Markov Chain Monte Carlo Methods

Chapter 10. Markov Chain Monte Carlo Methods

Simulated annealing

Simulated annealing is used to explore a distribution with multiple modes, or to search the maximum value of a function.

Typical case. Consider $V(x)$ be a nonnegative function defined on A, where x could be of multivariate dimension. We want to find $V^* = \max_{x \in A} V(x)$, and $M = \{x \in A, V(x) = V^*\}$.

Idea. Let $\lambda > 0$ and consider the following density function for $x \in A$,

$$
\begin{aligned}
p_\lambda(x) &= \frac{e^{\lambda V(x)}}{\sum_{x \in A} e^{\lambda V(x)}} \\
&= \frac{e^{\lambda(V(x) - V^*)}}{|M| + \sum_{x \neq M} e^{\lambda(V(x) - V^*)}} \\
&\to \frac{\delta(x, M)}{|M|} \quad \text{as } \lambda \to \infty,
\end{aligned}
$$

where

$$
\delta(x, M) = \begin{cases} 1, & \text{if } x \in M, \\ 0, & \text{otherwise.} \end{cases}
$$

Note:

1. If λ is too large, need a very large number of transitions before limiting distribution is approached.
2. Simulated annealing considers $\lambda_n = C \log(1+n)$ for $C > 0$.
3. If we generate $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ as successive states. We estimate V^* by $\max_{i=1,\ldots,m} V(x^{(i)})$. If the maximum occurs at $x^{(i*)}$, then $x^{(i*)}$ is an estimated point in M.

Example. Consider samples $X_i \overset{iid}{\sim} N(\mu, 1)$ for $i = 1, \ldots, n$. The likelihood is

$$L(\mu|x) \propto e^{-\sum_{i=1}^{n}(X_i-\mu)^2}.$$

The MLE estimator for μ is $\hat{\mu} = \sum_{i=1}^{n} X_i/n$.

A simple way to adopt the idea of simulated annealing to search the MLE of μ is to sample $\mu \propto \pi(\mu) = e^{-\sum_{i=1}^{n}(X_i-\mu)^2}$.

The procedures to find such an estimator are summarized below.

1. Start with an arbitrary $\mu^{(0)}$. Set $i = 1$.
2. Sample $y \sim N(\mu^{(i-1)}, 0.25)$. Calculate

   $$\alpha = \min(\frac{\pi(y)}{\pi(X^{(i-1)})}, 1).$$

   Update $\mu^{(i)} = y$ with probability α and $\mu^{(i)} = \mu^{(i-1)}$ with probability $(1 - \alpha)$.
3. Return to Step 2 until convergence.
4. Set $I = \arg max\{i : \pi(\mu^{(i)})\}$. Then we set $\hat{\mu}_{SA} = \mu^{(I)}$ to estimate the MLE of μ.

Example 10k. The traveling salesman problem. A sales man starts at city 0 and then sequentially visit all the cities, $1, \ldots, r$. A permutation $x = (x_1, \ldots, x_r)$ of means a path of the sales man. $v(i,j)$ is the reward if the salesman goes from i to j. Then the return is

$$V(x) = \sum_{i=1}^{r} v(x_{i-1}, x_i),$$

where $x_0 = 0$.

Two permutations are neighbors if one results from an interchange of two of the coordinates of the other.

An algorithm to find the best path using simulated annealing with $\lambda_n = \log(1+n)$.

1. Start with an arbitrary permutation $X^{(0)}$. Set $n = 0$.
2. Randomly select I and J from $\{1, \ldots, r\}$ with probability $1/C(r,2)$, and interchange the values of the I-th and the J-th values of X_n. Let the value be y. Set $X^{(n+1)} = y$ with probability

$$\alpha(X^{(n)}, y) = \min\left(\frac{(1+n)^{V(y)}}{(1+n)^{V(X^{(n)})}}, 1\right),$$

and set $X^{(n+1)} = X^{(n)}$ with probability $1 - \alpha(X^{(n)}, y)$.

3. Set $n = n+1$, return to Step 2 until convergence.

SIR

Idea:

1. Consider a random vector X with target mass function $f(x) = C_1 f_0(x)$.
2. Conduct a Markov chain Monte Carlo algorithm with limiting mass distribution $g(x) = C_2 g_0(x)$. Let $y_1, \ldots, y_m$ be the values of the Markov chain Monte Carlo algorithm.
3. Set $w_i = \frac{f_0(y_i)}{g_0(y_i)}$ for $i = 1, \ldots, m$.
4. Generate a random vector X such that $P(X = y_j) = \frac{w_j}{\sum_{i=1}^{m} w_i}$. Then X has a mass distribution f.

