Markov Chain and Monte Carlo

Summary of a paper by K B Athreya, M Delampady, T Krishnan
Adapted from *Resonance*, Volume 8, 2003

Nilava Metya

Chennai Mathematical Institute

September 29, 2021

Simple Monte Carlo

Riemann integration:

- Suppose we want to integrate $I := \int_D f$, and we are happy with an approximate answer.
- We divide D into smaller rectangles $\{D_i\}_{1\le i\le n}$ which are almost disjoint*.
- Evaluate $f(x_i)$ where $x_i \in D_i$.
- $I_n = \sum_{i=1}^{n} f(x_i)\,\mathrm{vol}\,(D_i)^{\dagger}$ approximates I.

*intersection has measure 0

†volume in the appropriate dimension

Method of statistical sampling (to find what proportion p of a population, size N, supports a given party):

- Select a sample of n individuals from the population of size N.
- Determine the proportion p_n in the sample that support the given party.
- Use this as an estimate of approximation for p.
- One can doubt the reliability of the estimate.

Theorem (LLN)

Let $X_1, X_2, \cdots$ be a sequence of iid RV's with $\mu = \mathbb{E}(X_1) < \infty$. Then the sample mean $\overline{X_n} = \frac{\sum_{i=1}^n X_i}{n}$ converges, in probability (weak law) or almost surely (strong law), to μ as $n \to \infty$.

So if h is a bounded function then $\frac{\sum_{i=1}^n h(X_i)}{n} \xrightarrow{a.s.} \mathbb{E}(h(X_1))$.

Let $X_1, X_2, \cdots$ be RV's which are independent uniformly distributed over a domain D (with density f) and h a bounded function. Then $\mathbb{E}(h(X_1)) = \int h(x)f(x)\,dx$. But $f(x) = \frac{1}{\text{vol}(D)}\ \forall x \in D$. By LLN,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{a.s.} \frac{1}{\text{vol}(D)}\int_D h(x)\,dx.$$

Therefore, an estimate of $I = \int_D h(x)\,dx$ is $I_n := \frac{\text{vol}(D)}{n}\sum_{i=1}^{n} h(X_i)$.
If we want to evaluate the integral wrt some mass distribution m then look at

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i)m(X_i) \xrightarrow{a.s.} \frac{1}{\text{vol}(D)}\int_D h(x)m(x)\,dx.$$

So our estimates are $J_n := \frac{\text{vol}(D)}{n}\sum_{i=1}^{n} h(X_i)m(X_i)$.

Example

Say we want to evaluate $J = \int_0^2 \sin x\, dx$. WolframAlpha's answer is 1.4161.

Here $D = [0, 2]$ so $\operatorname{vol} D = 2$. Generate n (large) many samples from $U(0, 2)$, say $X_1, \cdots, X_n$. Then the integral is close to $\frac{2}{n}\sum_{i=1}^{n} \sin X_i$.

Example

We again want to evaluate $J = \int_0^2 \sin x\, dx$. But this time, we integrate the function $\chi(x, y) = \begin{cases} 1 & \text{if } y \geq \sin x \\ 0 & \text{otherwise} \end{cases}$ over the domain $D = [0, 2] \times [0, 1]$. Now, we uniformly pick n points from D, say $X_1, X_2, \cdots$. Then the integral is close to $2 - \frac{2}{n}\sum_{i=1}^{n} \chi(X_i)$. Note that $\sum_{i=1}^{n} \chi(X_i) = \#\{\text{points above graph of } \sin x\} = n - \#\{\text{points below graph of } \sin x\}$.

We shall demonstrate an example of how to use the fact

$$\frac{\mathrm{vol}(D)}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{a.s.} \int_D f(x)h(x)\,dx$$

where $X_1, X_2, \cdots$ are all iid RV's with density f over D.

Example

We want to evaluate $J = \int_0^1 \sin x\,dx$. WolframAlpha's answer is 0.45970. Near $x = 0$, we have $\sin x \simeq x - \frac{x^3}{6}$. Hence we consider our density function to be $f(x) = \frac{24}{11}\left(x - \frac{x^3}{6}\right)$ and $h(x) = \frac{11}{24}$. Here $D = [0, 1]$ so $\mathrm{vol}\,D = 1$. Generate n (large) many samples from the above distribution, say $X_1, \cdots, X_n$. Then the integral is close to $\frac{1}{n}\sum_{i=1}^{n} h(X_i) = \frac{11}{24} \simeq 0.458$.

