Topics in Retrospective simulation

Gareth Roberts

Department of Statistics, University of Warwick

FoCSML, January 2025

Part 3: Perfect simulation

The problem

1. We are interested in the invariant distribution π of a given Markov chain P.
2. We have a given invariant distribution π and have one (or many) Markov chain for which it is the invariant distribution.

1 is steady state analysis, important in many fields, including stochastic geometry. 2 is MCMC.
Under suitable irreducibility, aperiodicity assumptions, $P(X_k \in A|X_0 = x_0)$ should converge to $\pi(A)$. But how large should k be?

We want to avoid the need to make the approximation: for "large enough" k, $P(X_k \in A | X_0 = x_0) \approx \pi(A)$.

Perfect simulation provides a collection of methods to do this.

It is closely related but different to coupling.

Will discuss:

- Coupling from the past
- Read-once CFTP
- Catalytic coupling
- One-shot coupling
- Some monotonicity methods and dominated CFTP

Coupling From The Past (CFTP) (Propp & Wilson 1996)

Imagine running an ergodic Markov chain started at time "$-\infty$". It is reasonable to hope that by time 0 it will have converged. Perhaps we can acertain its value at time 0 without going back as far as time $-\infty$?

We consider constructing the Markov chain using a collection of independent **random maps**:

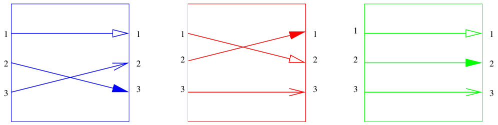
$$\dots, C^{[-1]}, C^{[0]}, C^{[1]}, C^{[2]}, \dots$$

with the property that $C^{[i]}(x)$ distributed according to $P(x, \cdot)$. Then set $C^{[i,j]}$ to be the composite random map:

$$C_{i,j} = C^{[j]} C^{[j-1]} \dots C^{[i+1]} C^{[i]}$$

so that given starting value, $X_0 = x$, $X_n = C^{[1,n]}(x)$ has distribution $P^n(x, \cdot)$.

What a random map looks like



P(🟦) = P(🟥) = P(🟩) = 1/3

$$P = \begin{pmatrix} 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{pmatrix}$$

But this construction is not unique - many constructions for random maps can lead to the same P.

Also $C^{[-n+1,0]}(x) \sim P^n(x, \cdot)$. So the limiting distribution of $C^{[-n+1,0]}(x)$ as $n \to \infty$ is π.

Moreover it **might** be true that $\lim_{n\to\infty} C^{[-n+1,0]}(x)$ exists or even that there exists a time N such that for $n \geq N$,

$$C^{[-n+1,0]}(x) = C^{[-N+1,0]}(x) \ .$$

If we can identify this limit, we have an exact observation from π.

For $N \leq n$ we have that

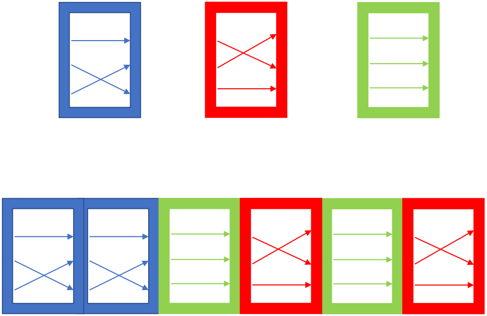
$$C^{[-n+1,0]} = C^{[-N+1,0]} C^{[-n+1,-N]}$$

Look at the **image** of the map $C^{[-N+1,0]}$, ie $I_N = \{C^{[-N+1,0]}(y), y \in \mathfrak{X}\}$. Suppose I_N contains just one value, z say (**coalescence**) then for all y, $C^{[-N+1,0]}(y) = z$ and in particular

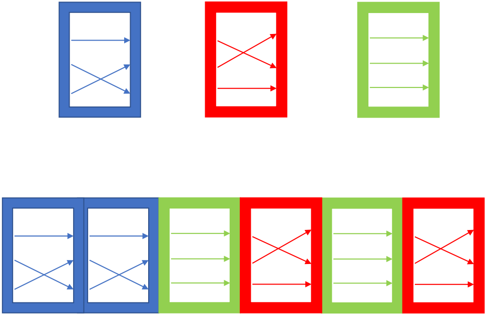
$$C^{[-N+1,0]} C^{[-n+1,N]}(x) = z$$

for all $n \geq N$, $x \in \mathfrak{X}$.