Proposition. The distribution of the vector X obtained by the SIR method converges as $m \to \infty$ to f.

The proof is left as homework.

To estimate $E_f[h(X)]$, suppose we first generate $Y_1, \ldots, Y_m$ from the Markov chain with limiting mass distribution g, and generate $X_1, \ldots, X_k$ based on

$$P(X = Y_j) = \frac{W_j}{\sum_{i=1}^{m} W_i},$$

where $W_j = f_0(Y_j)/g_0(Y_j)$. We can use the following two estimators for estimating $E_f[h(X)]$:

1. $\frac{1}{k}\sum_{i=1}^{k} h(X_i)$,
2. $\frac{1}{\sum_{i=1}^{m} W_i} h(Y_j)$.

Homework: Which estimator is better? Why?

Example 10I. Bayesian inference. Let sample $X \sim F(\theta)$, where $\theta = (\theta_1, \ldots, \theta_p)$ is a p-variate vector for some distribution F.

Bayesian inference considers:

1. Prior density: $p(\theta)$,
2. Likelihood: $f(x|\theta)$,
3. Posterior density: $p(\theta|X) = \frac{f(X|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta} \propto f(x|theta)p(\theta)$.

To explore the posterior density, SIR generates θ from $p(\theta)$, and set $w(\theta) = f(x|\theta)$. That is to say, SIR sets

1. Target mass function $f_0(\theta) = C_1 f_0(\theta) = C_1 f(x|\theta)p(\theta)$.
2. Generate θ with a Makov chain with limiting mass distribution $g(\theta) = C_2 g_0(\theta) = p(\theta)$.
3. Set $w(\theta) = \frac{f_0(\theta)}{g_0(\theta)} = \frac{f(x|\theta)p(\theta)}{p(\theta)} = f(x|\theta)$.

Generate a large number m of random vectors from $p(\theta)$. Let the values be $\theta^{(1)}, \ldots, \theta^{(m)}$. Then, we can estimate $E[h(\theta)|x]$ by

$$\sum_{j=1}^{m} \alpha_j h(\theta^{(j)}), \quad \alpha_j = \frac{f(x|\theta^{(j)})}{\sum_{i=1}^{m} f(x|\theta^{(i)})},$$

and estimate $P(\theta \in A|x)$ by $\sum_{j=1}^{m} \alpha_j \mathbf{1}_{\{\theta^{(j)} \in A\}}$.

The E-M Algorithm

Dempster, Laird and Rubin (1977). *JRSSB*.

– A general approach to iterative computation of **MLE**.

Each iteration consists of

i) **E-Step**: Expectation.

ii) **M-Step**: Maximization.

Ex. 197 animals are distributed into 4 categories. Let

$\mathbf{Y} = (y_1, y_2, y_3, y_4)$, such that

$$\mathbf{Y} \sim \mathit{Multinomial}\left(\sum_{i=1}^{4} y_i; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right), 0 \le \theta \le 1.$$

For observed $\mathbf{Y} = (125, 18, 20, 34)$, MLE of $\theta = ?$

We want to maximize

$$l(\theta|\mathbf{y}) \propto \left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1}{4}(1-\theta)\right)^{y_2}\left(\frac{1}{4}(1-\theta)\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4},$$

a polynomial in θ of degree 197, $\hat{\theta}=$???

Y indeed is the **incomplete** data.

Y indeed is the **incomplete** data.

Split y_1 into x_1 and x_2 (such that $x_1 + x_2 = y_1$) with cell probabilities $1/2$ and $\theta/4$, respectively.

$\mathbf{Y}$ indeed is the **incomplete** data.

Split y_1 into x_1 and x_2 (such that $x_1 + x_2 = y_1$) with cell probabilities $1/2$ and $\theta/4$, respectively.

Let $y_2 = x_3$, $y_3 = x_4$ and $y_4 = x_5$, then $\mathbf{X} = (x_1, x_2, x_3, x_4, x_5)$, is the **complete** data and

Y indeed is the **incomplete** data.

Split y_1 into x_1 and x_2 (such that $x_1 + x_2 = y_1$) with cell probabilities $1/2$ and $\theta/4$, respectively.

