4.3.1) ~~mean m~~ Variance v of Beta distribution $= \dfrac{ab}{(a+b)^2(a+b+1)}$

mean m of Beta distribution $= \dfrac{a}{a+b}$.

~~Let~~ $1-m = 1 - \dfrac{a}{a+b} = \dfrac{a+b-a}{a+b} = \dfrac{b}{a+b}$

Let $\eta = a+b$

$\Rightarrow \quad v = \dfrac{m(1-m)}{(\eta+1)}$ & $m = \dfrac{a}{a+b}$.

Solving for a & b we get:

$a = m\eta$ & $b = (1-m)\eta$.

$\Rightarrow \text{Beta}(a,b) = \dfrac{1}{B(a,b)} \theta_i^{a-1}(1-\theta_i)^{b-1}$

$$\boxed{\Rightarrow \text{Beta}(m,\eta) = \frac{1}{B(m\eta,(1-m)\eta)} \theta_i^{m\eta-1}(1-\theta_i)^{(1-m)\eta-1}}$$

4.3.2 m lies in the range $(0,1)$ and Beta distribution is defined on $[0,1]$ hence Beta distribution

v varies in the range $(0,\infty)$, on which Gamma distribution is defined. Hence we chose Gamma Prior for v.

4.3.3)

Approximating the distribution by ~~a~~ point mass at mode.

$$p(\theta_1, \dots \theta_d | D) \simeq p(\theta_1, \dots \theta_d | m_{MAP}, v_{MAP}).$$

without Approximation posterior distribution,

$$\boxed{p(\theta_1 \dots \theta_d | D) \simeq p(\theta_1, \dots \theta_d | m_{MAP}, v_{MAP})}$$

Assignment 1

2.1 $P(D|\theta) = P((H,H,T)|\theta)$

$= P(H) \cdot P(H) \cdot P(T)$

$= \boxed{\theta^2 (1-\theta)}$

2.2) Likelihood of seeing 2 Heads and a Tail as calculated above is :-

$\theta^2 (1-\theta)$

→ No. of ways of obtaining 2 Heads and a Tail are

H H T

H T H ⇒ Probability of 2 Heads and a Tail is $3 \times \theta^2 (1-\theta)$

T H H

$= \boxed{3\,\theta^2 (1-\theta)}$

2.3) $P(D|\theta) = \underbrace{p(H) \cdots p(H)}_{n_h \text{ times}} \times \underbrace{p(T) \times \cdots}_{n_t \text{ times}}$

$= \boxed{\theta^{n_h} \cdot (1-\theta)^{n_t}}$

2.4) Likelihood of observed sequence $= p(D|\theta)$
$\theta^{n_h}(1-\theta)^{n_t}$

To find $\hat{\theta}_{MLE}$.

$$\frac{d\,p(D|\theta)}{d\theta} = 0$$

$$\Rightarrow n_h \theta^{n_h-1}(1-\theta)^{n_t} + (-1)\times n_t(1-\theta)^{n_t-1}\times\theta^{n_h} = 0$$

$$\Rightarrow n_h \cancel{\theta^{n_h-1}}(1-\theta)^{\cancel{n_t}} = n_t \cancel{(1-\theta)}^{\cancel{n_t-1}} \times \theta^{\cancel{n_h}}$$

$$\Rightarrow n_h(1-\theta) = n_t(\theta) \Rightarrow n_h = \theta(n_h+n_t)$$

$$\Rightarrow \boxed{\theta = \frac{n_h}{n_h+n_t}}$$

$$\Rightarrow \boxed{\tilde{\theta}_{MLE} = \frac{n_h}{n_h+n_t}}$$

3.1) posterior $\propto$ likelihood $\times$ prior

$$= p(D|\theta) \times p(\theta)$$

$$= \theta^{h-1}(1-\theta)^{t-1} \times \theta^{n_h}(1-\theta)^{n_t}$$

$$= \theta^{n_h+h-1}(1-\theta)^{n_t+t-1}$$

$\Rightarrow$ posterior is a Beta distribution with parameter $(h+n_h, t+n_t)$

$\Rightarrow$ Posterior $P(\theta|D) \sim Beta(n_h+h, t+n_t)$.

3.2)

$$\boxed{\hat{\theta}_{MLE} = \frac{n_h}{n_h+n_t}}$$

$$\hat{\theta}_{MAP} = \frac{h+n_h-1}{h+n_h+t+n_t-2}$$ [Mode of posterior distribution].

$$\hat{\theta}_{Posterior\ Mean} = \frac{h+n_h}{h+n_h+t+n_t}$$

3.3) They converge to θ, the actual value of probability of Head. As we get more data, the effect of data on the posterior increases and the effect exerted by prior decreases.

3.4) MLE gives an unbiased estimate of θ. MAP and posterior mean are biased owing to the prior which assumes a distribution

3.5) MLE. since we have small data, MAP & posterior MODE would give an estimate biased under our prior. Since the coin is fair MLE is our best bet.