Will it work?

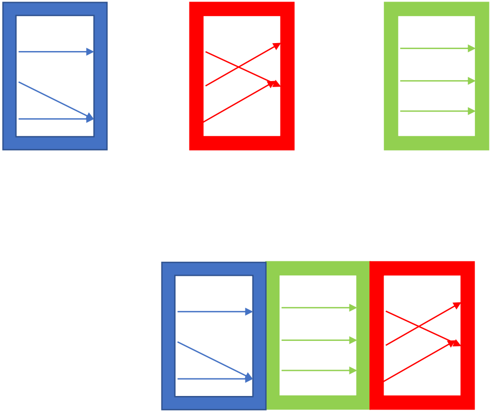


Will it work?

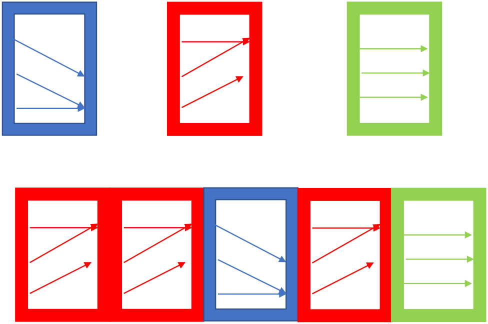


The coupling merely permutes states. Coupling never achieved.

Improved coupling



A monotone coupling



Monotonicity helps to keep track of coalescing states (more later) which becomes important in harder examples.

There is no reason why coalescence might be achieved for any N for any given construction of the random updates. However we have a great deal of **freedom** in choosing $\{C^{[i]}\}$. Some may be far more effective coalescers than others.

Practical considerations such as keeping track of coalescence become crucial in more complicated problems.

This construction does **not** require that $\{C^{[i]}\}$ are iterations from the **same** Markov chain, just independent iterations which are stationary with respect to π.

Uniform ergodicity

We say that a Markov chain X is uniformly ergodic if there exists a constant $\rho < 1$ and constant A such that for all $x \in \mathfrak{X}$

$$\|P^n(x,\cdot) - \pi\| \leq A\rho^n \ .$$

Here we use $\| \cdot \|$ to denote total variation distance (generally one of the hardest metric to work with) although similar theory can be derived for other metrics such as Wasserstein metrics.

This has enormous practical and theoretical advantages. There do not exist arbitrarily bad starting values.

Most MCMC algorithms are not uniformly ergodic. But we can sometime manipulate them to ensure uniform ergodicity.

Uniform ergodicity for a simple MCMC algorithm

Consider the independence sampler.

Interested in π but can simulate directly from q with the same support as π.

Given X_n, propose a new value $Y_{n+1} \sim q$ and accept wp

$$\min\left\{1, \frac{\pi(Y_{n+1})q(X_n)}{q(Y_{n+1})\pi(X_n)}\right\} \ .$$

Uniform ergodicity for a simple MCMC algorithm

Consider the independence sampler.

Interested in π but can simulate directly from q with the same support as π.

Given X_n, propose a new value $Y_{n+1} \sim q$ and accept wp

$$\min\left\{1, \frac{\pi(Y_{n+1})q(X_n)}{q(Y_{n+1})\pi(X_n)}\right\} \ .$$

Theorem (Mengersen and Tweedie 1994)
The independence sampler is uniformly ergodic if and only if π/q is bounded above.

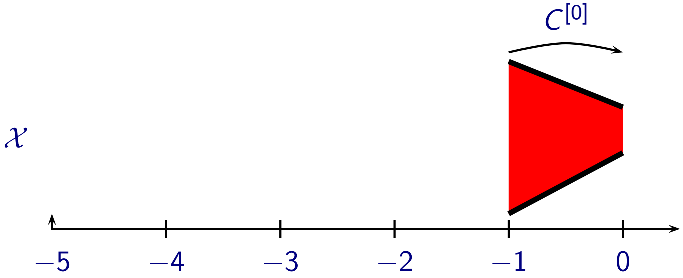
This coupling construction of putting together finite collections of random maps has **no chance** of success unless we have *uniform ergodicity*. The construction requiring that I_N be a singleton is a constructive minorisation condition:

