Mixture of kernels moralizing perfect simulation and perfect sampling for mixtures

Christian P. Robert

Université Paris Dauphine &

CREST, Insee, Paris

with George Casella, University of Florida

James P. Hobert, University of Florida

Kerrie L. Mengersen, Queensland University of Technology

D.M. Titterington, University of Glasgow

0. Perfect sampling in a nutshell

- Coupling from the past (CFTP)

[Propp & Wilson, 1996]

---

1. Start from every possible value at time $-t$
2. Run the m chains till time 0 *(coupling allowed)*
3. Check if the chains are equal at time 0
4. If not, start further back: $t \leftarrow 2 * t$

- Extension to continuous chains (multigamma coupling)

[Murdoch & Green, 1998]

1. The beginning of the story

Consider two Markov kernels K_1 and K_2

What of the mixture

$$K_3 = pK_1 + (1-p)K_2\,?$$

Example If K_1 kernel of a random walk with drift $1/p$,
K_2 kernel of a random walk with drift $-1/(1-p)$,
both kernels transient but K_3 recurrent since average drift 0.

2. Stability of the mixture

Theorem 1 *If* K_1 *and* K_2 *are recurrent kernels, the mixture kernel* K_3 *is recurrent.*

Theorem 2 *If* K_1 *and* K_2 *define positive recurrent chains with the same potential function* V*, that is, there exist a small set* C*,* $\lambda < 1$*,* $V \geq 1$ *and* V *bounded on* C *such that*

$$\mathbb{E}_{K_i}[V(x)|y] = \lambda V(y) + b\mathbb{I}_C(y)$$

then the mixture kernel K_3 *is also positive recurrent.*

3. Stationary measure

If $\pi_1 = \pi_2$ and K_3 is positive recurrent, π_1 is its stationary distribution.

Otherwise...

Special case: K_1 is an iid kernel π_1. Then

$$K_3 = p\pi_1 + (1-p)K_2$$

No assumption on K_2 (it can even be transient!) but, still,

Theorem 3 *K_3 is positive recurrent with stationary distribution*

$$\pi_3 = \sum_{i=0}^{+\infty} (1-p)^i p \, P_2^i \pi_1 \, ,$$

when $P_2^i \pi_1$ is the transform of π_1 under i transitions using K_2.

Special special case: K_3 is uniformly ergodic:

$$K_3(x,y) \geq \varepsilon\nu(y), \qquad \forall\, x \in \mathcal{X} ,$$

Mixture decomposition:

$$\begin{aligned} K_3(x,y) &= \varepsilon\nu(y) + (1-\varepsilon)\frac{K_3(x,y) - \varepsilon\nu(y)}{1-\varepsilon} \\ &= \varepsilon\nu(y) + (1-\varepsilon)K_2(x,y) \end{aligned}$$

Representation of the stationary distribution:

$$\sum_{i=0}^{+\infty} \varepsilon(1-\varepsilon)^i P_2^i \nu ,$$

where P_2 is associated with K_2

1. Simulate $x_0 \sim \nu$, $\omega \sim \mathcal{G}eo(\varepsilon)$.

2. Run the transition $x_{t+1} \sim K_2(x_t, y) \quad t = 0, \cdots, \omega - 1$,
   and take x_ω.

[Murdoch and Green, 1998]

4. General case

Minorizing condition

$$K_3(x,y) \geq \varepsilon \nu(y) \mathbb{I}_C(x) \qquad [MNRZ]$$

Splitting decomposition

$$\begin{aligned} K_3(x,y) &= \left\{ \varepsilon\nu(y) + (1-\varepsilon) \frac{K_3(x,y) - \varepsilon\nu(y)}{1-\epsilon} \right\} \mathbb{I}_C(y) + K_3(x,y)\mathbb{I}_{C^c}(y) \\ &= \{\varepsilon\nu(y) + (1-\varepsilon)K_2(x,y)\}\mathbb{I}_C(y) + K_3(x,y)\mathbb{I}_{C^c}(y) \end{aligned}$$

[Nummelin, 1984]

K_2 is the *depleted measure* of K_3. Simulation?