The Markov Chain Case

A Markov chain is a sequence of memoryless random variables, i.e., a sequence $X_0, X_1, \cdots$ of random variables such that $\mathbb{P}(A \cap B | X_n) = \mathbb{P}(A|X_n)\mathbb{P}(B|X_n)$ whenever A, B is an event defined in terms of the past $\{X_k : 0 \le k < n\}$ and the future $\{X_k : n < k\}$, respectively, for all $n \in \mathbb{N}_0$. We will be interested in Markov chains with a countable state space S. The above definition, in this case, is equivalent to demanding that $\mathbb{P}(X_{n+1} = a_{n+1} | X_n = a_n, \cdots, X_0 = a_0) = \mathbb{P}(X_{n+1} = a_{n+1} | X_n = a_n) \forall a_i \in S, n \ge 0$.

We define the transition probability matrix P where elements are $P_{ij} = \mathbb{P}(X_1 = a_j | X_0 = a_i)$. A stationary distribution π (which is a row vector) is a probability distribution such that $\pi_j = \sum_i \pi_i P_{ij} = \sum_i \pi_i \mathbb{P}(X_1 = j | X_0 = i)$, i.e., $\pi = \pi P$.

Finally we say that a Markov chain, with a countable state space, is *irreducible* if $\forall i, j \in S, \exists n = n_{ij} \ge 1$ such that $\mathbb{P}(X_n = j | X_0 = i) > 0$.

Theorem (LLN for Markov Chains on countable state space)

Let $\{X_i\}$ be an irreducible Markov chain with a countable state space S and a transition probability matrix P. Suppose $\pi = \begin{bmatrix} \pi_1 & \pi_2 & \cdots \end{bmatrix}$ is a stationary distribution of P. Then for any bounded function $h : S \to \mathbb{R}$ and for any initial distribution of X_0 we have that $\frac{1}{n} \sum_{i=0}^{n-1} h(X_i) \xrightarrow{P} \sum_i h(i)\pi_i$ *as* $n \to \infty$.

Using the LLN

Given a density π on (countable) S and a (bounded) function $h : S \to \mathbb{R}$, suppose we want $\sum_i h(i)\pi_i$.

- Find an irreducible Markov chain $\{X_i\}$ with state space S and stationary distribution π.
- Then starting a Markov chain from X_0 till X_{n-1}, we offer an estimate $\mu_n := \frac{1}{n}\sum_{i=0}^{n-1} h(X_i)$.

Example

Estimates for $\pi(A) = \sum_{i\in A} \pi_i$ for some $A \subseteq S$ are $\frac{1}{n}\sum_{i=0}^{n-1} \mathbf{1}_A(X_i)$ because $\pi(A) = \sum_i \mathbf{1}_A(i)\pi_i$.

If the Markov chain is aperiodic, then $\lim_{n\to\infty} \left|\mathbb{P}(X_n = j) - \pi_j\right| = 0$ holds additionally, for any X_0.
This means that instead of n (large) runs, we can do N independent runs of length m (s.t. $Nm = n$)

$$X_{0,1} \to X_{1,1} \to \cdots \to X_{m,1}$$

$$\vdots$$

$$X_{0,N} \to X_{1,2} \to \cdots \to X_{m,N}$$

and then we can offer the estimate $\mu_{N,m} = \frac{1}{N}\sum_{i=1}^{N} h(X_{m,i})$.

Metropolis-Hastings algorithm

The problem: To sample from a given (target) distribution.
The setup:

- A countable state space S.
- A (target) probability distribution π on S.
- A transition probability matrix $Q = (q_{ij})$, such that it is computationally easy to get a sample from the distribution $\{q_{ij}\}_{j \in S}$ for each $i \in S$.

We generate a Markov chain as follows (arbitrary $X_0 = i_0 \in S$):

1. If $X_n = i$, sample Y_n from the distribution $\{q_{ij}\}_{j \in S}$, i.e., $\mathbb{P}(Y_n = j | X_n = i) = q_{ij}$.
2. Define the acceptance probability $\rho(i,j) = \rho_{ij} := \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}$ whenever $\pi_i q_{ij} > 0$.
3. Choose X_{n+1} according to the distribution $\mathbb{P}(X_{n+1} = Y_n | X_n, Y_n) = 1 - \mathbb{P}(X_{n+1} = X_n | X_n, Y_n) = \rho(X_n, Y_n)$.

