Bayesian statistics. Bayesian theory of statistics is based on the principle that a "degree of belief" regarding a parameter θ of model is expressed by a probability distribution for θ over a parameter space Ω. This differs from frequentists who interpret θ as the parameter uniquely identifying a population of interest. In the population model a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ follows a density function $f(\boldsymbol{x}; \theta)$ with parameter $\theta \in \Omega$. A Bayesian model begins with a **prior distribution** $\pi(\theta)$ over the parameter space Ω. Furthermore, $f(\boldsymbol{x}; \theta)$ is viewed as the conditional distribution of $\boldsymbol{X}$ given θ. By the Bayes' rule the conditional density $\pi(\theta|\boldsymbol{x})$ can be derived from

$$\pi(\theta|\boldsymbol{x}) = \begin{cases} \pi(\theta)f(\boldsymbol{x};\theta) \Big/ \displaystyle\sum_{\theta\in\Omega} \pi(\theta)f(\boldsymbol{x};\theta) & \text{if } \Omega \text{ is discrete;} \\ \pi(\theta)f(\boldsymbol{x};\theta) \Big/ \displaystyle\int_{\Omega} \pi(\theta)f(\boldsymbol{x};\theta)\,d\theta & \text{if } \Omega \text{ is continuous.} \end{cases}$$

and called the **posterior distribution**.

Posterior distribution. Regardless of whether the parameter space Ω is discrete or continuous, the posterior distribution $\pi(\theta \mid \boldsymbol{x})$ is "proportional" to $\pi(\theta)f(\boldsymbol{x};\theta)$ up to the constant. Thus, we write

$$\pi(\theta \mid \boldsymbol{x}) \propto \pi(\theta)f(\boldsymbol{x};\theta)$$

The exact posterior distribution $\pi(\theta \mid \boldsymbol{x}) = c(\boldsymbol{x})\pi(\theta)f(\boldsymbol{x};\theta)$ can be obtained by calculating the **normalizing constant**

$$c(\boldsymbol{x}) = \begin{cases} 1 \Big/ \displaystyle\sum_{\theta\in\Omega} \pi(\theta)f(\boldsymbol{x};\theta) & \text{if } \Omega \text{ is discrete;} \\ 1 \Big/ \displaystyle\int_{\Omega} \pi(\theta)f(\boldsymbol{x};\theta)\,d\theta & \text{if } \Omega \text{ is continuous.} \end{cases}$$

The calculation of $c(\boldsymbol{x})$ may be intractable but unnecessary in light of MCMC algorithms for posterior distribution.

Conjugate family of distributions. The normalizing constant $c(\boldsymbol{x})$ can be tractable in a special occasion when both the prior density function $\pi(\theta)$ and the posterior density

$$\pi(\theta \mid \boldsymbol{x}) \propto \pi(\theta)f(\boldsymbol{x};\theta)$$

belong to the same family of density function $\pi(\theta;\eta)$ with **hyperparameter** η. Then $\pi(\theta;\eta)$ is called **conjugate** to $f(\boldsymbol{x};\theta)$. For example, a prior density

$$\pi(\theta;\eta) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}, \quad 0 \le \theta \le 1$$

of beta distribution with hyperparameter $\eta = (\alpha_1, \alpha_2)$ is conjugate to a binomial distribution

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \dots, n,$$

since the posterior density has a beta distribution with parameter $(\alpha_1 + x, \alpha_2 + n - x)$.