Introduction of the *split chain* $\Phi^\star = \{(X_n, \delta_n)\}_n$, on $\mathcal{X} \times \{0,1\}$, with transition kernel

$$P'\left[(x,0), A \times \delta\right] = \begin{cases} \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] K_3(x,A) & x \notin C \\ \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] K_2(x,A) & x \in C \end{cases}$$

and

$$P'\left[(x,1), A \times \delta\right] = \begin{cases} \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] K_3(x,A) & x \notin C \\ \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] \nu(A) & x \in C \end{cases}$$

where $\delta \in \{0,1\}$ *(renewal indicator)*

[Athreya and Ney, 1984]

Then $\alpha := C \times \{1\}$ is an *accessible atom*

1. `Simulate` $X_n \sim K_3(x_{n-1}, \cdot\,)$

2. `Simulate` δ_{n-1} `conditional on` (x_{n-1}, x_n)

$$\Pr(\delta_{n-1} = 1 | x_{n-1}, x_n) = \frac{\varepsilon\nu(x_n)}{K_3(x_{n-1}, x_n)}$$

[Mykland, Tierney and Yu, 1995]

5. General Mixture Representation

Let τ_α be the first return time to α

$$\tau_\alpha = \min\{n \geq 1 : (X_n, \delta_n) \in \alpha\} \ .$$

and

$$\Pr_\alpha(\cdot) \quad \text{and} \quad \mathrm{E}_\alpha(\cdot)\,,$$

probability and expectation conditional on $(X_0, \delta_0) \in \alpha$

Tail renewal time T^*

$$\Pr(T^* = t) = \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\mathrm{E}_\alpha(\tau_\alpha)}$$

If the chain is recurrent, $\mathrm{E}_\alpha(\tau_\alpha) < \infty$

Theorem 4 *If $(X_n)_n$ is μ-irreducible, aperiodic, and Harris recurrent with invariant probability distribution π, with a minorization condition [MNRZ], then*

$$\pi(A) = \sum_{t=1}^{\infty} \Pr(N_t \in A) \Pr(T^* = t)$$

where N_t is equal in distribution to X_t given $X_1 \sim \nu(\cdot)$ and given no regenerations before time t.

Follows from Kac's theorem

$$\pi(A) = \frac{1}{\mathrm{E}_\alpha(\tau_\alpha)} \sum_{t=1}^{\infty} \Pr_\alpha(X_t \in A, \tau_\alpha \geq t)$$

Can be extended to stationary measures

6. Estimating the Mass Function of T^* via iid Samples

$$\begin{aligned} p_t &= \Pr(T^* = t) = \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\mathrm{E}_\alpha(\tau_\alpha)} \\ &= \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\sum_{s=1}^{\infty} \Pr_\alpha(\tau_\alpha \geq s)} \end{aligned}$$

$\tau_{\alpha,1}, \ldots, \tau_{\alpha,n}$ iid sample of T^*'s
$F_n(t)$ corresponding empirical cdf

Estimate of p_t

$$\hat{p}_t = \frac{1 - F_n(t-1)}{1 + \sum_{s=1}^{\infty}[1 - F_n(s)]} = \frac{1 - F_n(t-1)}{\overline{\tau}_\alpha}$$

with $\overline{\tau}_\alpha$ sample mean.

Evaluation of the error

$$\left\| \sum_{t=1}^{\infty} N_t(\cdot)\hat{p}_t - \pi(\cdot) \right\| = \left\| \sum_{t=1}^{\infty} N_t(\cdot)\hat{p}_t - \sum_{t=1}^{\infty} N_t(\cdot)p_t \right\| \leq \sum_{t=1}^{\infty} |\hat{p}_t - p_t|$$

and

$$\frac{1}{2}\sum_{t=1}^{\infty} |\hat{p}_t - p_t| \leq \sum_{t=1}^{\infty} |F_n(t) - F(t)| = \int_{-\infty}^{\infty} |F_n(t) - F(t)|\, dt = d_1(F_n, F)$$

