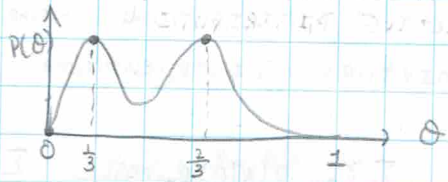
Lecture 19

Suppose $x|\theta \sim \text{Bin}(n,\theta)$ & $\theta \sim \text{Beta}(\alpha,\beta)$. We know that this implies ① $\theta|x \sim \text{Beta}(x+\alpha,\ n-x+\beta)$ &
② $X \sim \text{BetaBin}(n,\alpha,\beta)$

What if instead, we had $x|\theta \sim \text{Bin}(n,\theta)$ & $\theta \not\sim \text{Beta}(\alpha,\beta)$.
I.e. if



Clearly $\theta \not\sim \text{Beta}(\alpha,\beta)$ since Beta is unimodal & this is bimodal.

What would be our posterior? Well

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

So $k(\theta|x) = p(x|\theta)p(\theta)$. Hence

① we can sample from $\theta \in \Theta_G$ OR
② get $k(\theta_g|x)$ via <u>grid sampling</u>
③ use a "TYPE OF CONJUGACY"

Let us explore ③.

Assume $p(\theta) \approx \gamma_1 \text{Beta}(\alpha_1,\beta_1) + \gamma_2 \text{Beta}(\alpha_2,\beta_2)$ where $\gamma_1 + \gamma_2 = 1$

Suppose, in particular,
$$p(\theta) = \tfrac{1}{2}\text{Beta}(3,3) + \tfrac{1}{2}\text{Beta}(2,4)$$

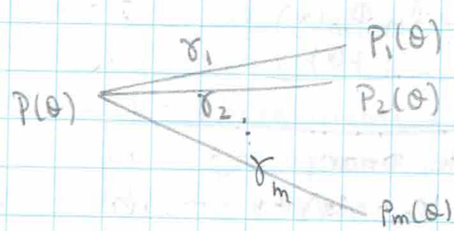
} CONJUGATE MIXTURE PRIOR

In general a CONJUGATE MIXTURE PRIOR looks like

$$p(\theta) = \sum_{m=1}^{M} \gamma_m p_m(\theta) \quad \text{s.t.} \quad \sum \gamma_m = 1 \ \&$$

$p_m(\theta)$ is a <u>density</u>.

Visually we see $p(\theta)$ is broken down into M parts:



We can think of each γ_i as our <u>PRIOR WEIGHTS</u>.

If we have a CONJUGATE MIXTURE PRIOR then

$$p(x) = \int_\theta p(x,\theta)\, d\theta = \int_\theta p(x|\theta) p(\theta)\, d\theta$$

$$= \int_\theta p(x|\theta) \sum \gamma_m p_m(\theta)\, d\theta$$

$$= \int_\theta p(x|\theta) \left(\gamma_1 p_1(\theta) + \gamma_2 p_2(\theta) + \dots + \gamma_m p_m(\theta) \right) d\theta = \sum \gamma_m \underbrace{\int_\theta p(x|\theta) p_m(\theta)\, d\theta}_{p_m(x)}$$

$= \sum \gamma_m p_m(x) \rightarrow$ this is the PRIOR PREDICTIVE DISTRIBUTION UNDER THAT SINGLE PRIOR IN THE MIXTURE DISTRIBUTION.

Hence,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)\sum \gamma_m p_m(\theta)}{p(x)} = \frac{\sum \gamma_m p(x|\theta) p_m(\theta)}{p(x)} = \frac{\sum \gamma_m p(x|\theta)(1) p_m(\theta)}{p(x)}$$

$$= \frac{\sum \gamma_m p(x|\theta) \frac{p_m(x)}{p_m(x)} p_m(\theta)}{p(x)} = \frac{\sum \gamma_m \frac{p(x|\theta)p_m(\theta)}{p_m(x)} p_m(x)}{p(x)}$$

$$= \frac{\sum \gamma_m p_m(\theta|x) p_m(x)}{p(x)} = \frac{\sum \gamma_m p_m(x)}{p(x)} \cdot p_m(\theta|x)$$

Now $p(x)$ & $p_m(x)$ is a constant so $= \boxed{\sum \gamma_m' \; p_m(\theta|x)}$

Thm: If $p(\theta) = \sum \gamma_m p_m(\theta)$ then

$$p(\theta|x) = \sum \gamma_m' p_m(\theta|x)$$

where γ_i is your prior weight on mixture model i, γ_i' is your posterior weight on model i, $p_m(\theta|x)$ is your posterior model per single mixture & where $\sum_{m=1}^{M} \frac{\gamma_m p_m(x)}{p(x)}$

Hence this looks CONJUGATE. The prior & posterior are of the same form.