Bernoulli trials and their conjugate. Consider an observation $\mathbf{z} = (z_1, \ldots, z_n)$ of independent Bernoulli trials z_i's according to the probability mass

$$f(j;\theta) = \begin{cases} \theta & \text{if } j = 1; \\ 1-\theta & \text{if } j = 2, \end{cases}$$

for $i = 0, 1, \ldots, n$, and a prior density $\pi(\theta)$ of beta distribution with parameter (α_1, α_2). Particularly the prior distribution

$$\pi(\theta) \propto 1$$

is called **uninformative** (or flat) prior when $\alpha_1 = \alpha_2 = 1$. Let $m_j(\mathbf{z})$ be the number of z_i's such that $z_i = j$ for $j = 1$ or 2. In the Bernoulli trial model we immediately observe that

$$\pi(\theta \mid \mathbf{z}) \propto \pi(\theta) f(\mathbf{z};\theta) \propto \theta^{\alpha_1 + m_1(z) - 1}(1-\theta)^{\alpha_2 + m_2(z) - 1}$$

which is proportional to a beta distribution with parameter $(\alpha_1 + m_1(z), \alpha_2 + m_2(z))$.

Bayesian model of simple mixture. We assume that the pdf f_1 and f_2 of two components are entirely known, and that a data set $\mathbf{x} = (x_1, \ldots, x_n)$ is an independent sample from the simple mixture

$$f(x_i;\theta) = \theta f_1(x_i) + (1-\theta) f_2(x_i), \quad i = 1, \ldots, n.$$

Then the posterior density $\pi(\theta|\mathbf{x})$ of weight parameter θ with a prior $\pi(\theta) \propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$ is proportional to

$$\pi(\theta|\mathbf{x}) \propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1} \prod_{i=1}^{n} [\theta f_1(x_i) + (1-\theta) f_2(x_i)].$$

The numerical analysis of the posterior density $\pi(\theta|\mathbf{x})$ is very hard. Alternatively, we can devise an MCMC method to perform a Monte Carlo integration.

Data augmentation algorithm. Let $\mathbf{z} = (z_1, \ldots, z_n)$ be a vector of latent variables z_i's of Bernoulli trial identifying the first or the second component the i-th observation x_i belongs to. Then the posterior density $\pi(\boldsymbol{\theta}|\mathbf{x})$ of interest is viewed as the marginal density of the joint density

$$\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x}) \propto \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \prod_{i=1}^{n} \theta_{z_i} f_{z_i}(x_i)$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\theta, 1-\theta)$. The corresponding Gibbs sampler is called **data augmentation algorithm**, and it is formed by two successive probability transitions—then the first one from $\boldsymbol{\theta} = (\theta_1, \theta_2)$ to $\mathbf{z} = (z_1, \ldots, z_n)$ and the second one from $\mathbf{z}$ to $\boldsymbol{\theta}$. Given a parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)$ of interest we obtain the first probability transition by

$$\pi(\mathbf{z}|\boldsymbol{\theta}, \mathbf{x}) = \prod_{i=1}^{n} \left[\frac{\theta_{z_i} f_{z_i}(x_i)}{\theta_1 f_1(x_i) + \theta_2 f_2(x_i)} \right]$$

Given a latent vector $\mathbf{z} = (z_1, \ldots, z_n)$ the first probability transition $\pi(\boldsymbol{\theta}|\mathbf{z}, \mathbf{x})$ is proportional to

$$\pi(\theta|\mathbf{z}, \mathbf{x}) \propto \theta^{m_1(z) + \alpha_1 - 1}(1-\theta)^{m_2(z) + \alpha_2 - 1}$$

and it is a beta distribution with $(m_1(\mathbf{z}) + \alpha_1, m_2(\mathbf{z}) + \alpha_2)$ where $m_j(\mathbf{z})$ denotes the number of z_i's such that $z_i = j$ for $j = 1$ or 2. Then the Gibbs sampler updates a Markov chain $(\boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)})$ as follows:

1. Generate each latent variable $z_i^{(t)}$ by setting $z_i^{(t)} = j$ independently according to

$$\pi_i(j|\boldsymbol{\theta}^{(t)}, \mathbf{x}) = \frac{\theta_j^{(t)} f_j(x_i)}{\theta_1^{(t)} f_1(x_i) + \theta_2^{(t)} f_2(x_i)}, \quad j = 1, 2$$

2. Generate the updated state $\boldsymbol{\theta}^{(t+1)} = (\theta, 1-\theta)$ by setting θ according to a beta distribution with $(m_1(\mathbf{z}^{(t)}) + \alpha_1, m_2(\mathbf{z}^{(t)}) + \alpha_2)$.