L_1-Wasserstein distance between probability distributions

Moreover,

$$\sqrt{n}\, d_1(F_n, F) \xrightarrow{d} \sum_{t=1}^{\infty} |B(F(t))| \approx \sum_{t=1}^{\infty} |B(F_m(t))|$$

7. Example

Bayesian hierarchical version of one-way random effects model

$$Y_{ij}|\theta_i, \lambda_e \sim \mathrm{N}(\theta_i, \lambda_e^{-1})$$

$$\theta|\mu, \lambda_\theta \sim \mathrm{N}(\mu\mathbf{1}, \lambda_\theta^{-1}\mathbf{I}) \quad \text{and} \quad \lambda_e \sim \mathrm{Exp}(1),$$

$$\mu \sim \mathrm{N}(0,1) \quad \text{and} \quad \lambda_\theta \sim \mathrm{Exp}(1).$$

Parameters $\xi = (\theta, \mu)$ and $\lambda = (\lambda_e, \lambda_\theta)$

Simulation from

$$\pi(\xi, \lambda|y) \propto f(y|\theta, \lambda_e) f(\theta|\mu, \lambda_\theta) f(\lambda_e) f(\mu) f(\lambda_\theta)$$

Two-stage Gibbs sampler [Data Augmentation]

$$(\lambda', \xi') \longrightarrow (\lambda, \xi') \longrightarrow (\lambda, \xi).$$

Markov transition density

$$k(\lambda, \xi|\lambda', \xi') = \pi(\lambda_\theta|\xi', y)\ \pi(\lambda_e|\xi', y)\ \pi(\xi|\lambda, y)$$

where $\pi(\lambda_\theta|\xi', y)$ and $\pi(\lambda_e|\xi', y)$ gamma densities and $\pi(\xi|\lambda, y)$ multivariate normal

[Jones and Hobert, 2000]

Minorization condition

$$
\begin{aligned}
\mathrm{SSE} &= \sum_{i,j}(y_{ij}-\overline{y}_i)^2, \; M_T = \sum_i m_i, \\
v_1(\xi) &= \sum_{i=1}^{K}(\theta_i-\mu)^2 \\
v_2(\xi) &= \sum_{i=1}^{K} m_i(\theta_i-\overline{y}_i)^2 \, .
\end{aligned}
$$

Then, for $c_1, c_2 > 0$

$$\pi(\lambda_\theta|\xi', y)\pi(\lambda_e|\xi', y)\pi(\xi|\lambda, y) \geq \varepsilon q(\lambda, \xi) \quad \forall(\lambda', \xi') \in C$$

where

$$C = \{(\lambda, \xi) : v_1(\xi) < c_1 \;\text{ and }\; v_2(\xi) < c_2\}$$

Define

$$g(\lambda_\theta) = \begin{cases} \text{Gamma}\left(\frac{K}{2}+1, 1; \lambda_\theta\right) & 0 < \lambda_\theta < \lambda_\theta^* \\ \text{Gamma}\left(\frac{K}{2}+1, \frac{c_1}{2}+1; \lambda_\theta\right) & \lambda_\theta \geq \lambda_\theta^* \end{cases}$$

and

$$h(\lambda_e) = \begin{cases} \text{Gamma}\left(\frac{M_T}{2}+1, \frac{\text{SSE}}{2}+1; \lambda_e\right) & 0 < \lambda_e < \lambda_e^* \\ \text{Gamma}\left(\frac{M_T}{2}+1, \frac{\text{SSE}+c_2}{2}+1; \lambda_e\right) & \lambda_e \geq \lambda_e^* \end{cases}$$

where

$$\lambda_\theta^* = \frac{(K+2)}{c_1}\log\left(1+\frac{c_1}{2}\right) \quad \text{and} \quad \lambda_e^* = \frac{(M_T+2)}{c_2}\log\left(1+\frac{c_2}{2+\text{SSE}}\right).$$

Then

$$q(\lambda, \xi) \propto g(\lambda_\theta)\, h(\lambda_e)\, \pi(\xi | \lambda, y)$$

and

$$\epsilon = \int_0^\infty g(\lambda_\theta)\, d\lambda_\theta \int_0^\infty h(\lambda_e)\, d\lambda_e$$