(1- probability of click)

4.1) $p(D_i|\theta_i) = (\text{prob. of click})^{\text{no. of click}} \times (\text{prob. of non-click})^{\text{no. of non-click}}$

$$= (\theta_i)^{x_i} \times (1-\theta_i)^{n_i - x_i}$$

4.2) Sum of prob. of diff. values of $\theta_i = 1$.

$$\Rightarrow \int p(\theta_i) = 1$$

$$\Rightarrow \int \frac{1}{B(a,b)} \cdot \theta_i^{a-1} \cdot (1-\theta_i)^{b-1} \, d\theta_i = 1$$

$$\Rightarrow \boxed{\int \theta_i^{a-1} (1-\theta_i)^{b-1} \, d\theta_i = B(a,b)} \; *$$

4.3) $$p(\theta_i | D_i) = \frac{p(\theta_i) \times p(D_i|\theta_i)}{p(D_i)}$$

$$= \frac{1}{B(a,b) \times p(D_i)} \times \theta_i^{a-1} \times (1-\theta_i)^{b-1} \times \theta_i^{x_i} \times (1-\theta_i)^{n_i - x_i}$$

$$= \frac{1}{B(a,b) \times p(D_i)} \times \theta_i^{a+x_i-1} (1-\theta_i)^{b+n_i-x_i-1}$$

$\Rightarrow \sim \text{Beta}(a+x_i,\ b+n_i-x_i).$

$\Rightarrow p_i$: density must integrate to 1

$$\boxed{B(a,b) \times p(D_i) = B(a+x_i,\ b+n_i-x_i)}$$

$\hookrightarrow$ by definition of $B(a+x_i, b+n_i-x_i)$

4.4) From above Expression,

~~$p(\theta_i$~~

$$B(a,b) \times p(D_i) = B(a+x_i,\ b+n_i-x_i).$$

$$\Rightarrow \boxed{p(D_i) = \frac{B(a+x_i,\ b+n_i-x_i)}{B(a,b)}.}$$

4.5) $\hat{\theta}_i$ MLE maximizes the value of ~~$p(D_i)$~~ $p(D_i \mid \theta_i)$. for any given value of θ_i.

$P(D_i)$ is weighted average of $p(D_i \mid \theta_i)$ for different values of $\theta_i \in [0,1]$ where weights are $p(\theta_i)$.

The weighted average $p(D_i)$ ~~takes~~ assumes maximum value when, weight of maximum component is 1 and 0 otherwise which is the case for $p_{MLE}(D_i)$.

Hence $p_{MLE}(D_i)$ is larger than $p(D_i)$ for any other prior we put on θ_i.

4.2.1) $p(D \mid a, b) = p(D_1) \cdot p(D_2) \cdots p(D_d)$

$$= \prod_{i=1}^{d} p(D_i)$$

$$= \prod_{i=1}^{d} \frac{B(a + x_i, b + n_i - x_i)}{B(a, b)}$$

4.2.2) The dataset for i^{th} app D_i depends only on CTR of the i^{th} app.
Also, the CTR for i^{th} app θ_i is independent of CTR for anyother app.

$\Rightarrow$ D_i's are independent of each other

$\Rightarrow$ information about one app does not help with information about other apps

$\Rightarrow$ $p(\theta_i \mid D) = p(\theta_i \mid D_i)$.
(posterior θ_i is influenced only by D_i data for app i.)

4.1.6) ~~f~~ As seen above $p(D_i)$ can be interpreted as the weighted average of $p(D_i | \theta_i)$, where weights are $p(\theta_i)$.

- Also $p(D_i | \theta_i = x_i/n_i) \geq p(D_i | \theta_i)$ $\forall\ \theta_i \in (0,1)$.

- Therefore, the higher concenteration of ~~the~~ prior is around $\theta_i = x_i/n_i$, higher will $p(D_i)$ be.

- Also, we can have Beta distributions with expected value x_i/n_i (i.e concentrated around x_i/n_i), and of ~~diminishing~~ diminishing variance, i.e (greater concentration around x_i/n_i).
  ↳ Property of Beta distribution

- Therefore we can keep increasing the Likelihood without bounds.

Hence Maximum Likelihood cannot be used.

4.2.3 ~~poster~~ posterior for app i : $p(\theta_i | D) = p(\theta_i | \ldots$

posterior ~ $Beta(a + x_i, b + n_i - x_i)$

$\Rightarrow$ posterior Mean $= \dfrac{a + x_i}{a + x_i + b + n_i - x_i} = \boxed{\dfrac{a + x_i}{a + b + n_i}}$

$$MAP = \frac{a + x_i - 1}{(a + x_i) + (b + n_i - x_i) - 2} = \frac{a + x_i - 1}{(a + b + n_i - 2)}$$

$$= \boxed{\frac{a + x_i - 1}{a + b + n_i - 2}}$$

posterior SD $= \sqrt{Var}$

$$Var = \frac{ab}{(a+b)^2 (a + b + 1)} = \frac{(a + x_i)(b + n_i - x_i)}{(a + b + n_i)^2 (a + b + n_i + 1)}$$

$\Rightarrow$ For App 1

$$MAP = \frac{6.47 + 50 - 1}{(6.47 + 50) + (1181.4 + \cancel{6.47} + 10000 - 50) - 2}$$

$= 0.49\%$

posterior mean $= \dfrac{6.47+50}{56.47+1181.4+10000-50}$

$= 0.0050474$

$= 0.50474\%$

~~posterior SD =~~ $\cancel{\sqrt{6.47 \times 50\ 1181.4}}$

$SD = \left(\dfrac{(6.47+50) \times (1181.4+10000-50)}{(6.47+1181.4+10000)^2 (6.47+1181.4+10000+1)} \right)^{1/2}$

$= 0.00066699$

$= 0.06699\%$

4.2.4

	MAP	Posterior Mean	Posterior SD
App 1	0.5%	0.49%	0.07%
App 2	0.78%	0.79%	0.06%
App 3	0.3%	0.3%	0.02%
App 4	0.42%	0.5%	0.19%
App 5	0.459%	0.54%	0.21%
App 6	0.544%	0.62%	0.22%