Lecture 10 - More Gibbs sampling

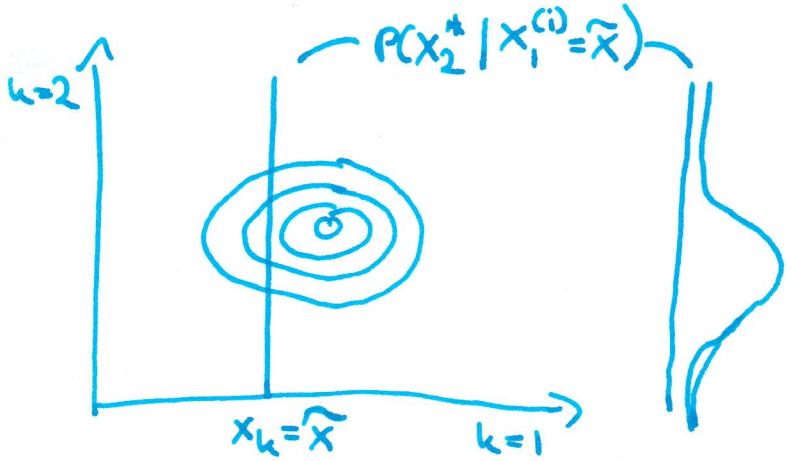
Let's prove that Gibbs sampling converges to our target distribution.

Notation: $x_{-k}^{(i)}$ all components of $x^{(i)}$ except k

$x_k^{(i)}$ k^{th} component of $x^{(i)}$

We said that Gibbs sampling is a form of Metropolis-Hastings where we always accept the proposed new position, and this new position is sampled from the conditional:

$$q_k(x^* \mid x^{(i)}) = \begin{cases} P(x_k^* \mid x_{-k}^{(i)}) & \text{for } x_{-k}^* = x_{-k}^{(i)}, \\ 0 & \text{otherwise} \end{cases}$$



For the reverse direction we have:

$$q_k(x^{(i)} \mid x^*) = \begin{cases} p(x_k^{(i)} \mid x_{-k}^*) & \text{for } x_{-k}^{(i)} = x_{-k}^* \\ 0 & \text{otherwise} \end{cases}$$

Lets look at our M-H acceptance probability:

$$A = \min\left\{1, \frac{p(x^*)}{p(x^{(i)})} \cdot \frac{q_k(x^{(i)} \mid x^*)}{q_k(x^* \mid x^{(i)})}\right\}$$

we know that: $p(x^*) = p(x_{-k}^*, x_k^*)$

$$= p(x_{-k}^* \mid x_k^*) \cdot p(x_k^*)$$

$$= p(x_k^* \mid x_{-k}^*) \cdot p(x_{-k}^*)$$

$$\Rightarrow A = \min\left\{1, \frac{p(x_k^* \mid x_{-k}^*) \cdot p(x_{-k}^*) \cdot q_k(x^{(i)} \mid x^*)}{p(x_k^{(i)} \mid x_{-k}^{(i)}) \cdot p(x_{-k}^{(i)}) \cdot q_k(x^* \mid x^{(i)})}\right\}$$

Now we put in the definition of

$q_k(x^{(i)} \mid x^*)$ and $q_k(x^* \mid x^{(i)})$

$$A = \min\left\{1, \frac{p(x_k^* \mid x_{-k}^*) \cdot p(x_{-k}^*) \cdot p(x_k^{(i)} \mid x_{-k}^*)}{p(x_k^{(i)} \mid x_{-k}^{(i)}) \cdot p(x_{-k}^{(i)}) \cdot p(x_k^* \mid x_{-k}^{(i)})}\right\}$$

Next we realize that $x_{-k}^{(i)} = x_{-k}^*$ because the new position for x^* is computed componentwise, so x^* differs from $x^{(i)}$ only in the new k^{th} component.

$$\frac{p(x_k^* \mid x_{-k}^{(i)}) \cdot p(x_{-k}^{(i)}) \cdot p(x_k^{(i)} \mid x_{-k}^{(i)})}{p(x_k^{(i)} \mid x_{-k}^{(i)}) \cdot p(x_{-k}^{(i)}) \cdot p(x_k^* \mid x_{-k}^{(i)})} = 1$$

$$\Rightarrow A = \min\{1, 1\} = 1$$

So we can rest assured that Gibbs sampling is a form of Metropolis-Hastings with acceptance probability 1.

Rat tumor example

- We are conserned about the tumor rates of rats in control group studies.
- Θ is the probability that a rat receiving no treatment develops a tumor.
- Data: $n = 14$ rats $\quad y = 4$ rats develop tumor

$\Rightarrow$ We can solve this analog to the infected people in a city example.