Illustration

Table 1: Simulated Data

Group	1	2	3	4	5	6
m_i	5	5	5	10	10	10
$\bar{y}_i$	-0.107	-0.169	.0052	-0.229	-0.259	-0.149

$$M_T = \sum_{i=1}^{6} m_i = 45$$

$$\text{SSE} = \sum_{i=1}^{6} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2 = 14.928$$

Choice of $c_1 = 2$ and $c_2 = 3$

Recall

$$\sqrt{n}\, d_1(F_n, F) \xrightarrow{d} \sum_{t=1}^{\infty} |B(F(t))| \approx \sum_{t=1}^{\infty} |B(F_m(t))|$$

Determination of the sample size for

$$\Pr\left(\left\| \sum_{t=0}^{\infty} N_t(\cdot)\hat{p}_t - \pi(\cdot) \right\| < 0.01 \right) > 0.95$$

Initial sample of $m = 100,000$ iid τ_α's for the empirical distribution leads to

$$n^\star = \frac{4(7.211)^2}{(0.01)^2} \approx 2.08 \times 10^6$$

Additional 2.08×10^6 iid τ_α's lead to new estimate of F and of mass function of T^*.

Table 2: Estimate of the mass function of T^* (to seven decimal places)

t	$\hat{p}_t$	t	$\hat{p}_t$
1	.2541461	26	.0001774
2	.1884236	27	.0001357
3	.1406596	28	.0001031
4	.1051206	29	.0000789
5	.0786641	30	.0000593
6	.0588587	31	.0000441
$\vdots$	$\vdots$	$\vdots$	$\vdots$
25	.0002363	50	.0000001

8. Simulation of N_t

Solution 1: Brute force accept-reject implementation

`If` $T^* = t$`,`

1. `Run the original Markov kernel` P `for` $\Phi^\star$ t `steps`
2. `If no regeneration has occured, accept` Φ_t`.`
   `Otherwise, repeat.`

Solution 2: Recycle the sequences $x_0^i, \ldots, x_{\tau_{\alpha,i}}^i$ $(i = 1, \ldots, I)$ produced during the simulation of the $\tau_{\alpha,i}$'s.

$$
\begin{aligned}
\mathrm{E}^{\pi}[h(X)] &= \sum_{t=0}^{\infty} P(T^* = t)\mathrm{E}[h(N_t)] \\
&= \frac{\sum_{t=0}^{\infty} \mathrm{E}\left[h(X_t)\delta_{T \geq t}\right]}{1 + \mathrm{E}[T]} \\
&\approx \frac{\sum_{t=0}^{\infty} \sum_{i=1}^{I} h(x_t^i)\delta_{\tau_{\alpha,i} \geq t}}{\sum_{i=1}^{I}(\tau_{\alpha,i} + 1)} \\
&= \frac{\sum_{i=1}^{I} \sum_{j=0}^{\tau_{\alpha,i}} h(x_j^i)}{\sum_{i=1}^{I}(\tau_{\alpha,i} + 1)}
\end{aligned}
$$

9. Perfect slice samplers

Generation from

$$\mathcal{U}\left(\{\omega\,;\,\pi(\omega) \geq u\pi(\omega_0)\}\right)$$

and $u \sim \mathcal{U}([0,1])$

- Slice samplers do not require normalising constants
- Slice samplers induce a natural order

  If $\quad \pi(\omega_1) \leq \pi(\omega_2)$

  $$\mathcal{A}_2 = \{\omega\,;\,\pi(\omega) \geq u\pi(\omega_2)\} \subset \mathcal{A}_1 = \{\omega\,;\,\pi(\omega) \geq u\pi(\omega_1)\}$$

- Slice samplers induce a natural discretization of continuous state space

[Mira, Møller & Roberts, 1999]

- Slice samplers preserve monotonicity :

(a) Start from $\tilde{0} = \arg\min \pi(\omega)$ and $\tilde{1} = \arg\max \pi(\omega)$