   X_n is then a Markov chain with transition matrix $P = (p_{ij})$ given by $p_{ij} = \begin{cases} q_{ij}\rho_{ij} & \text{if } i \neq j \\ 1 - \sum_{k \neq i} p_{ik} & \text{if } i = j \end{cases}$.

Claim

1. (Detailed balance) $\pi_i p_{ij} = \pi_j p_{ji} \forall i,j \in S$.
2. π is a stationary probability distribution for P.

Metropolis-Hastings: An attempted generalization

The problem: To sample from a given (target) distribution.
The setup:

- An uncountable state space S.
- A (target) probability distribution π on S.
- A transition probability function q.

We generate a Markov chain as follows (arbitrary $X_0 = \theta_0 \in S$):

1. If $X_n = \theta$, sample Y_n from the distribution $\{q(x|\theta)\}_{x \in S}$, i.e., $\mathbb{P}(Y_n = \theta' | X_n = \theta) = q(\theta'|\theta)$.
2. Define the acceptance probability $\rho\left(\theta, \theta'\right) := \min\left\{ \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}, 1 \right\}$ whenever $\pi(\theta)q(\theta'|\theta) > 0$.
3. Choose X_{n+1} according to the distribution $\mathbb{P}(X_{n+1} = X_n | X_n, Y_n) = 1 - \mathbb{P}(X_{n+1} = Y_n | X_n, Y_n) = \rho(X_n, Y_n)$.

   X_n is a Markov chain with transition kernel $P(X_{n+1} \in A | X_n = \theta) = \begin{cases} \int_A q(x|\theta)p(\theta, x)\,dx & \text{if } \theta \notin A \\ 1 - \int_{A^c} q(x|\theta)p(\theta, x)\,dx & \text{if } \theta \in A \end{cases}.$

Conjecture

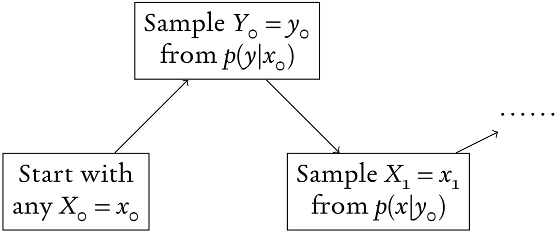
1. (Detailed balance) $\pi(B) \displaystyle\int_B P(X_{n+1} \in A | X_n = x)\,dx = \pi(A) \int_A P(X_{n+1} \in B | X_n = x)\,dx \forall A, B \subseteq S$.
2. π is a stationary probability distribution for P.

Gibb's Sampler

The problem: To sample from a multivariate distribution (MH was for single variable distribution).
Why and when: $p(x, y)$ is difficult to simulate, but not $p(x|y)$ or $p(y|x)$.
The setup (2 dimensional): Let π be the distribution of a bivariate random vector (X, Y).

We generate a Markov chain $Z_n = (X_n, Y_n)$:

If π is discrete, then the transition probability matrix Q of Z_n is given by $q_{(ij),(kl)} = \mathbb{P}(y_j \twoheadrightarrow x_k \twoheadrightarrow y_l) = p(x_k|y_j)p(y_l|x_k) = \frac{\pi_{kj}}{\sum_t \pi_{tj}} \cdot \frac{\pi_{kj}}{\sum_t \pi_{kt}}$.



Claim

1. Z_n is irreducible and aperiodic.
2. The stationary distribution of $\{Z_n\}$ is π.

Corollary

If h is bounded then $\sum_{ij} h(i,j)\pi_{ij}$ can be approximated by $\frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_j)$

Example (Bivariate normal)

Want $(X,Y) \sim N(\boldsymbol{\mu},\Sigma)$ where $\boldsymbol{\mu} = \mathbf{0}, \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ (ρ is the correlation coefficient).