Let $y_2 = x_3$, $y_3 = x_4$ and $y_4 = x_5$, then $\mathbf{X} = (x_1, x_2, x_3, x_4, x_5)$, is the **complete** data and

$$X \sim Multinomial\left(197; 1/2, \theta/4, (1-\theta)/4, \frac{1}{4}(1-\theta), \theta/4\right).$$

Thus given the *complete* data $\mathbf{X} = (\mathbf{Y}, Z) = \mathbf{x}$,

Thus given the *complete* data $\mathbf{X} = (\mathbf{Y}, Z) = \mathbf{x}$,

$$l(\theta|\mathbf{x}) \propto (\frac{1}{2})^{x_1}(\frac{\theta}{4})^{\mathbf{x_2}}(1-\theta)^{x_3+x_4}\theta^{x_5} \propto \theta^{\mathbf{x_2}+x_5}(1-\theta)^{x_3+x_4},$$

Thus given the *complete* data $\mathbf{X} = (\mathbf{Y}, Z) = \mathbf{x}$,

$$l(\theta|\mathbf{x}) \propto (\frac{1}{2})^{x_1}(\frac{\theta}{4})^{\mathbf{x_2}}(1-\theta)^{x_3+x_4}\theta^{x_5} \propto \theta^{\mathbf{x_2}+x_5}(1-\theta)^{x_3+x_4},$$

and

$$\hat{\theta} = \frac{\mathbf{x_2} + x_5}{\mathbf{x_2} + x_3 + x_4 + x_5}.$$

Thus given the *complete* data $\mathbf{X} = (\mathbf{Y}, Z) = \mathbf{x}$,

$$l(\theta|\mathbf{x}) \propto (\frac{1}{2})^{x_1}(\frac{\theta}{4})^{\mathbf{x_2}}(1-\theta)^{x_3+x_4}\theta^{x_5} \propto \theta^{\mathbf{x_2}+x_5}(1-\theta)^{x_3+x_4},$$

and

$$\hat{\theta} = \frac{\mathbf{x_2} + x_5}{\mathbf{x_2} + x_3 + x_4 + x_5}.$$

But, what is the augmented data $x_2 = ?$

(i) **E-Step**: "Estimate" the sufficient statistics of $\mathbf{x}$ based on $\mathbf{y}$.

(i) **E-Step**: "Estimate" the sufficient statistics of **x** based on **y**.

Note that $x_1 + x_2 = y_1$, $x_2 = 0, 1, \ldots, y_1$, so

$$\mathbf{X_2|y_1} \sim \ bin(y_1, \frac{\theta/4}{\frac{1}{2}+\frac{\theta}{4}}) \equiv \mathbf{bin(y_1, \frac{\theta}{2+\theta})}.$$

Hence,

$$E(X_2|\mathbf{y}, \theta) = y_1\theta/(2+\theta).$$

(i) **E-Step**: "Estimate" the sufficient statistics of $\mathbf{x}$ based on $\mathbf{y}$.

Note that $x_1 + x_2 = y_1$, $x_2 = 0, 1, \ldots, y_1$, so

$$\mathbf{X_2}|\mathbf{y_1} \sim \; bin(y_1, \frac{\theta/4}{\frac{1}{2}+\frac{\theta}{4}}) \equiv \mathbf{bin}(\mathbf{y_1}, \frac{\theta}{\mathbf{2}+\theta}).$$

Hence,

$$E(X_2|\mathbf{y}, \theta) = y_1\theta/(2+\theta).$$

$$\therefore \; \hat{\mathbf{x}}_\mathbf{2} = \frac{y_1\theta}{2+\theta}.$$

(ii) **M-Step**: Estimate θ by the MLE as though the **estimated 'complete' data** were the observed data.

(ii) **M-Step**: Estimate θ by the MLE as though the **estimated 'complete' data** were the observed data.

$$\therefore \hat{\theta} = \frac{\hat{\mathbf{x}}_2 + x_5}{\hat{\mathbf{x}}_2 + x_3 + x_4 + x_5}.$$

Starting with $\theta^{(0)} = 0.5$, say, then cycling back and forth between (i) and (ii). i.e.

Starting with $\theta^{(0)} = 0.5$, say, then cycling back and forth between (i) and (ii). i.e.