$$P^N(x, \cdot) \geq \epsilon \nu(\cdot)$$

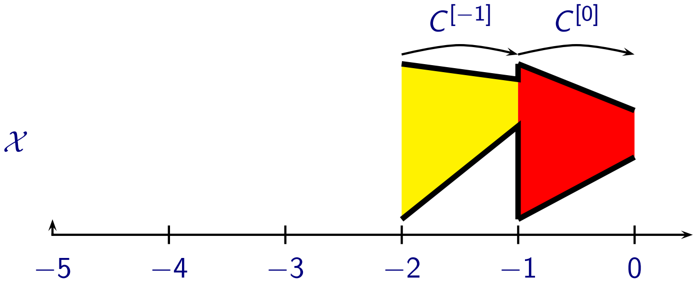
where ϵ denotes the probability of success of the coalescence after N iterations, and ν is the probability measure describing the distribution of the value coalesced to.

From now on we shall take $C^{[i]}$ to be IID draws from a composite Markov chain iteration with invariant distribution π and satisfying the minorisation condition above, often constructed by composing different Markov chain mechanisms.

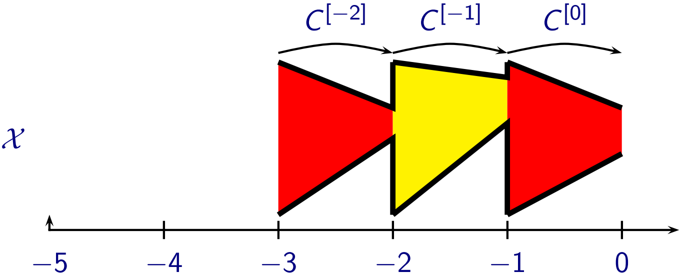
1. Generate independent realisations $C^{[0]}, C^{[-1]}, C^{[-2]}, \dots$ of C "back in time"...



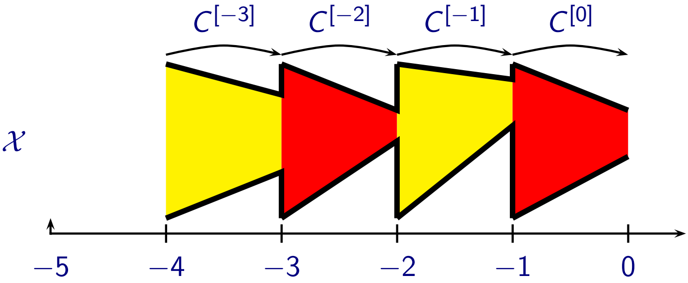
1. Generate independent realisations $C^{[0]}, C^{[-1]}, C^{[-2]}, \dots$ of C "back in time"...



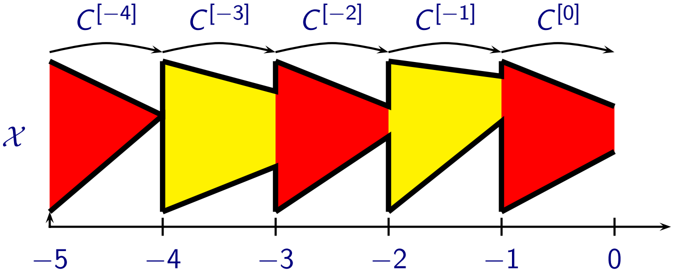
1. Generate independent realisations $C^{[0]}, C^{[-1]}, C^{[-2]}, \ldots$ of C “back in time”...



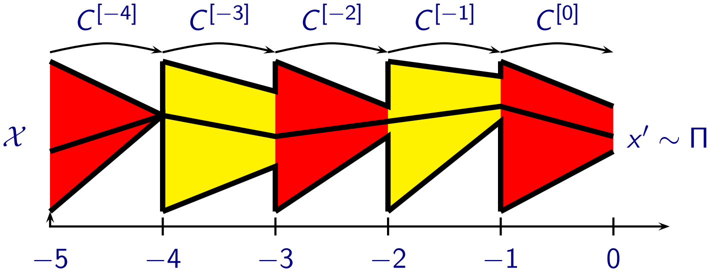
1. Generate independent realisations $C^{[0]}, C^{[-1]}, C^{[-2]}, \ldots$ of C "back in time"...