Then $p(X|Y=y) = N(\rho y, 1-\rho^2), p(Y|X=x) = N(\rho x, 1-\rho^2)$. Here the second parameter denotes variance.
We do the following:

1. Start with any $X_0 = x_0$.
2. Sample $Y_0 = y_0 \sim N(\rho x_0, 1-\rho^2)$.
3. Sample $X_1 = x_1 \sim N(\rho y_0, 1-\rho^2)$.
4. Sample $Y_1 = y_1 \sim N(\rho x_1, 1-\rho^2)$.

$$\vdots$$

Statistical Concepts

Suppose a statistical test involves n identical and independent trials, with k possible outcomes, and probability of i^{th} outcome is p_i. Let N_i be the random variable denoting the number of times the i^{th} outcome happens. Then

$$f(\boldsymbol{n}|\boldsymbol{p}) := \mathbb{P}\left(N_1 = n_1, \cdots, N_k = n_k | p_1, \cdots, p_k\right) = \binom{\sum_i n_i}{n_1, \cdots, n_k} \prod_i p_i^{n_i}.$$

Given the data $\boldsymbol{n}$, the likelihood function of $\boldsymbol{p}$ is $l(\boldsymbol{p}) := f(\boldsymbol{n}|\boldsymbol{p})$. The $\boldsymbol{p}$ which explains the data "best" is the $\boldsymbol{p}$ for which $l(\boldsymbol{p})$ is maximal. Now, maximizing $l(\boldsymbol{p})$ is equivalent to maximizing $\log l(\boldsymbol{p}) = k + \sum_i n_i \log p_i$ for some constant k (because $\boldsymbol{n}$ is fixed). Recall that the constraint is $\sum_i p_i = 1$. This can be solved using Lagrange multipliers.
So the problem at hand is: Maximize $g(\boldsymbol{p}) = \sum_i n_i \log p_i$ constraint to $h = 1$ where $h(\boldsymbol{p}) = \sum_i p_i$.
$\nabla h(\boldsymbol{p}) = (1, \cdots, 1)$. $\nabla g(\boldsymbol{p}) = \left(\frac{n_1}{p_1}, \cdots, \frac{n_k}{p_k}\right)$. So $\frac{n_i}{p_i} = \lambda \forall i$ for some contant λ. Adding, $n = \lambda$. Hence, $\hat{p}_i = \frac{n_i}{n}$.

Sufficiency

We start with a probability density function $f(x|\theta)$ with unknown parameter θ. Further suppose we have a sample from a population with this distribution.

Definition (Statistic)

Any function of the sample is called a statistic.

Definition (Sufficient statistic)

A statistic $g(\boldsymbol{X})$ (where $\boldsymbol{X}$ is the sample) is said to be sufficient for θ if the conditional distribution of $\boldsymbol{X}$ given $g(\boldsymbol{X})$ does not involve θ.

EXAMPLE

Say $X_1, \cdots, X_n$ are iid Poisson RV's with mean λ. We can look at the parameter $\theta = \lambda$. A sufficient statistic is $g(\boldsymbol{X}) = \frac{1}{n}\sum X_i$. Note that $T := X_1 + \cdots + X_n$ is a Poisson RV with mean $n\theta$. Hence

$$
\begin{aligned}
\mathbb{P}\left(X_i = x_i \forall i | g(\boldsymbol{X}) = \mu\right) &= \frac{\mathbb{P}\left(X_i = x_i \forall i, T = n\mu\right)}{\mathbb{P}(T = n\mu)} \\
&= \frac{\mathbb{P}\left(X_i = x_i \forall 1 \le i < n, X_n = n\mu - \sum_{1 \le i < n} x_i\right)}{\mathbb{P}(T = n\mu)} \\
&= \left(\prod_{i=1}^{n-1} \frac{e^{-\theta}\theta^{x_i}}{x_i!}\right) \times \frac{e^{-\theta}\theta^{n\mu - \sum_{1 \le i < n} x_i}}{\left(n\mu - \sum_{1 \le i < n} x_i\right)!} \times \frac{(n\mu)!}{e^{-n\theta}(n\theta)^{n\mu}} \\
&= \frac{(n\mu)!}{n^{n\mu} \cdot \prod_i x_i!}
\end{aligned}
$$

Theorem

Let $\hat{\theta}(X_1, \cdots, X_n)$ be an estimator of θ with finite variance. Say T is a sufficient statistic for θ. Define $\hat{\theta}^*(t) := \mathbb{E}\left(\hat{\theta}(X_1, \cdots, X_n)|T = t\right)$. Then

$$\mathbb{E}\left[\left(\hat{\theta}^*(T) - \theta\right)^2\right] \le \mathbb{E}\left[\left(\hat{\theta}(X_1, \cdots, X_n) - \theta\right)^2\right]$$

with equality iff $\hat{\theta} = \hat{\theta}^*$.