(i) $\mathbf{x_2^{(k)}} = \dfrac{125\theta^{(\mathbf{k})}}{2+\theta^{(\mathbf{k})}}$,

Starting with $\theta^{(0)} = 0.5$, say, then cycling back and forth between (i) and (ii). i.e.

(i) $\mathbf{x_2^{(k)}} = \dfrac{125\theta^{(\mathbf{k})}}{2+\theta^{(\mathbf{k})}}$,

(ii) $\theta^{(\mathbf{k})} = \dfrac{\mathbf{x_2^{(k)}} + 34}{\mathbf{x_2^{(k)}} + 72}$, $k = 0, 1, 2, \ldots$, until $\theta^{(k+1)} \approx \theta^{(k)}$.

Starting with $\theta^{(0)} = 0.5$, say, then cycling back and forth between (i) and (ii). i.e.

(i) $\mathbf{x_2^{(k)}} = \dfrac{125\theta^{(\mathbf{k})}}{2 + \theta^{(\mathbf{k})}}$,

(ii) $\theta^{(\mathbf{k})} = \dfrac{\mathbf{x_2^{(k)}} + 34}{\mathbf{x_2^{(k)}} + 72}$, $k = 0, 1, 2, \ldots$, until $\theta^{(k+1)} \approx \theta^{(k)}$.

Answer: $\theta^{(8)} = 0.626821484$.

The solution of θ indeed is

$$\theta^\star = \frac{\frac{125\theta^\star}{2+\theta^\star} + 34}{\frac{125\theta^\star}{2+\theta^\star} + 73}$$

The solution of θ indeed is

$$\theta^\star = \frac{\frac{125\theta^\star}{2+\theta^\star} + 34}{\frac{125\theta^\star}{2+\theta^\star} + 73} \iff a{\theta^\star}^2 + b\theta^\star + c = 0.$$

The solution of θ indeed is

$$\theta^\star = \frac{\frac{125\theta^\star}{2+\theta^\star} + 34}{\frac{125\theta^\star}{2+\theta^\star} + 73} \Longleftrightarrow a{\theta^\star}^2 + b\theta^\star + c = 0.$$

$\theta^\star = 0.6268214980.$ $\square$

Formal Definition:

Formal Definition:

$\mathbf{Y} \sim f(\cdot|\theta)$ and $\mathbf{Y} = \mathbf{y}$, observed (**incomplete**) data;

Formal Definition:

$\mathbf{Y} \sim f(\cdot|\theta)$ and $\mathbf{Y} = \mathbf{y}$, observed (**incomplete**) data;

Z: **latent** (augmented) data, so that

$\mathbf{X} = (\mathbf{Y}, Z)$ is the **complete** data.

Assume $p(z|\theta, y)$ is *known*.

Formal Definition:

$\mathbf{Y} \sim f(\cdot|\theta)$ and $\mathbf{Y} = \mathbf{y}$, observed (**incomplete**) data;

Z: **latent** (augmented) data, so that

$\mathbf{X} = (\mathbf{Y}, Z)$ is the **complete** data.

Assume $p(z|\theta, y)$ is *known*.

Define

$$
\begin{aligned}
Q(\theta, \theta^i) &= \int \log l(\theta|\mathbf{y}, z) p(z|\theta^i, \mathbf{y})\, dz \\
&= E^{\mathbf{Z}|\theta^{\mathbf{i}}, \mathbf{y}} \left[\log l(\theta|\mathbf{y}, \mathbf{Z})\right].
\end{aligned}
$$

(i) **E-Step**:

(i) **E-Step**: Given θ^i and $\mathbf{y}$, compute

$$Q(\theta, \theta^i) = E^{Z|\theta^i, \mathbf{y}} \log\left[l(\theta|\mathbf{y}, Z)\right].$$

(i) **E-Step**: Given θ^i and $\mathbf{y}$, compute

$$Q(\theta, \theta^i) = E^{Z|\theta^i, \mathbf{y}} \log\left[l(\theta|\mathbf{y}, Z)\right].$$

(ii) **M-Step**:

(i) **E-Step**: Given θ^i and $\mathbf{y}$, compute

$$Q(\theta, \theta^i) = E^{Z|\theta^i, \mathbf{y}} \log\left[l(\theta|\mathbf{y}, Z)\right].$$

(ii) **M-Step**: Maximize $Q(\theta, \theta^i)$ with respect to θ, i.e. solve

(i) **E-Step**: Given θ^i and $\mathbf{y}$, compute

$$Q(\theta, \theta^i) = E^{Z|\theta^i, \mathbf{y}} \log\left[l(\theta|\mathbf{y}, Z)\right].$$

(ii) **M-Step**: Maximize $Q(\theta, \theta^i)$ with respect to θ, i.e. solve

$$\frac{\partial Q(\theta, \theta^i)}{\partial \theta} = 0$$

for θ, call it θ^{i+1}.