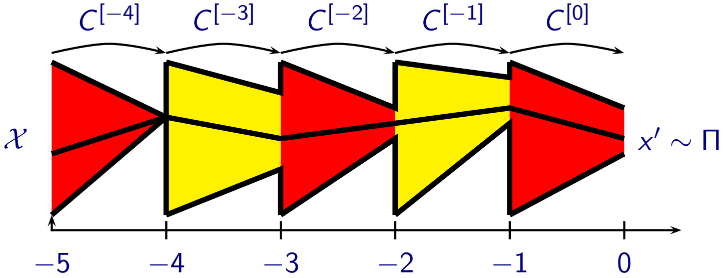
1. Generate independent realisations $C^{[0]}, C^{[-1]}, C^{[-2]}, \ldots$ of C "back in time"...
2. ...until $C^{[-l]}$ is coalescent. Then $C^{[0]} \circ \cdots \circ C^{[-l]}$ is coalescent.



1. Generate independent realisations $C^{[0]}, C^{[-1]}, C^{[-2]}, \ldots$ of C "back in time"...
2. ...until $C^{[-l]}$ is coalescent. Then $C^{[0]} \circ \cdots \circ C^{[-l]}$ is coalescent.
3. ...and $x' = C^{[0]} \circ \cdots \circ C^{[-l]}(\mathcal{X})$ is a sample from Π.



1. Generate independent realisations $C^{[0]}, C^{[-1]}, C^{[-2]}, \ldots$ of C "back in time"...
2. ...until $C^{[-l]}$ is coalescent. Then $C^{[0]} \circ \cdots \circ C^{[-l]}$ is coalescent.
3. ...and $x' = C^{[0]} \circ \cdots \circ C^{[-l]}(\mathcal{X})$ is a sample from Π.
4. Waiting time Geometric(ϵ).



Read-once CFTP (Wilson 2000): derivation from minorisation conditions

Suppose $P(x,\cdot) \geq \epsilon\nu(\cdot)$ (ie $\mathfrak{X}$ is *1-small*). Of course $\pi = \pi P$ so

$$\pi = \pi(\epsilon\nu + (1-\epsilon)Q)$$

where Q is the **residual** kernel ($P^N = \epsilon\nu + (1-\epsilon)Q$). So

$$\pi[I - (1-\epsilon)Q] = \nu$$

and

$$\pi = \nu(1 + (1-\epsilon)Q + ((1-\epsilon)Q)^2 + \ldots)$$

which implies that we can construct π exactly by introducing an auxiliary variable G ~Geom(ϵ) and then:

1. starting from $\nu(\cdot)$
2. running $G-1$ steps of the Markov chain with kernel Q.

The value after $G-1$ steps has distribution π.

Magic! But how can this be implemented?

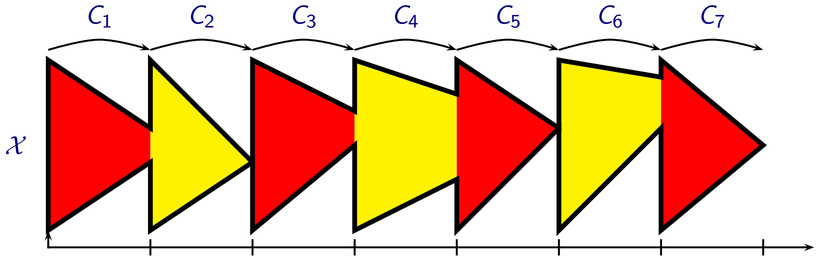
Assume that $C : \mathfrak{X} \to \mathfrak{X}$ is a random function that preserves stationarity w.r.t. target distribution Π, ie.

$$\int_{\mathfrak{X}} \mathbb{P}(C(x) \in A)\Pi(dx) = \Pi(A).$$

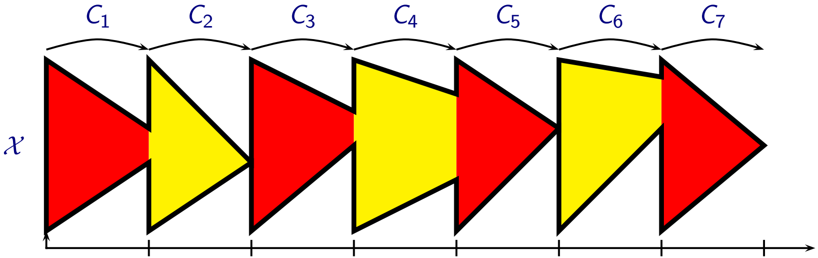
Assume there is a positive probability of C being coalescent, ie. $\# C(\mathfrak{X}) = 1$.