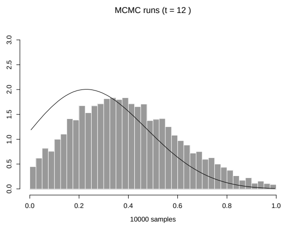
Markov chain Monte Carlo. Data augmentation algorithm generates a Markov chain

$$(\boldsymbol{\theta}^{(0)}, \mathbf{z}^{(0)}), (\boldsymbol{\theta}^{(1)}, \mathbf{z}^{(1)}), (\boldsymbol{\theta}^{(2)}, \mathbf{z}^{(2)}), \ldots$$

which is ergodic. Since it is a Gibbs sampler, it has the stationary distribution $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$, and $\boldsymbol{\theta}^{(t)} = (\theta^{(t)}, 1 - \theta^{(t)})$ satisfies

$$\theta^{(t)} \xrightarrow{\mathcal{L}} \pi(\theta|\mathbf{x}) \quad \text{as} \quad t \longrightarrow \infty.$$

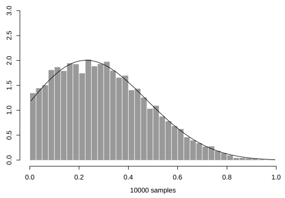
Therefore, when the running time t is "large enough" we can sample $\theta^{(t)}$ "*approximately*" from $\pi(\theta|\mathbf{x})$. But... how long is long enough?



Time to stationarity. Convergence theorem. An ergodic Markov chain $\mathbf{X}$ has a unique stationary distribution π such that

$$\mathbf{X}_t \xrightarrow{\mathcal{L}} \pi \quad \text{as} \quad t \longrightarrow \infty.$$

Strong stationary time (Aldous and Diaconis, 1987). For any ergodic Markov chain $\mathbf{X}$, there exists a "self-verifying" time T so that $\mathbf{X}_T \sim \pi$.



Question. Does there exist a practical algorithm to construct such a "self-verifying" time?

Coupling from the past (CFTP). Let $\mathbf{P}$ be a transition matrix on a discrete state space S. Then a function ϕ from $S \times V$ to S is called a **transition rule** if there exists a V-valued r.v. $\mathbf{U}$ such that

$$\mathbf{P}(x, y) = P(\phi(x, \mathbf{U}) = y) \quad \text{for } x, y \in S.$$

Backward coupling. With independent copies $\mathbf{U}_{-1}, \mathbf{U}_{-2}, \dots$ of $\mathbf{U}$, we can construct a trajectory $(\mathbf{X}_{-t}^{(z)}, \dots, \mathbf{X}_{-1}^{(z)}, \mathbf{X}_0^{(z)})$ starting from $\mathbf{X}_{-t}^{(z)} = z$ at time $(-t)$ in the past and moving forward by

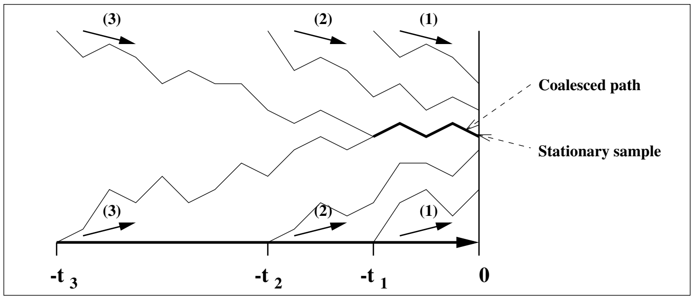
$$\mathbf{X}_{s+1}^{(z)} = \phi(\mathbf{X}_s^{(z)}, \mathbf{U}_s) \quad \text{for } s = -t, \dots, -1.$$

until the time $s = 0$.