(i) **E-Step**: Given θ^i and $\mathbf{y}$, compute

$$Q(\theta, \theta^i) = E^{Z|\theta^i, \mathbf{y}} \log\left[l(\theta|\mathbf{y}, Z)\right].$$

(ii) **M-Step**: Maximize $Q(\theta, \theta^i)$ with respect to θ, i.e. solve

$$\frac{\partial Q(\theta, \theta^i)}{\partial \theta} = 0$$

for θ, call it θ^{i+1}.

If $|\theta^{i+1} - \theta^i|$ is **small**, stop; otherwise, return to (i).

Standard errors in EM:

$$(1)\ Var\ \hat{\theta} \approx -\left(\frac{\partial^2 \log l(\theta|\mathbf{y})}{\partial\theta^2}\Big|_{\theta=\hat{\theta}}\right)^{-1}.$$

$$(2)\ -\frac{\partial^2 \log l(\mathbf{y}|\theta)}{\partial\theta^2} =$$
$$-\int \frac{\partial^2 \log l(\theta|\mathbf{y},z)}{\partial\theta^2} p(z|\theta,\mathbf{y})\,dz\ -\ Var^Z\left[\frac{\partial \log l(\theta|\mathbf{y},Z)}{\partial\theta}\right].$$

Ex(Cont'd). $l(\theta|\mathbf{y}, z) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}$. So

<u>**Ex**</u>(Cont'd). $l(\theta|\mathbf{y}, z) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}$. So

$$\frac{\partial \log l(\theta|\mathbf{y}, z)}{\partial\theta} = \frac{x_2 + x_5}{\theta} - \frac{x_3 + x_4}{1-\theta},$$

while

Ex(Cont'd). $l(\theta|\mathbf{y}, z) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}$. So

$$\frac{\partial \log l(\theta|\mathbf{y}, z)}{\partial \theta} = \frac{x_2 + x_5}{\theta} - \frac{x_3 + x_4}{1-\theta},$$

while

$$-\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} = \frac{E(\mathbf{X_2}|\hat{\theta}, \mathbf{y}) + x_5}{\hat{\theta}^2} + \frac{x_3 + x_4}{(1-\hat{\theta})^2},$$

and

Ex(Cont'd). $l(\theta|\mathbf{y}, z) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}$. So

$$\frac{\partial \log l(\theta|\mathbf{y}, z)}{\partial \theta} = \frac{x_2+x_5}{\theta} - \frac{x_3+x_4}{1-\theta},$$

while

$$-\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta^2}|_{\theta=\hat{\theta}} = \frac{E(\mathbf{X_2}|\hat{\theta}, \mathbf{y}) + x_5}{\hat{\theta}^2} + \frac{x_3+x_4}{(1-\hat{\theta})^2},$$

and

$$Var\left(\frac{\partial \log l(\theta|\mathbf{y}, Z)}{\partial \theta}\right)|_{\theta=\hat{\theta}} =$$

Ex(Cont'd). $l(\theta|\mathbf{y}, z) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}$. So

$$\frac{\partial \log l(\theta|\mathbf{y}, z)}{\partial \theta} = \frac{x_2 + x_5}{\theta} - \frac{x_3 + x_4}{1-\theta},$$

while

$$-\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta^2}|_{\theta=\hat{\theta}} = \frac{E(\mathbf{X_2}|\hat{\theta}, \mathbf{y}) + x_5}{\hat{\theta}^2} + \frac{x_3 + x_4}{(1-\hat{\theta})^2},$$

and

$$Var\left(\frac{\partial \log l(\theta|\mathbf{y}, Z)}{\partial \theta}\right)|_{\theta=\hat{\theta}} = \frac{Var(\mathbf{X_2}|\hat{\theta}, \mathbf{y})}{\hat{\theta}^2} =$$