(b) Generate $u_{-t}, \ldots, u_0$

(c) Get the successive images of $\tilde{0}$ for $t = -T, \ldots, 0$

(d) Check if those are acceptable as successive images of $\tilde{1}$
If not, generate the corresponding images

- **But....** slice samplers are real hard to implement:

$$\mathcal{U}\left(\left\{\theta\,;\,\prod_{i=1}^{n}\sum_{j=1}^{k}p_j f(x_i \mid \theta_j) \geq \epsilon\right\}\right)$$

impossible to simulate

10. Duality Principle (is back !)

`Dual marginalization:` integrate out the parameters (θ, p) in

$$\mathbf{z}, \theta \mid \mathbf{x} \sim \pi(\theta, p) \prod_{i=1}^{n} p_{z_i} f(x_i \mid \theta_{z_i})$$

Easily done in conjugate (exponential) settings.

Use the slice sampler on the marginal posterior of $\mathbf{z}$

- Finite state space
- Link with Rao–Blackwellisation
- Perfect sampling on $\mathbf{z}$ equivalent to perfect sampling on θ

Exponential example ($k = 2, p$ known)

Joint distribution

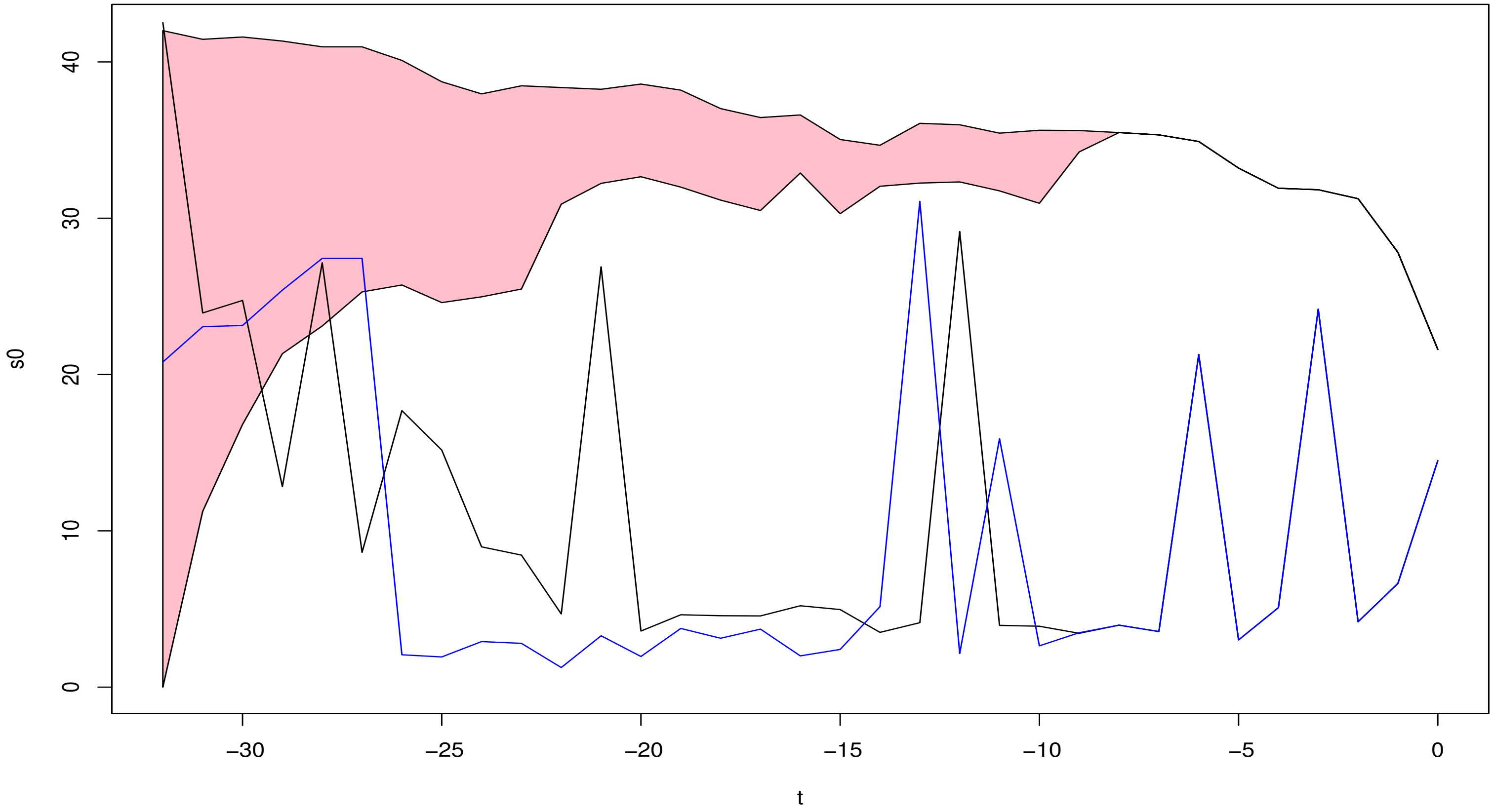
$$\prod_{i=1}^{n} p^{(1-z_i)}(1-p_i)^{z_i}\lambda_{z_i}\exp(-\lambda_{z_i}x_i)\prod_{j=1}^{k}\lambda_j^{\alpha_j-1}\exp(-\lambda_j\beta_j)$$