Naive estimate: $\frac{4}{14} = 0.286$

Bayes estimate:

Likelihood: $P(y_i|\Theta_i, n_i) = \text{Binomial}(n_i, y_i, \Theta_i)$

$$= \binom{n_i}{y_i} \cdot \Theta_i^{y_i} \cdot (1-\Theta_i)^{n_i - y_i}$$

Conjugate prior: $P(\Theta_i) = \text{Beta}(\Theta_i, \alpha, \beta)$

$$= \frac{\Theta_i^{\alpha-1}(1-\Theta)^{\beta-1}}{B(\alpha,\beta)}$$

↳ Beta function

To solve this we need to choose α and β

⇒ In the beginning of the lecture we needed an expert to give us values.

⇒ This time we look for more data instead.

⇒ We find 70 studies in the literature

Now our data looks like this:

$n: [n_1, n_2, \dots, n_{70}]$

$y: [y_1, y_2, \dots, y_{70}]$

⇒ We are saving our new data point for next week.

What can we do with this data?

Likelihood: $p(y|\theta, n) = \prod_i p(y_i|\theta_i, n_i)$

We need to think a bit more about θ_i, α and β.

We have three choices:

- Complete pooling
- unpooling
- partial pooling

Complete pooling:

- We assume there is only one tumor rate and $\theta_i = \hat{\theta}$, $\alpha_i = \hat{\alpha}$, $\beta_i = \hat{\beta}$.

- This is likely to underfit the data, as we have experiments from different labs, times, and rat lineages.

Unpooling:

- We are completely flexible, allowing different θ_i, α_i and β_i for each experiment:

$$\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$$

- These are a lot of parameters!
- We are likely to overfit with this model.

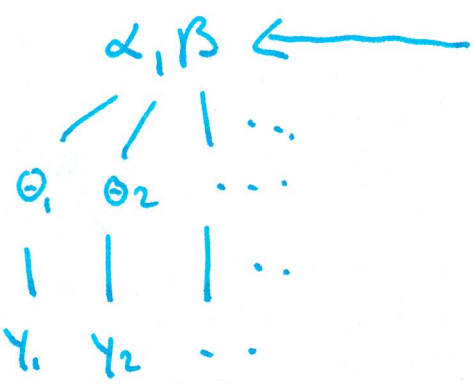
Partial pooling:

- While conditions vary a bit, all experiments still come labs with similar conditions, and all experiments are from rats.

⇒ We allow different θ_i for each experiment, but we assume they all come from the same distribution:

$$\theta_{gi} \sim \text{Beta}(\alpha, \beta)$$

⇒ This is a hierarchical Bayes model. The Bioassay was a simplified version of this where the relationship between θ and α, β was deterministic.



Hyper parameters: parameters that control the parameters of our model.

⇒ We now need to find 70 tumor rates θ_i and α, β

⇒ We have a 72 dimensional problem.

Our complete posterior:

$$P(\theta, \alpha, \beta \mid Y, n) \propto \underbrace{P(Y \mid \theta, \alpha, \beta, n)}_{\text{Likelihood}} \cdot \underbrace{P(\theta, \alpha, \beta)}_{\text{Prior}}$$

We already have our likelihood as a Binomial so we are back to thinking about the prior

Prior:

$$P(\theta, \alpha, \beta) = P(\alpha, \beta) \cdot \prod_i P(\theta_i \mid \alpha, \beta)$$

$$\hookrightarrow \text{Beta}(\theta_i, \alpha, \beta)$$

What about $p(\alpha, \beta)$?

- In the past we often just made it uniform
- In hierarchical models this can be tricky
- For large values of α and β our posterior goes to infinity and becomes non-integratable.

Let's think about α and β in a more intuitive way:

The mean of the Beta distribution is:

$$\mu = \frac{\alpha}{\alpha + \beta}$$

- It makes sense to have a uniform prior on the mean tumor rate, which corresponds to μ.

- In addition we put a uniform prior on

$$\nu = (\alpha+\beta)^{-\frac{1}{2}}$$

where $\alpha+\beta$ can be understood as our sample size.

- Note: to see this look at $\alpha+\beta$ for the Beta and the result when using it as a conjugate prior for our Binomial:

Before	After
$\text{Beta}(\alpha,\beta)$	$\text{Beta}(\alpha+y, \beta+N-y)$
$\alpha+\beta$	$\rightarrow \quad \alpha+y+\beta+N-y = \alpha+\beta+N$

- What do these uniform priors on μ and ν mean for α and β?

$$\mu = \frac{\alpha}{\alpha+\beta} \qquad \nu = \frac{1}{\sqrt{\alpha+\beta}} \qquad \nu^2 = \frac{1}{\alpha+\beta}$$

For α we have:

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\Leftrightarrow \quad \frac{\mu}{\alpha} = \frac{1}{\alpha+\beta}$$

$$\Leftrightarrow \frac{\mu}{\alpha} = v^2$$

$$\Leftrightarrow \alpha = \frac{\mu}{v^2}$$

For β we have:

$$\frac{1}{v^2} = \alpha + \beta \quad \Leftrightarrow \quad \beta = \frac{1}{v^2} - \alpha \quad \Leftrightarrow \quad \beta = \frac{1}{v^2} - \frac{\mu}{v^2}$$

$$\Leftrightarrow \beta = \frac{1-\mu}{v^2}$$

So we got our inverse transform now, the mapping from $\mu, v \to \alpha, \beta$.