Thm: For the mixture Model conjugate prior,

$$p(\theta|x) \propto \sum \gamma_m p_m(x) \cdot p_m(\theta|x)$$

Proof: follows straight from $p(\theta|x) = \sum \gamma_m' p_m(\theta|x)$

Thm: In the special case all prior weights are equal, i.e. $\gamma_m = \frac{1}{M}$ then $p(\theta|x) = \sum \frac{p_m(x)}{p_1(x) + \dots + p_M(x)} \cdot p_m(\theta|x)$

$$\propto \sum p_m(x) p_m(\theta|x)$$

With these theorems, let us do an example: Suppose $X \sim \text{Bin}(n, \theta)$ for a fixed n & suppose $\theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$, $\theta_2 \sim \text{Beta}(\alpha_2, \beta_2)$, $\gamma_1 = \frac{1}{2} = \gamma_2$.

Then $p(\theta) = \sum \gamma_m p_m(\theta)$

$$= \tfrac{1}{2}\text{Beta}(\alpha_1, \beta_1) + \tfrac{1}{2}\text{Beta}(\alpha_1, \beta_1) \quad \&$$

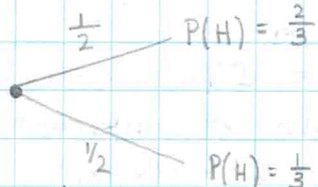
$$p_1(x) = \text{BetaBin}(n, \alpha_1, \beta_1), \quad p_1(\theta|x) = \text{Beta}(\alpha_1 + x, \beta_1 + (n-x))$$

$$p_2(x) = \text{BetaBin}(n, \alpha_2, \beta_2), \quad p_2(\theta|x) = \text{Beta}(\alpha_2 + x, \beta_2 + (n-x))$$

Thus $$p(\theta|x) = \frac{\sum p_m(x) \cdot p_m(\theta|x)}{p_1(x) + p_2(x)} = \frac{p_1(x)\,p_1(\theta|x) + p_2(x)\,p_2(\theta|x)}{p_1(x) + p_2(x)}$$

$$\boxed{= \frac{\text{BetaBin}(n, \alpha_1, \beta_1)\,\text{Beta}(\alpha_1 + x, \beta_1 + (n-x)) + \text{BetaBin}(n, \alpha_2, \beta_2)\,\text{Beta}(\alpha_2 + x, \beta_2 + (n-x))}{\text{BetaBin}(n, \alpha_1, \beta_1) + \text{BetaBin}(n, \alpha_2, \beta_2)}}$$

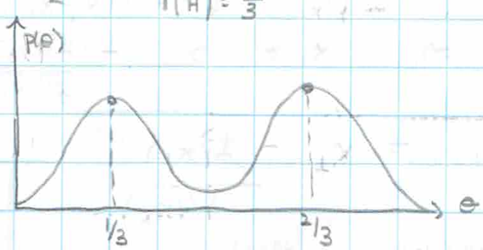
Now suppose a magician shaves off any side of a coin w.p. $\frac{1}{2}$. He then spins the coin. So we have



AND SUPPOSE FROM OUR PRIOR KNOWLEDGE,

$p(\theta) \approx \frac{1}{2}\text{Beta}(10, 20) + \frac{1}{2}\text{Beta}(20, 10)$

Then $p(\theta)$ looks like



Suppose we spin a coin 10 times & get 3 heads. What is the posterior?

Recall in R,

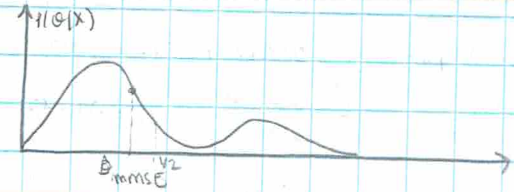
```
x = rbetabin(n, α, β)
p = pbetabinom(x, n, α, β)
x = qbetabinom(p, n, α, β)
d = betabin(x, n, α, β)    ← for discrete distributions this is p(x)
```

using these commands & the formula presented above becomes

$$p(\theta|x) = p(\theta|x=3) = \frac{\text{dbb}(3,10,10,20)\,\text{Beta}(13,27) + \text{dbb}(3,10,20,10)\,\text{Beta}(23,17)}{\text{dbb}(3,10,10,20) + \text{dbb}(3,10,20,10)}$$