leads to

$$\mathbf{z} \mid \mathbf{x} \sim p^{n_0}(1-p)^{n_1}\frac{\Gamma(\alpha_0+n_0-1)\Gamma(\alpha_1+n_1-1)}{(\beta_0+s_0)^{\alpha_0+n_0}(\beta_1+s_1)^{\alpha_1+n_1}}\,.$$

- Closed form computable expression (up to constant)
- Factorises through (n_0, s_0), sufficient statistic
- Maximum $\tilde{1}$ and minimum $\tilde{0}$ can be derived

- **But...** slice sampler still difficult to implement because of number of values of s_0 : $\binom{n}{n_0}$
  Still, feasible for small values of n $(n \leq 40)$



Fixed n_0, 40 observations

Perfect sampling is possible!

Idea: Use Breyer and Roberts' (1999) automatic coupling:

If

$$x_1^{(t+1)} = \begin{cases} y_t \sim q(y|x_1^{(t)}) & \text{if } u_t \leq \dfrac{\pi(y_t)\, q(x_1^{(t)}|y_t)}{\pi(x_1^{(t)})\, q(y_t|x_1^{(t)})}, \\ x_1^{(t)} & \text{otherwise.} \end{cases}$$

generate

$$x_2^{(t+1)} = \begin{cases} y_t & \text{if } u_t \leq \dfrac{\pi(y_t)\, q(x_2^{(t)}|x_1^{(t)})}{\pi(x_2^{(t)})\, q(y_t|x_1^{(t)})}, \\ x_2^{(t)} & \text{otherwise.} \end{cases} \qquad (1)$$