Here $C(\mathfrak{X}) = \{C(x) : x \in \mathfrak{X}\}$ is the image of C
and $\#$ denotes cardinality.

1. Generate independent realisations $C_1, C_2, \ldots$ of C
2. Let $T_0, T_1, \ldots$ denote indices of coalescent functions, ie. $C_{T_i}(\mathcal{X})$ is coalescent.
3. For $i = 1, 2, \ldots$ let $x_i = C_{T_{i-1}} \circ \cdots \circ C_{T_{i-1}}(\mathcal{X})$

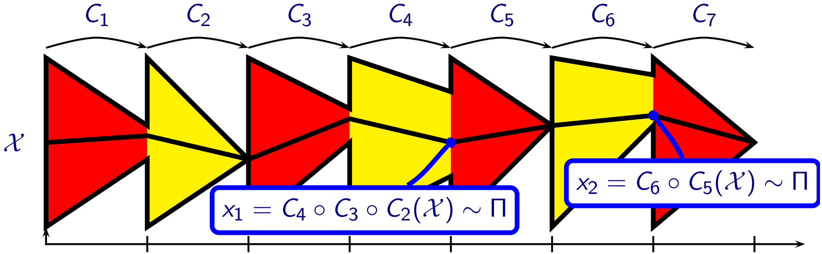


1. Generate independent realisations $C_1, C_2, \ldots$ of C
2. Let $T_0, T_1, \ldots$ denote indices of coalescent functions, ie. $C_{T_i}(\mathcal{X})$ is coalescent.
3. For $i = 1, 2, \ldots$ let $x_i = C_{T_i - 1} \circ \cdots \circ C_{T_{i-1}}(\mathcal{X})$



1. Generate independent realisations $C_1, C_2, \ldots$ of C
2. Let T_0, $T_1, \ldots$ denote indices of coalescent functions, ie. $C_{T_i}(\mathcal{X})$ is coalescent.
3. For $i = 1, 2, \ldots$ let $x_i = C_{T_i-1} \circ \cdots \circ C_{T_{i-1}}(\mathcal{X})$

Then $x_1, x_2, \ldots$ are an IID sample from Π.



$x_1 = C_4 \circ C_3 \circ C_2(\mathcal{X}) \sim \Pi$

$x_2 = C_6 \circ C_5(\mathcal{X}) \sim \Pi$

Coupling vs coalescence

Coupling vs coalescence

Coalescence is **harder** than coupling because

- It is intrinsically more demanding, ie it needs more things to couple.
- It is practically much more complicated. How to keep track of all the trajectories.

For example consider the 3 state chain:

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

Then there is **no** 1-step minorisation, but the whole space is **1-pseudo small** (R+R, 2000, Stochastic models).

When does it work?

In its pure form, both CFTP and RoCFTP requires that $\mathfrak{X}$ is 1-small.

We can extend this to requiring that the state space is n-small by just letting $C^{[i]}$ denote the random map obtained by n steps of P, ie $C^{[i]}(x) \sim P^n(x, \cdot)$.

But this requires P to be uniformly ergodic, and this is rather restrictive.

A further problem is we might not be able to work out whether coalescence has been achieved.

Strategies for making coalescence work

Two issues:

- constructing a suitable uniformly ergodic Markov chain with adequate minorisation probability;
- constructing the random map and the coalescence.

Composite chains

Instead of using n steps of a given Markov chain kernel P, we use $\prod_{i=1}^{n} P_i$, where different P_is have ‘different roles’.
For example:

- **Bounded collapsing** P_1 is a uniformly ergodic ergodic chain for which $\mathfrak{X}$ is 1-small. This is typically possible with an algorithm such as an independence sampler with very fat-tailed proposal. It needn’t be an efficient stand alone method and typically isn’t!
- **Shrinking** $P_2, \ldots P_{n-1}$ are moves of an algorithm typically with some kind of monotonicity and/or contraction property.
- **Coalescence** P_n attempts the coalescence.

This is a **one-shot** strategy (see R+Rosenthal, SPA, 2002).