The crux of the above theorem is that $\hat{\theta}^*$ is a better estimator than $\hat{\theta}$. In fact, it gives a constructive way to improve the estimator.
(For the special case $\mathbb{E}(\hat{\theta}) = \theta$) The key point to keep in mind, to realize why this theorem works, is the variance formula for two RV's S, T on the same probability space with $|\operatorname{Var}(S)| < \infty$:

$$\operatorname{Var}(S) = \operatorname{Var}(\mathbb{E}(S|T)) + \mathbb{E}(\operatorname{Var}(S|T)).$$

A more general version of this is Jensen's inequality.

A proof of $\mathrm{Var}(S) = \mathrm{Var}(\mathbb{E}(S|T)) + \mathbb{E}(\mathrm{Var}(S|T))$

We will repeatedly use that $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$.

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}(\mathbb{E}(S^2|T)) \\
&= \mathbb{E}\left(\mathrm{Var}(S|T) + [\mathbb{E}(S|T)]^2\right) \\
\implies \mathrm{Var}(S) = \mathbb{E}(S^2) - [\mathbb{E}(S)]^2 &= \mathbb{E}\left(\mathrm{Var}(S|T) + [\mathbb{E}(S|T)]^2\right) - [\mathbb{E}\left(\mathbb{E}(S|T)\right)]^2 \\
&= \mathbb{E}(\mathrm{Var}(S|T)) + \mathbb{E}\left(\boxed{\mathbb{E}(S|T)}^2\right) - \left[\mathbb{E}\left(\boxed{\mathbb{E}(S|T)}\right)\right]^2 \\
&= \mathbb{E}(\mathrm{Var}(S|T)) + \mathrm{Var}\left(\boxed{\mathbb{E}(S|T)}\right) \\
&= \mathrm{Var}(\mathbb{E}(S|T)) + \mathbb{E}(\mathrm{Var}(S|T))
\end{aligned}$$

Recall the example of Gibbs sampling from bivariate normal. We have a sample of size n, from the algorithm described in Gibbs sampling.

EXAMPLE

Let's try to estimate $\mu_x = \mathbb{E}(X)$. One obvious guess is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Note that $y_1, \cdots, y_n$ also carry some information about x_i which has not been used: $\mu_x = \mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$. Hence,

$$\hat{\mu}_x := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X|y_i) = \frac{\rho}{n}\sum_{i=1}^{n} y_i$$

is also an estimator. In fact, it is better because $\mathrm{Var}(X) - \mathrm{Var}(\mathbb{E}(X|Y)) = \mathbb{E}(\mathrm{Var}(X|Y)) \geq 0$.

Statistical Applications

Metropolis Hastings Algorithm

Consider a continuous and strictly monotone distribution F. Then $F(x) = P(X \leq x)$ for a random variable X iff $F(X) \sim U(0,1)$. So we can sample from a given distribution F if we sample $U \sim U(0,1)$. Just take $X = F^{-1}(U)$. F^{-1} makes sense because F is continuous and strictly monotone.

Example

Gamma $\Gamma(\alpha,\beta)$ with parameters α,β has density $f_{\alpha,\beta}(x) = \dfrac{\beta^\alpha e^{-\alpha x} x^{\alpha-1}}{\Gamma(\alpha)}$, where $\Gamma(\alpha) = \displaystyle\int_0^\infty e^{-x} x^{\alpha-1}\, dx$.

It is known that if $\alpha = k \in \mathbb{Z}$ then $X_{k,\beta} = \frac{-1}{\beta} \sum_{i=1}^{k} \ln U_i$ has $\Gamma(\alpha = k, \beta)$ distribution, where U_i are iid $U(0,1)$. In case $\alpha \notin \mathbb{Z}$, we use *rejection sampling*.