Theorem *In the special case*

$$q(y|x) = h(y),$$

if $(x_1^{(t)})$ *starts from*

$$\tilde{0} = \arg\min \pi/h,$$

if $(x_2^{(t)})$ *starts from*

$$\tilde{1} = \arg\max \pi/h,$$

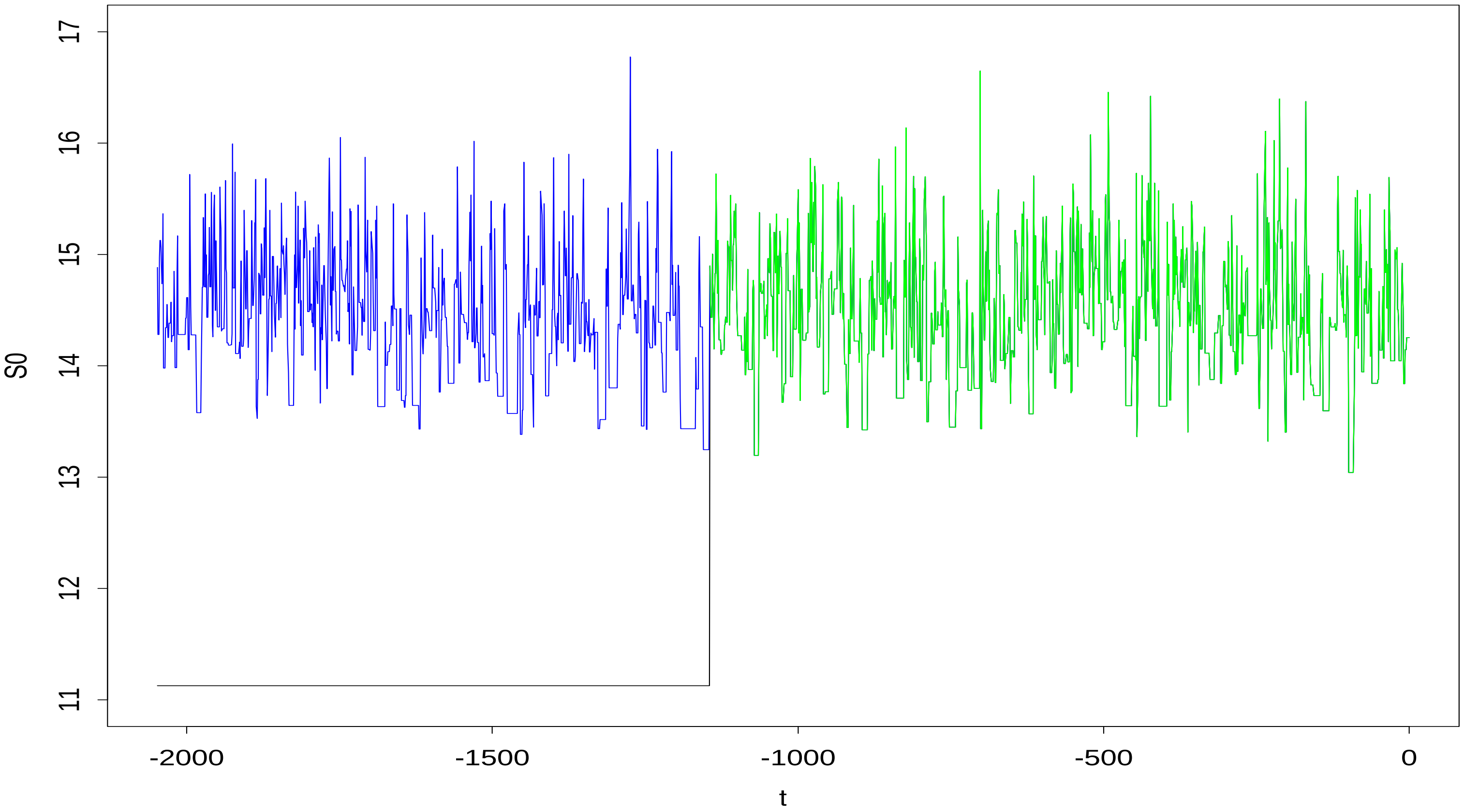
the coupling (1) preserves the ordering.

[Now, this is a result from Corcoran and Tweedie!!!]