Keeping track

Typically interested in $C^{[i]}(A)$ where A is a set (often infinite) stored in a suitable format.
Would like to be able to find $B \supset C^{[i]}(A)$ without looking at the image of **all** the elements of A.
Most useful trick is **monotonicity**: there exists a partial ordering on $\mathfrak{X}$, $\leq$ typically with a kind of **well-ordered** property that suitable sets (eg bounded ones) possess infima and suprema. Then the property we require is that

- If $x \leq y$ then there exists a random update function construction F such that

$$F(x) \leq F(y) \quad a.s.$$

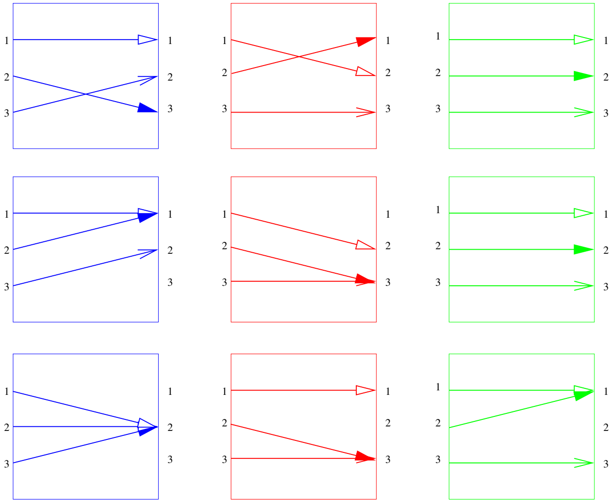
Is this ever possible for real MCMC?

MCMC is typically on complex high-dimensional spaces, not known for their monotonicity.

However the simple structure of many MCMC algorithms reveals many usable monotonicity properties, at least for suitable target densities.

Eg: random walk Metropolis, slice samplers, independence samplers, Langevin methods and Gibbs samplers.

Some random block choices



Example (Beskos and Roberts, 2005)

Consider a d-dimensional Gaussian density with precision matrix $Q = (q_{ij})$ which is positive-definite and $q_{ij} \leq 0$ for $i \neq j$ (a **Stiljes matrix**).

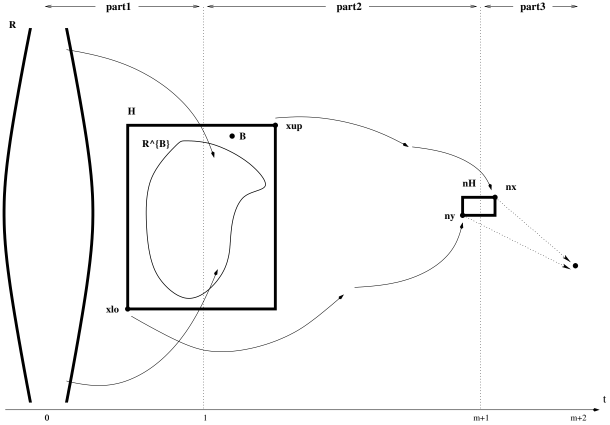
Such a distribution is in the **positive-association** class, and one consequence of this is that X_i and X_j are positively correlated for all i, j.

The partial ordering: $\mathbf{x} \leq \mathbf{y}$ if and only if

$$x_i \leq y_i, \qquad \text{for all } 1 \leq i \leq d$$

is preserved by the Gibbs sampler which uses the same standard normal random seed to update the state starting from each starting value.

Example (Beskos and Roberts, 2005)



Example (Beskos and Roberts, 2005)

As with all coalescence methods, block length is important for algorithm efficiency.

Ths method was used to simulate from multi-dimensional truncated normals. Works well (eg hundreds of IID observations per second in 100 dimensions on a 1500MHz machine).

Computing cost scales well (possibly linearly) with dimension.

A version of the method for general Gaussian densities exists. This is less efficient.

Extensions to other Gibbs samplers possible.

Catalysts (Breyer and R, 2001)

Various specific methods for coalescence in continuous spaces have been proposed (eg Murdoch and Green 1998 and Wilson, 2000) (eg for RWM with symmetric unimodal proposals). See Huber (2016) for recent developments

The catalyst approach has the advantage of being generic.

It starts with a random map and improves it.

Catalysts (Breyer and R, 2001)

Let $F(x) \sim P(x, \cdot)$ be a random map (as part of a coalescence block). F may have very poor coalescence properties. But we can attempt to improve it!