Say an RV X has density f. We use the help of another *proposal distribution* g such that $cg(x) \geq f(x) \forall x$, for some constant $c > 0$ and it is easy to sample from g.

The idea is as follows: Say T is sampled from h. The points $(T, cg(T) \cdot U)$ are uniformly distributed over $\{t\} \times [0, cg(t)]$ for each t from the sample T. Next, we accept only those $(t, cug(t))$ which lie below G_f, i.e., $u \leq \frac{f(t)}{cg(t)}$. These points are uniformly distributed over $\{t\} \times [0, f(t)]$. Clearly the probability of drawing an accepted value over $(t, t + dt)$ is $\propto h(t)\,dt \cdot \frac{f(t)}{cg(t)}$. Hence it is wise to choose $h = g$ so that the infinitesimal probability of an accepted value is $\propto \frac{1}{c} f(t)\,dt$.

Hence we have the following algorithm:

1. Do the following until one T is accepted:
    1. Sample $T \sim g$.
    2. Sample $U \sim U(0, 1)$.
    3. Accept T if $U \leq \frac{f(T)}{cg(T)}$, else reject.
2. Repeat the previous step unless the required number of samples are accepted.

Example ($\Gamma(\alpha, 1)$)

Let's complete the example of drawing $X \sim \Gamma(\alpha, \beta)$ when $\alpha \notin \mathbb{Z}$ and restricting to $\beta = 1$. $f = f_{\alpha,\beta}$. The proposed distribution is $g \leftarrow \Gamma\left(\lfloor\alpha\rfloor, \frac{\lfloor\alpha\rfloor}{\alpha}\right)$, with $c = \dfrac{\alpha^{\alpha} e^{-\alpha} / \Gamma(\alpha)}{\lfloor\alpha\rfloor^{\lfloor\alpha\rfloor} e^{-\lfloor\alpha\rfloor} / \Gamma(\lfloor\alpha\rfloor)}$.

So our algorithm becomes (repeat until desired):

1. Sample $T \sim \Gamma\left(\lfloor\alpha\rfloor, \frac{\lfloor\alpha\rfloor}{\alpha}\right)$. We know how to do this because $\lfloor\alpha\rfloor \in \mathbb{Z}$. Or use MH. [a]
2. Sample $U \sim U(0, 1)$.
3. Accept T if $U \le \dfrac{f(T)}{cg(T)} = \left(\dfrac{eT}{e^{T}\alpha}\right)^{\alpha - \lfloor\alpha\rfloor}$, else reject.

[a] An independent MH procedure, with $q(y) = g(y)$

Statistical Applications

Gibbs Sampling

In case of inference for the θ in binomial distribution, we take the "prior" to be $g(\theta) \leftarrow \beta(\alpha,\gamma)$ having density $\propto \theta^{\alpha-1}(1-\theta)^{\gamma-1}$. So the density of θ given data x is just, by Bayes' theorem, $\pi(\theta|x) \propto f(x|\theta)g(\theta) \propto \theta^{x}(1-\theta)^{n-x} \cdot \theta^{\alpha-1}(1-\theta)^{\gamma-1}$.

In the multinomial case, as a generalization, we'd like our prior to be $\propto \prod_{i=1}^{n} p_i^{\alpha_i-1}$, where α_i are the parameters. This is known as the *Dirichlet distribution with parameters* $\alpha_1, \cdots, \alpha_n$, and along with the correct proportionality constant, the distribution becomes $\boxed{\pi(\boldsymbol{p}) = \frac{\Gamma\left(\sum_{i=1}^{n}\alpha_i\right)}{\prod_{i=1}^{n}\Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{\alpha_i-1}}$.

Bayes' theorem gives the posterior $\pi(\boldsymbol{p}|\boldsymbol{n}) \propto f(\boldsymbol{n}|\boldsymbol{p})\pi(\boldsymbol{p}) \propto \prod_{i=1}^{n} p_i^{n_i} \cdot \prod_{i=1}^{n} p_i^{\alpha_i-1} = \prod_{i=1}^{n} p_i^{n_i+\alpha_i-1}$.

Here the correct constant of proportionality is $\frac{\prod_{i=1}^{n}\Gamma(n_i+\alpha_i)}{\Gamma\left(n+\sum_{i=1}^{n}\alpha_i\right)}$.