CFTP algorithm (Propp and Wilson [4]). Starting from $t = (-t_k)$ in the past, we can run a coupled chain $(\mathbf{X}_{-t}^{(z)}, \dots, \mathbf{X}_{-1}^{(z)}, \mathbf{X}_0^{(z)})$ from every state $z \in S$. If

$$\mathbf{X}_0^{(z)} = \mathbf{X}_0^{(w)} \quad \text{for any } z, w \in S,$$

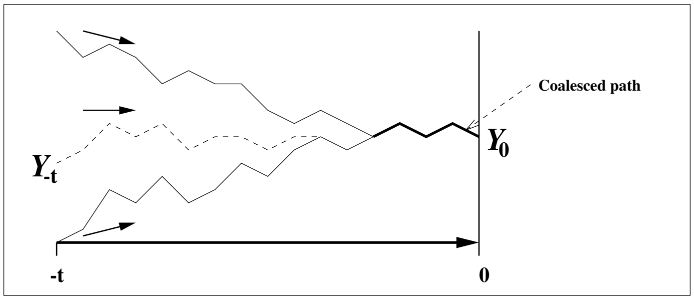
then we return the value $\mathbf{X}_0^{(z)}$. Otherwise, restart with a further step $(-t_{k+1})$ in the past.



How does CFTP work? Assuming $\mathbf{Z} \sim \pi$, construct a stationary trajectory $(\mathbf{Y}_{-t}, \dots, \mathbf{Y}_{-1}, \mathbf{Y}_0)$ with $\mathbf{Y}_{-t} = \mathbf{Z}$ via

$$\mathbf{Y}_{s+1} = \phi(\mathbf{Y}_s, \mathbf{U}_s) \quad \text{for } s = -t, \dots, -1.$$

If the backward coupling have coalesced, then $\mathbf{Y}_0 = \mathbf{X}_0^{(z)}$ for an arbitrary $z \in S$. Thus, we can observe a stationary path $\mathbf{Y}_0$ at $t = 0$ distributed as π.



Moving backward: Time-reversed chain. Let $\mathbf{P}$ be an irreducible and aperiodic transition probability with stationary distribution π. Define the **time-reversal** $\tilde{\mathbf{P}}$ of $\mathbf{P}$ by

$$\tilde{\mathbf{P}}(x, y) := \pi(y)\mathbf{P}(y, x)/\pi(x). \tag{4.1}$$

If $\tilde{\mathbf{P}} = \mathbf{P}$ then $\mathbf{P}$ is said to be **time-reversible**. Let $\mathbf{X}$ and $\tilde{\mathbf{X}}$ be chains from $\mathbf{P}$ and $\tilde{\mathbf{P}}$, respectively. Then

1. $\tilde{\mathbf{X}}_t \xrightarrow{\mathcal{L}} \pi$ as $t \longrightarrow \infty$;

2. $\mathcal{L}(\tilde{\mathbf{X}}_t, \dots, \tilde{\mathbf{X}}_0 \mid \tilde{\mathbf{X}}_0 = x, \tilde{\mathbf{X}}_t = y)$
$$= \mathcal{L}(\mathbf{X}_0, \dots, \mathbf{X}_t \mid \mathbf{X}_0 = y, \mathbf{X}_t = x),$$

where $\mathcal{L}(\cdots \mid \mathbf{X}_0 = y, \mathbf{X}_t = x)$ represents the law of distribution conditionally given $\mathbf{X}_0 = y$ and $\mathbf{X}_t = x$. In other words, a sample path $\mathbf{X}_0, \dots, \mathbf{X}_t$ is viewed as a trajectory of $\tilde{\mathbf{X}}_t, \dots, \tilde{\mathbf{X}}_0$ moving backward when they have shared the pair (x, y) of terminating states.

Imputation of coupled trajectory.