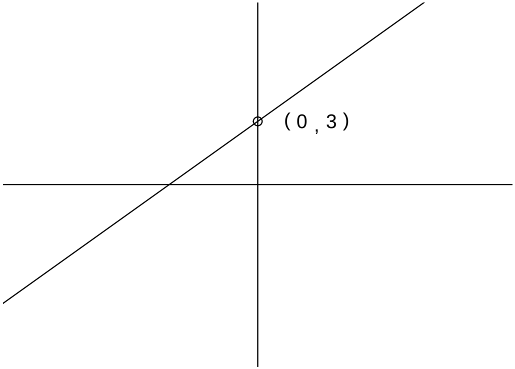
- Choose a location for applying a catalyst, x^* say.
- Simulate $Y \sim P(x^*, \cdot)$.
- Simulate $U \sim U(0,1)$.
- For any x, set $\tilde{F}(x) = Y$ with probability

$$\min\left\{1, \frac{P(x, dY)P(x^*, dF(x))}{P(x^*, dY)P(x, dF(x)}\right\}$$

- Otherwise set $\tilde{F}(x) = F(x)$.

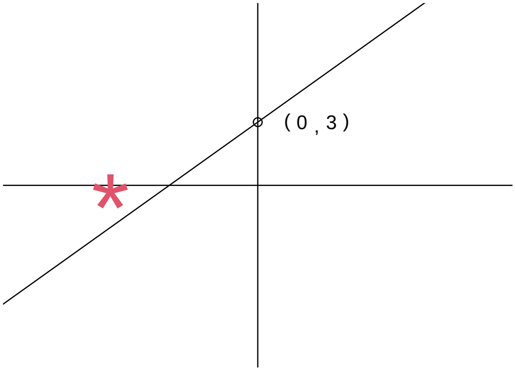
Then $\tilde{F}(x) \sim P(x, \cdot)$ too.
This follows since the algorithm merely carries out (for each x) an independence sampler with invariant distribution $P(x, \cdot)$ and proposal $P(x^*, \cdot)$.
The point is that we use **the same** Y for each x. So this will have a coupling effect.
For suitably continuous P, x values close to x^* will couple with x^*.
Such x are in the **basin of attraction** of (x^*, Y).

Simple illustration of catalyst



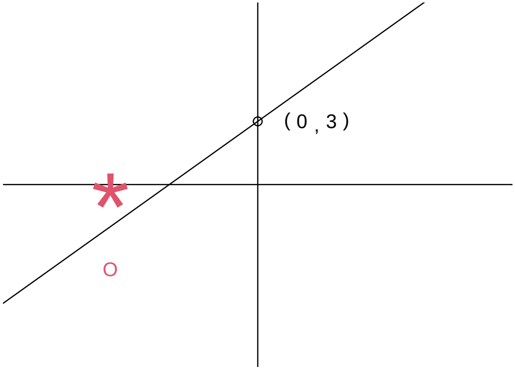
$P(x,\cdot) \sim N(x,1)$, take $F(x) = x + Z$ for $Z \sim N(0,1)$

Simple illustration of catalyst



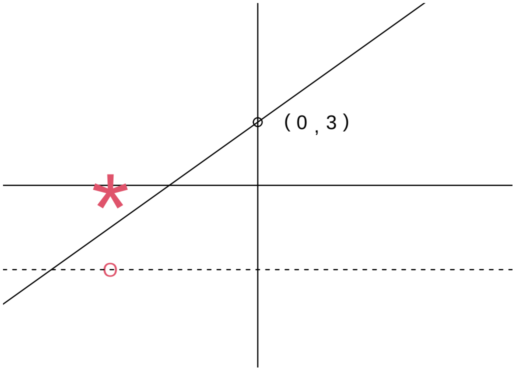
Propose a catalyst at $x^* = -5$.

Simple illustration of catalyst



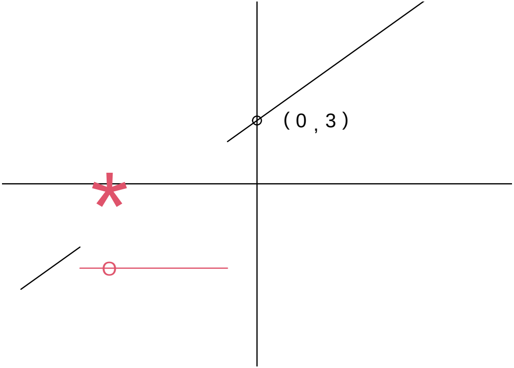
Propose $Y \sim N(x^*, 1)$ in this case $Y = -4$.