We need the Jacobian to get the factor by which this mapping expands or shrinks volumes.

Note: Wikipedia has really great articles for non-Physicists who would like to understand this in more depth.

The Jacobian matrix for the inverse transform is:

$$\begin{vmatrix} \frac{\partial \alpha}{\partial \mu} & \frac{\partial \alpha}{\partial v} \\ \frac{\partial \beta}{\partial \mu} & \frac{\partial \beta}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{v^2} & -2\frac{\mu}{v^3} \\ -\frac{1}{v^2} & \frac{2(\mu-1)}{v^3} \end{vmatrix}$$

For the determinant we get:

$$\frac{\partial \alpha}{\partial \mu} \cdot \frac{\partial \beta}{\partial \nu} - \frac{\partial \alpha}{\partial \nu} \cdot \frac{\partial \beta}{\partial \mu} = \frac{2(\mu-1)}{\nu^5} - \frac{2\mu}{\nu^5} = -\frac{2}{\nu^5}$$

- For us the relevant factor is the magnitude of the Jacobian, which is dominated by the denominator; so we go with $\frac{1}{\nu^5}$

- We defined ν as $(\alpha+\beta)^{-\frac{1}{2}}$

⇒ So finally for our prior we have:

$$P(\alpha, \beta) \sim (\alpha+\beta)^{-5/2}$$

⇒ We have everything together for our posterior now $\forall_o$!

⇒ We can sample from this using M-H.

⇒ We are in 72 dimensions, it is going to be painful.

⇒ If we can get the conditionals we can do Gibbs sampling!

Reminder: Our posterior is

$$P(\theta, \alpha, \beta \mid Y, n) \propto P(\alpha, \beta) \cdot \prod_i P(\theta_i \mid \alpha, \beta) \cdot \prod_i P(Y_i \mid \theta_i, \alpha, \beta, n_i)$$

$P(\theta_i \mid Y_i, n_i, \alpha, \beta)$ is easy!

$\Rightarrow$ $P(\alpha, \beta)$ is just a number for given α and β, and we have our conjugate prior setup with a Binomial and a Beta, so we get another Beta.

$\Rightarrow P(\theta_i \mid Y_i, n_i, \alpha, \beta) \propto \text{Beta}(\alpha + Y_i, \beta + n_i - Y_i)$

We still need the conditionals to sample α and β.

It is not easy, but doable:

$$\underbrace{\frac{\theta_i^{\alpha-1} \cdot (1-\theta_i)^{\beta-1}}{B(\alpha, \beta)}}_{\text{Beta}} \cdot \underbrace{\theta_i^{Y_i} \cdot (1-\theta_i)^{n_i - Y_i}}_{\substack{\text{Binomial} \\ \text{(dropped normalization)}}}$$

On Wikipedia we read:

$$\Gamma(\alpha) + \Gamma(\beta) = \Gamma(\alpha + \beta) \cdot B(\alpha, \beta)$$

$$\Rightarrow B(\alpha,\beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

This gives us:

$$\frac{\Gamma(\alpha+\beta)\cdot\theta_i^{\alpha-y_i}\cdot(1-\theta_i)^{\beta-1+n_i-y_i}}{\Gamma(\alpha)\cdot\Gamma(\beta)}$$

Conditioning on all except α we get:

$$P(\alpha \mid Y, n, \theta, \beta) \propto P(\alpha,\beta)\cdot\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\right)^N\cdot\prod_i\theta_i^{\alpha}$$

And for β:

$$P(\beta \mid Y, n, \theta, \alpha) \propto P(\alpha,\beta)\cdot\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\right)^N\cdot\prod_i(1-\theta_i)^{\beta}$$

Sampling from these conditionals is not straight forward.

$\Rightarrow$ We can still do Gibbs sampling, using M-H to sample from the conditionals for α and β.

$\Rightarrow$ We have a nice sampling function for our 70 θ_i s.

The 71st data point:

- In the beginning we had a data point in addition to the data we got from our literature review:

$$n_{71} = 14 \qquad y_{71} = 4$$

- Let's compare how our additional data influences our estimate for θ_{71} compared to the naive estimate $\frac{4}{14}$.

$$P(\theta_{71}, \alpha, \beta \mid y, n, y_{71}, n_{71})$$

$$\propto P(\alpha, \beta) \cdot \underbrace{P(\theta_{71} \mid \alpha, \beta) \cdot P(y_{71} \mid \theta_{71}, \alpha, \beta, n_{71})}_{\text{Beta}(\alpha+4, \beta+10)}$$

$P(\alpha, \beta)$ → We have samples for this from our Gibbs sampling!