Coupling forward. With independent copies $\mathbf{U}_1, \mathbf{U}_2, \dots$ of a random variable $\mathbf{U}$, we can construct a trajectory $(\mathbf{X}_0^{(z)}, \dots, \mathbf{X}_t^{(z)})$ starting from $\mathbf{X}_0^{(z)} = z$ at time 0, and moving forward by

$$\mathbf{X}_s^{(z)} = \phi(\mathbf{X}_{s-1}^{(z)}, \mathbf{U}_s) \quad \text{for } s = 1, \dots, t.$$

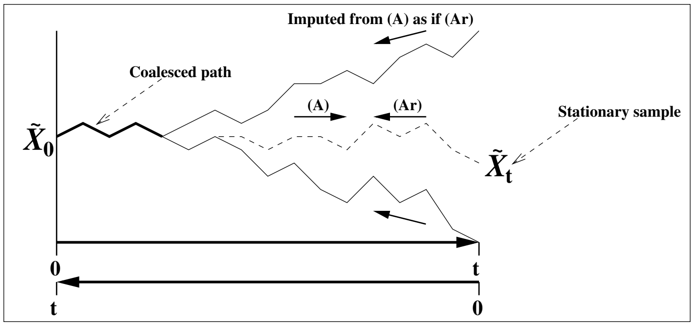
Imputation. Given a trajectory $(\mathbf{Y}_0, \dots, \mathbf{Y}_t) = (y_0, \dots, y_t)$, we can impute a forward coupling via

$$\mathbf{X}_s^{(z)} = \phi(\mathbf{X}_{s-1}^{(z)}, \mathbf{U}'_s) \quad \text{for } s = 1, \dots, t.$$

where $\mathbf{U}'_s$ is imputed from $\mathcal{L}(\mathbf{U}_s \mid \phi(y_{s-1}, \mathbf{U}_s) = y_s)$. So that it can be viewed as if the two trajectories of $\mathbf{X}_s^{(z)} = \phi(\mathbf{X}_{s-1}^{(z)}, \mathbf{U}_s)$ and $\mathbf{Y}_s = \phi(\mathbf{Y}_{s-1}, \mathbf{U}_s)$, $s = 1, \dots, t$, were coupled forward by common random variables $\mathbf{U}_s$'s.

Fill's algorithm [1].

1. **(A)** Run the time-reversed chain $\tilde{\mathbf{X}}$ for t steps;
2. **(Ar)** View $(\tilde{\mathbf{X}}_t, \dots, \tilde{\mathbf{X}}_0)$ as a trajectory $(\mathbf{Y}_0, \dots, \mathbf{Y}_t)$ moving backward.
3. Given $(\mathbf{Y}_0, \dots, \mathbf{Y}_t)$, impute $(\mathbf{X}_0^{(z)}, \dots, \mathbf{X}_t^{(z)})$ which runs backward from every state $z \in S$ at $\mathbf{X}_0^{(z)}$.
4. If the entire trajectories of a forward coupling have coalesced then we return the value $\tilde{\mathbf{X}}_t$.



How does Fill's algorithm work? Consider the following "hypothetical" construction: Assuming $\mathbf{Z} \sim \pi$, construct a stationary trajectory $(\mathbf{Y}_0, \dots, \mathbf{Y}_t)$ with $\mathbf{Y}_0 = \mathbf{Z}$ via

$$\mathbf{Y}_s = \phi(\mathbf{Y}_{s-1}, \mathbf{U}_s) \quad \text{for } s = 1, \dots, t.$$

Define the **coalescence event** by

$$C := \left\{ \mathbf{X}_t^{(z)} = \mathbf{X}_t^{(w)} \text{ for any } z, w \in S \right\}.$$

Since $\{\mathbf{Y}_0 = y\}$ and $\{\mathbf{Y}_t = x\} \cap C$ are independent, we have