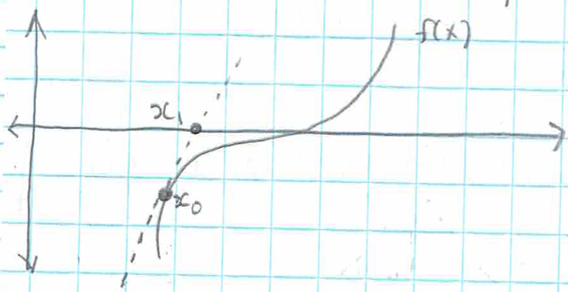
$$= .892\,\text{Beta}(13,27) + .108\,\text{Beta}(23,17)$$

Plotting yields



* which is STILL BIMODAL b/c the data has yet to overpower the prior *

We now take a small detour: Suppose you have a function $f(x)$ s.t. f is differentiable & you want to find $\tilde{x}$ s.t. $f(\tilde{x}) = 0$. (i.e. you want to find a ROOT) How do you do this? Here is an algorithm:



1. Guess where the solution is & call your guess x_0
2. Draw tangent line from x_0
3. Find the x-intercept of x_0
4. Repeat ①-③ until $|x_{t+1} - x_t| < \varepsilon$ for ε a predefined tolerance level.

NOTE IN STEP ② WE ARE DRAWING A LINE through the point $(x_0, f(x_0))$ (our notation here doesn't agree) By the point slope formula, this is the line

$$y - f(x_0) = f'(x_0)(x - x_0)$$

Now in step ③ we are getting the x-intercept, i.e. the value for where $y = 0$. Hence

$$-f(x_0) = f'(x_0)(x_1 - x_0)$$

$$\Rightarrow \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Generally we have

$$\boxed{x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}}$$

NEWTON-RAPHSON ALGORITHM. This is an example of an ITERATIVE NUMERICAL ALGORITHM

NOTE: YOU MUST CHOOSE x_0 & CHOOSE AN ε (SO YOU ENTER, x_0, ε & YOUR FUNCTION)

We will soon use this result.

We know by the Theorem which is boxed in on PAGE 2 that if $p(\theta) = \sum \gamma_m p_m(\theta)$ then $p(\theta|x) = \sum \gamma_m' p_m(\theta|x)$.
What is $\hat{\theta}_{mmse}$?

Thm: $\hat{\theta}_{mmse} = \int_\theta \theta p(\theta|x)\,d\theta = \int_\theta \theta \sum \gamma_m' p_m(\theta|x)\,d\theta$

$= \sum \gamma_m' \int_\theta \theta p_m(\theta|x)\,d\theta = \sum \gamma_m' E_m[\theta|x]$ so

$$\boxed{\hat{\theta}_{mmse} = E[\theta|x] = \sum \gamma_m' E_m[\theta|x]}$$

This is a very nice result which matches our intuition! The expected value of $\theta|x$, or the expected value of the sum $\sum \gamma_m' p_m(\theta|x)$, is the sum of $\hat{\theta}_p^{mmse}$ times the posterior weight γ_m' where $\hat{\theta}_p^{mmse} = E_p[\theta|x]$, i.e. the expected value of $p_m(\theta|x)$.

As an example consider the $n=10$, $x=3$ heads example. We found $p(\theta|x) = p(\theta|x=3) = .892\,\text{Beta}(13,27) + .108\,\text{Beta}(23,17)$.

Recall for $Y \sim \text{Beta}(\alpha,\beta)$ $E[Y] = \frac{\alpha}{\alpha+\beta}$

$$\hat{\theta}_{mmse} = E[\theta|x]$$
$$= .892\left(\tfrac{13}{40}\right) + .108\left(\tfrac{23}{40}\right)$$
$$= .352!$$

A more difficult question to answer is what is $\hat{\theta}_{MAP}$? Well,

$$\hat{\theta}_{MAP} = \underline{\text{argmax}}\{p(\theta|x)\} = \underline{\text{argmax}}\{k(\theta|x)\}.$$

So to find $\hat{\theta}_{MAP}$, it suffices to maximize $k(\theta|x)$. To find the max of $k(\theta|x)$ we now have 3 ways:

1) From Lecture 15, $k(\theta|x) \approx \propto$ to $\mathcal{N}\left(\hat{\theta}_{MAP}, \left(\frac{1}{\sqrt{g''(\hat{\theta}_{MAP}|x)}}\right)^2\right)$

2) From Lecture 15, $k(\theta|x)$ can grid sample.

3) We can do this the old fashioned way, take the derivative of $k(\theta|x)$ & set it equal to 0.