Example: application of Gibbs Sampler

The problem: We want to estimate probabilities p, q, r from data n_o, n_a, n_b, n_{ab} when the following is known:

x	$f(x\|p,q,r)$
$n_o = n_{oo}$	r^2
n_{aa}	p^2
n_{ao}	$2pr$
n_{bb}	q^2
n_{bo}	$2qr$
n_{ab}	$2pq$

Here $n_a = n_{aa} + n_{ao}, n_b = n_{bb} + n_{bo}$.
An attempt: The Bayesian estimation of p, q, r with a Dirichlet prior with parameters α, β, γ involves the likelihood function $L(p,q,r) = r^{2n_o}(p^2 + 2pr)^{n_a}(q^2 + 2qr)^{n_b}(2pq)^{n_{ab}}$.
Also, the posterior $\propto r^{2n_o+\gamma-1}(p^2 + 2qr)^{n_a}(q^2 + 2pr)^{n_b} p^{n_{ab}+\alpha-1} q^{n_{ab}+\beta-1}$.
These are too complicated to do computations with. Gibbs Sampler method comes to rescue, here.

Denote $\boldsymbol{n} = (n_{oo}, n_{aa}, n_{ao}, n_{bb}, n_{bo}, n_{ab}), \tilde{\boldsymbol{n}} = (n_o, n_a, n_b, n_{ab}), \boldsymbol{N} = (n_{aa}, n_{bb}), \boldsymbol{P} = (p, q, r)$.
With the data $\boldsymbol{n}$, the likelihood $\propto r^{2n_{oo}+n_{ao}+n_{bo}} \cdot q^{2n_{bb}+n_{bo}+n_{ab}} \cdot p^{2n_{aa}+n_{ab}+n_{ao}}$.
We let $N_a := 2n_{aa} + n_{ab} + n_{ao}, N_b := 2n_{bb} + n_{bo} + n_{ab}, N_o := 2n_{oo} + n_{ao} + n_{bo}$ so that the above is $p^{N_a} q^{N_b} r^{N_o}$.
So the posterior $\propto p^{N_a+\alpha-1} q^{N_b+\beta-1} r^{N_o+\gamma-1}$ when the prior is Dirichlet with parameters α, β, γ.
Clearly, for given $\tilde{\boldsymbol{n}}, \boldsymbol{P}$ the probability of $\{n_{aa} = k\}$ is simply $\binom{n_a}{k} \left(\frac{p^2}{p^2+2pr}\right)^k \left(\frac{2pr}{p^2+2pr}\right)^{n_a-k}$. So,
$(n_{aa} | \tilde{\boldsymbol{n}}, \boldsymbol{P}) \sim \mathrm{Bin}\left(n_a, \frac{p^2}{p^2+2pr}\right)$. Similarly, $(n_{bb} | \tilde{\boldsymbol{n}}, \boldsymbol{P}) \sim \mathrm{Bin}\left(n_b, \frac{q^2}{q^2+2qr}\right)$. These are independent.
Finally, $(p, q, r | \boldsymbol{n}) \sim \mathrm{Dirichlet}(N_a + \alpha, N_b + \beta, N_o + \gamma)$.
In the description Gibb's sampler given earlier, we shall sample $(\boldsymbol{N} | \tilde{\boldsymbol{n}}, \boldsymbol{P})$ and $(\boldsymbol{P} | \tilde{\boldsymbol{n}}, \boldsymbol{N})$ in turns.
Suppose we have run the Gibbs sampler algorithm k independent times (each time, long enough to be close to the "limit"). We have $\boldsymbol{P}_1, \cdots, \boldsymbol{P}_k, \boldsymbol{N}_1, \cdots, \boldsymbol{N}_k$.
The posterior mean of $\boldsymbol{P}$ can be estimated by Rao-Blackwellization:

$$\frac{1}{k} \sum_{i=1}^{k} \mathbb{E}(\boldsymbol{P}_i | \tilde{\boldsymbol{n}}, \boldsymbol{N}_i) = \frac{(\alpha + n_{ab} + n_{ao}, \beta + n_{bo}, n_{ab}, \gamma + n_{ao} + n_{bo}) + 2\sum_{i=1}^{k} \boldsymbol{N}_i}{k(\alpha + \beta + \gamma + 2n)}$$

THE END