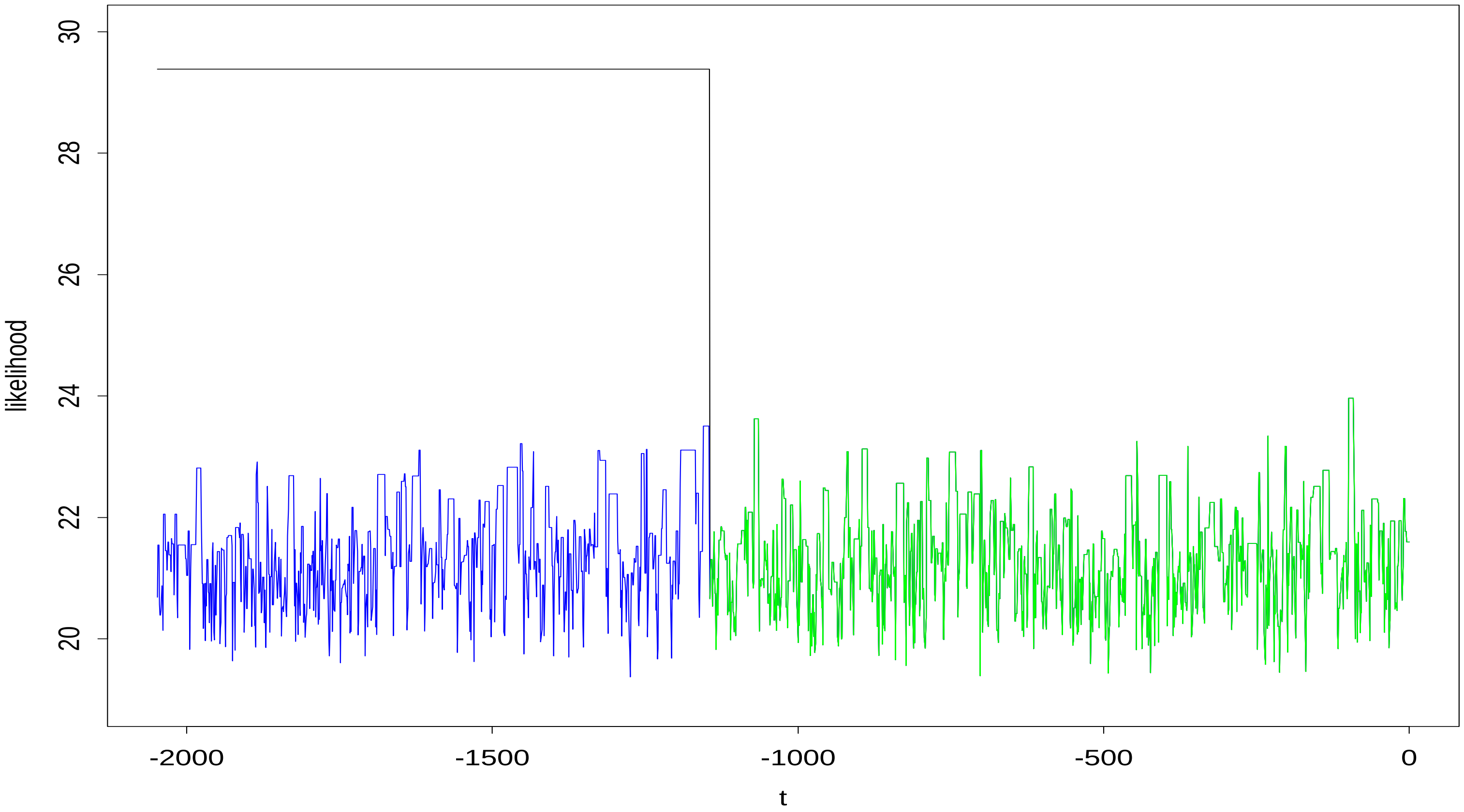
Example When state space $\mathcal{X}$ compact,
use for h the uniform distribution on $\mathcal{X}$.

Extremal elements $\tilde{0}$ and $\tilde{1}$ then induced by π only.

Implementation: start from arbitrary value for $x_1^{(0)}$ and keep proposing for $x_2^{(0)} = \tilde{1}$



Coupling history



Corresponding likelihoods

11. Back to Basics!

When $\mathcal{X}$ compact, and $\pi(x) \leq \pi(\tilde{1})$, independent Metropolis–Hasting coupling **is accept–reject**, based on uniform proposals.

Reason: When coupling occurs, $x_2^{(t)} = y_t$,

$$u_t \leq \frac{\pi(y_t)}{\pi(\tilde{1})} = \frac{\pi(y_t)}{\max \pi}$$

and therefore the chain is in stationnary regime at coupling time.

This extends to the general case, with accept–reject based on proposal h.

- In this case, the accept–reject algorithm could have been conceived independently from perfect sampling (?)
- Fill's (1998) algorithm is an accept–reject algorithm in disguise, but it could not have been conceived independently from perfect sampling

[Casella, Lavine and Robert, 2001]

- Other cases?