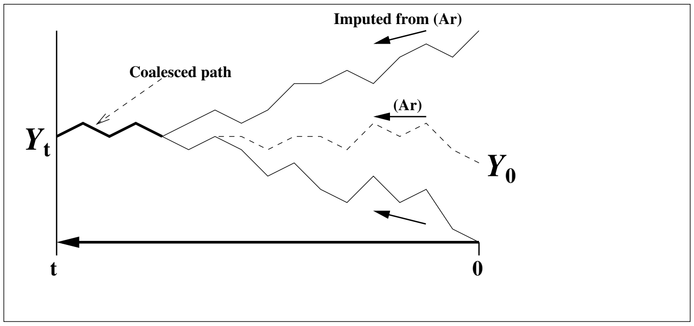
$$P(\mathbf{Y}_0 = y \mid \mathbf{Y}_t = x, C) = P(\mathbf{Y}_0 = y) = \pi(y).$$

Observe that the time-reversibility of (4.1) implies $\mathcal{L}(\tilde{\mathbf{X}}_1 \mid \tilde{\mathbf{X}}_0 = x) = \mathcal{L}(\mathbf{Y}_0 \mid \mathbf{Y}_1 = x)$ when $\mathbf{Y}_0$ is distributed as π. In general we can establish

$$\mathcal{L}(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_t \mid \tilde{\mathbf{X}}_0 = x) = \mathcal{L}(\mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_0 \mid \mathbf{Y}_t = x),$$

which allows us to construct a trajectory of $\mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_0$ running backward conditionally given $\mathbf{Y}_t = x$. Consequently we obtain

$$P(\tilde{\mathbf{X}}_t = y \mid \tilde{\mathbf{X}}_0 = x, C) = P(\mathbf{Y}_0 = y \mid \mathbf{Y}_t = x, C) = \pi(y).$$



Fill's algorithm: Comparison with CFTP.

- Fill's algorithm is a rejection algorithm, and independent of stopping time T (CFTP is not); thus, it allows us to construct a strong stationary time.
- Because of independence between a returned sample and a stopping time, complete transition rules are not required in storage (CFTP does require) by terminating it earlier.
- If the time-reversed $\tilde{\mathbf{P}}$ is stochastically monotone then the forward coupling can be reduced to a bivariate coupling. It can consider a cross-monotone case [2]. Whereas, realizable monotonicity is required for CFTP.
- Fill's algorithm requires a (potentially intricate) procedure for imputation, which may not be implemented in practice.

Monotonicity by data augmentation algorithm. Assuming uninformative prior $\pi(\theta) \equiv 1$ in the simple mixture model the MCMC can be implemented as follows. Given a current state $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$:

1. With iid uniforms $U_1, \ldots, U_n$ on $[0, 1)$, set

$$z_i^{(t)} = \begin{cases} 1 & \text{if } U_i < \pi_i(1|\boldsymbol{\theta}^{(t)}, \mathbf{x}); \\ 2 & \text{otherwise.} \end{cases}$$

2. With iid exponentials $V_1, \dots, V_{n+2}$, set

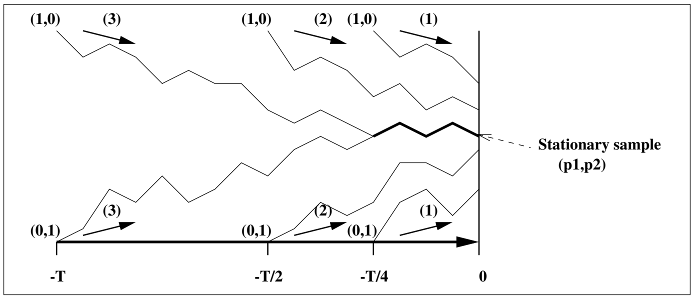
$$\theta_1^{(t+1)} = \left(V_1 + \cdots + V_{m_1(\mathbf{z}^{(t)})+1}\right) / \left(V_1 + \cdots + V_{n+2}\right) ;$$
$$\theta_2^{(t+1)} = 1 - \theta_1^{(t+1)}.$$

CFTP for simple mixture model (Hobert et al., [3]).