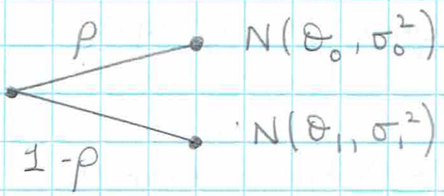
Let us do (3): $p(\theta|x) = \sum \gamma_m \binom{n}{x} \frac{B(x+\alpha_m, n-x+\beta_m)}{B(\alpha_m,\beta_m)} \cdot \frac{1}{B(x+\alpha_m, n-x+\beta_m)} \cdot \theta^{x+\alpha_m-1}(1-\theta)^{n-x+\beta_m-1}$

$$\Rightarrow k(\theta|x) = \sum \gamma_m \cdot \frac{1}{B(\alpha_m,\beta_m)} \cdot \theta^{x+\alpha_m-1}(1-\theta)^{n-x+\beta_m-1}$$

$$\Rightarrow k'(\theta|x) = \sum \frac{\gamma_m}{B(\alpha_m,\beta_m)}\left[(x+\alpha_m-1)\theta^{x+\alpha_m-2}(1-\theta)^{n-x+\beta_m-1} - (n-x+\beta_m-1)\theta^{x+\alpha_m-1}(1-\theta)^{n-x+\beta_m-2}\right.$$

Setting $k'(\theta|x) = 0$ & solving for θ is EXTREMELY DIFFICULT.
If only we had an algorithm to do this ..... oh wait, we do!
Let us use the Newton-Raphson Algorithm.
Remember, to use the NRA we must specify our initial guess & specify some $\varepsilon > 0$. BASED on our graph on the top of the last page, it seems $\hat{\theta}_{MAP}$ IS NEAR $\hat{\theta}_{mmse}$. So using the ε from Professor Kapeler, we enter in the computer "$\theta_0 = \hat{\theta}_{mmse} = .352$" & enter "$\theta_{t+1} = \theta_t - \frac{k'(\theta|x=3)}{k''(\theta|x=3)}$"
Note the computer also calculates the next derivative of k'.
We get: $\hat{\theta}_{MAP} = .31577 \neq \hat{\theta}_{mmse} = .352$.

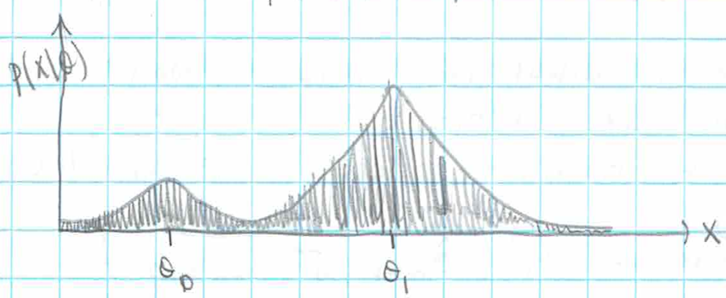
We now begin our next unit: Let us generalize our coin flipping example: Suppose we have the following mixture



Where $X_1, \ldots, X_n \overset{\text{each}}{\sim} \rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)$
$= p(x|\theta)$

So we have a bunch of r.v.'s each being a mixture r.v.

So we have $p(x|\theta) = \rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)$ looks like this:



Now $p(x|\theta)$ is really $p(x|\rho, \theta_0, \theta_1, \sigma_0^2, \sigma_1^2)$ since given θ means you are given $\rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)$. Thus we have 5 PARAMETERS!

The Bayesian Question is what is $p(\theta|x)$, i.e what is $p(\rho, \theta_0, \theta_1, \sigma_0^2, \sigma_1^2 | x)$ which is VERY, VERY, DIFFICULT. We saw how messy things were for the normal distribution, i.e. $p(\theta, \sigma^2|x)$. This now is even messier! Hence to do things like compute credible regions for θ_0 & such are difficult. A more reasonable question is what is $\hat{\theta}_{MLE}$?

That is, what is the solution to $\nabla \ln \ell(\theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho : X) = 0$
That is, what is the solution to

$$\nabla \ln \left(\prod_{i=1}^{n} \left(\frac{1}{\rho \sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(x_i-\theta_0)^2} + \frac{1}{(1-\rho)\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)^2} \right) \right) = 0$$

We could grid search... BUT the amount of points to plug in can lead to a quadrillion plus samples which takes forever. Why? For

$\underbrace{\rho \in [0,1]}_{1000}$, $\underbrace{\theta_0 \in [\ldots]}_{1000}$, $\underbrace{\theta_1 \in [\ldots]}_{1000}$, $\underbrace{\sigma_0^2 \in [\ldots]}_{1000}$, $\underbrace{\sigma_1^2 \in [\ldots]}_{1000}$

1000^5 samples... impossible for a computer to sample! So how do we sample?
STAY TUNED FOR THE NEXT LECTURE.