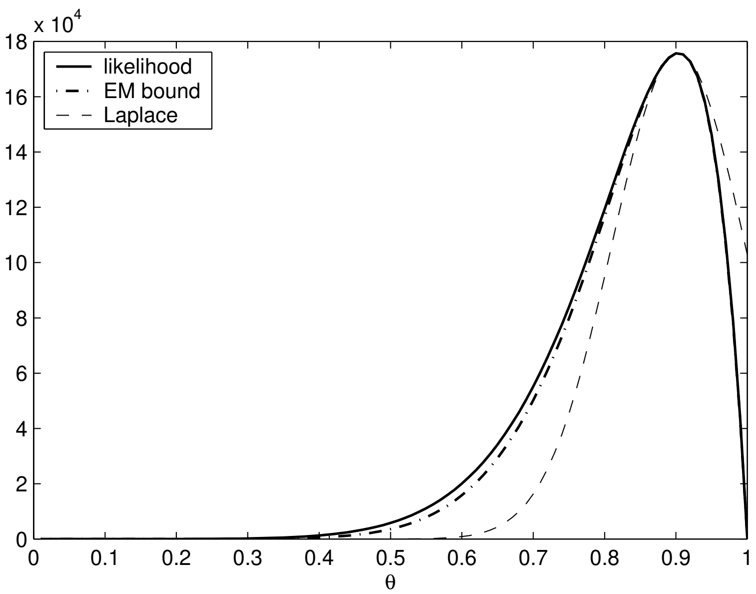

Figure 3: The likelihood along with the Laplace approximation and the final EM bound.