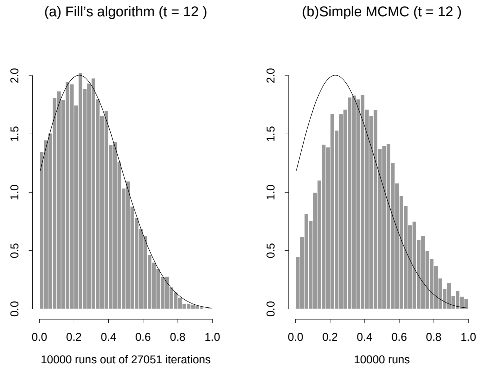
Realizable monotonicity. For $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, define $\boldsymbol{\theta} \le \boldsymbol{\eta}$ if $\theta_1 \le \eta_1$. The MCMC achieves

$$\boldsymbol{\theta}^{(t)} \le \boldsymbol{\eta}^{(t)} \Longrightarrow \boldsymbol{\theta}^{(t+1)} \le \boldsymbol{\eta}^{(t+1)}$$

Thus, CFTP algorithm applies.

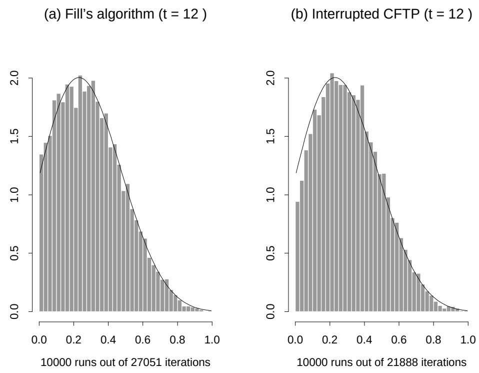


An example of perfect sampling. We took the two components $\mathcal{N}(0, 0.9)$ and $\mathcal{N}(0.8, 0.9)$ weighed by $\theta_1 = 0.18$ and $\theta_2 = 0.82$, respectively. A choice of initial state and an earlier termination ($t = 12$ steps) does not affect the Fill's algorithm.


(a) Fill's algorithm (t = 12)

(b)Simple MCMC (t = 12)

Independence of early termination. CFTP cannot allow us an early termination, and has to continue until the coalescence, while Fill's algorithm can. The comparison below illustrates this subtle distinction of two algorithms.



Comparison of Monte Carlo integration. Here we compare Fill's algorithm, CFTP, and MCMC for the quality of Monte Carlo integration

$$\bar{\theta} = \int_0^1 \theta \pi(\theta|\mathbf{x}) d\theta \approx \frac{1}{M} \sum_{m=1}^{M} \theta_m^{(t)}$$

when we stopped the algorithms earlier at $t = 12$.

Method	$\bar{\theta}$
Numerical integration	0.3058
Fill's algorithm†	0.3031
Interrupted CFTP†	0.3162
Simple MCMC†	0.3969

† For $M = 10000$, samples from $\theta_m^{(12)}$ were obtained.

Activities and report questions.

1. Explain data augmentation algorithm for Bayesian simple mixture model.
2. Discuss how a Bayesian approach is related with Monte Carlo methodology. What did you learn about it from this lecture series?
3. Explain coupling and coalescence in the context of perfect sampling algorithms (CFTP and Fill).
4. What are the limitations of perfect sampling algorithms in MCMC?

References

[1] James Allen Fill. An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Probab.*, 8(1):131–162, 1998.

[2] James Allen Fill, Motoya Machida, Duncan J. Murdoch, and Jeffrey S. Rosenthal. Extension of Fill's perfect rejection sampling algorithm to general chains. In *Proceedings of the Ninth International Conference "Random Structures and Algorithms" (Poznan, 1999)*, volume 17, pages 290–316, 2000.

[3] James P Hobert, Christian P Robert, and DM Titterington. On perfect simulation for some mixtures of distributions. *Statistics and Computing*, 9(4):287–298, 1999.

[4] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 223–252, 1996.