Ex(Cont'd). $l(\theta|\mathbf{y}, z) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}$. So

$$\frac{\partial \log l(\theta|\mathbf{y}, z)}{\partial \theta} = \frac{x_2+x_5}{\theta} - \frac{x_3+x_4}{1-\theta},$$

while

$$-\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} = \frac{E(\mathbf{X}_2|\hat{\theta}, \mathbf{y}) + x_5}{\hat{\theta}^2} + \frac{x_3+x_4}{(1-\hat{\theta})^2},$$

and

$$Var\left(\frac{\partial \log l(\theta|\mathbf{y}, Z)}{\partial \theta}\right)\Big|_{\theta=\hat{\theta}} = \frac{Var(\mathbf{X}_2|\hat{\theta}, \mathbf{y})}{\hat{\theta}^2} = \frac{y_1 \frac{\hat{\theta}}{2+\hat{\theta}} \frac{2}{2+\hat{\theta}}}{\hat{\theta}^2}.$$

$\therefore\ Var\ \hat{\theta} \approx (435.8 - 57.8)^{-1} = .05.$ $\square$

Note: 1. θ^i converge to a *stationary point* of $l(\theta|\mathbf{y})$ if the limit exists.

2. The convergence could be a **local** maximum, so multiple starting points are recommended.

Ex. (Cont'd.) **Bayesian Method**.

Ex. (Cont'd.) **Bayesian Method**.

Consider the uniform prior on θ, i.e. $\theta \sim U(0,1)$. Then the posterior of θ given the data $\mathbf{y}$ is

Ex. (Cont'd.) **Bayesian Method**.

Consider the uniform prior on θ, i.e. $\theta \sim U(0,1)$. Then the posterior of θ given the data $\mathbf{y}$ is

$$\pi(\theta|\mathbf{y}) \propto l(\theta|\mathbf{y}) \propto \left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1}{4}(1-\theta)\right)^{y_2}\left(\frac{1}{4}(1-\theta)\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}.$$

However, given the *complete* data $\mathbf{X} = (\mathbf{Y}, Z) = \mathbf{x}$,

$$\pi(\theta|\mathbf{x}) \propto (\frac{1}{2})^{x_1} (\frac{\theta}{4})^{\mathbf{x_2}} (1-\theta)^{x_3+x_4} \theta^{x_5} \propto \theta^{\mathbf{x_2}+x_5} (1-\theta)^{x_3+x_4}.$$

However, given the *complete* data $\mathbf{X} = (\mathbf{Y}, Z) = \mathbf{x}$,

$$\pi(\theta|\mathbf{x}) \propto (\frac{1}{2})^{x_1}(\frac{\theta}{4})^{\mathbf{x_2}}(1-\theta)^{x_3+x_4}\theta^{x_5} \propto \theta^{\mathbf{x_2}+x_5}(1-\theta)^{x_3+x_4}.$$

Q: How to impute the *latent variable* x_2?

Recall: Given $\mathbf{y}$, the *conditional distribution* of

x_2 given θ is

Recall: Given $\mathbf{y}$, the *conditional distribution* of

x_2 given θ is $bin(y_1, \theta/(2+\theta))$;

Recall: Given $\mathbf{y}$, the *conditional distribution* of

x_2 given θ is $bin(y_1, \theta/(2+\theta))$;

and the *conditional distribution* of

θ given x_2 is $beta(x_2 + x_5 + 1, x_3 + x_4 - 1)$.

Recall: Given $\mathbf{y}$, the *conditional distribution* of

x_2 given θ is $bin(y_1, \theta/(2+\theta))$;

and the *conditional distribution* of

θ given x_2 is $beta(x_2 + x_5 + 1, x_3 + x_4 - 1)$.

Hence, one can perform the **Gibbs sampler** by generating θ and x_2 from their **conditional posteriors** iteratively to approximate the posterior distribution of θ, namely

Recall: Given $\mathbf{y}$, the *conditional distribution* of

x_2 given θ is $bin(y_1, \theta/(2+\theta))$;

and the *conditional distribution* of

θ given x_2 is $beta(x_2 + x_5 + 1, x_3 + x_4 - 1)$.

Hence, one can perform the **Gibbs sampler** by generating θ and x_2 from their **conditional posteriors** iteratively to approximate the posterior distribution of θ, namely

$$p(\theta|\mathbf{y}, x_2) \sim beta(x_2 + 35, 39),$$

$$p(x_2|\mathbf{y}, \theta) \sim bin(125, \theta/(2+\theta)).$$