Simple illustration of catalyst



Propose Y for every x value.

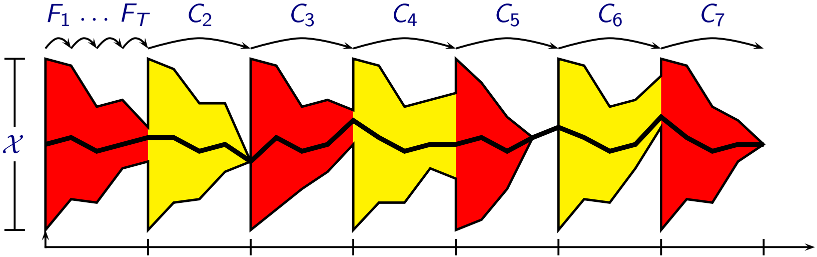
Simple illustration of catalyst



All xs which accept have $\tilde{F}(x) = Y$.

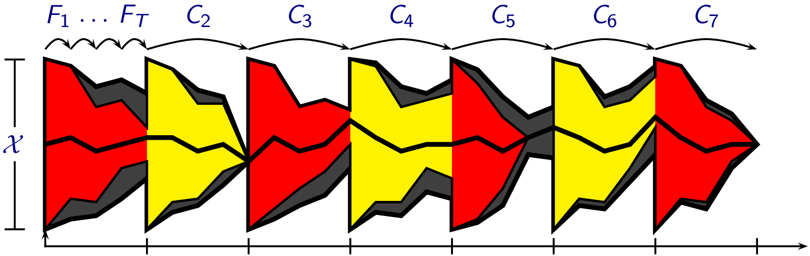
In practise:

- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\mathcal{X})$.
- Redefine T_i, so W_{T_i} is “coalescent”.
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



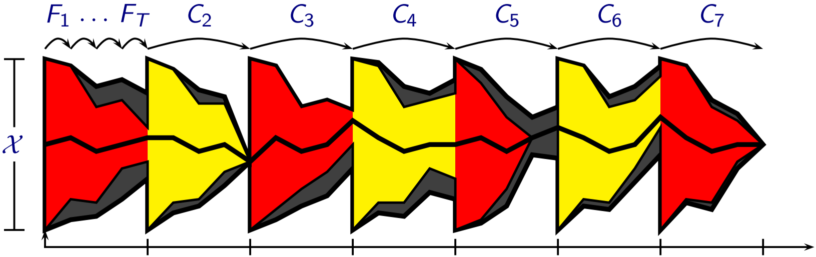
In practise:

- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\mathcal{X})$.
- Redefine T_i, so W_{T_i} is "coalescent".
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



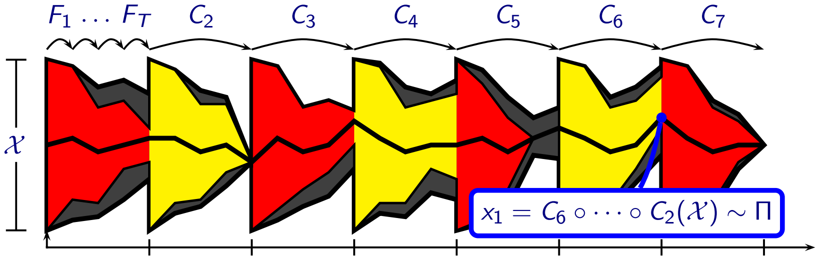
In practise:

- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\mathcal{X})$.
- Redefine T_i, so W_{T_i} is "coalescent".
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



In practise:

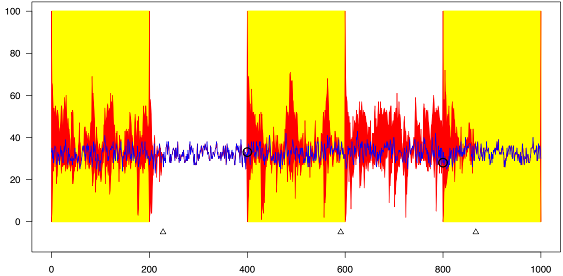
- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\mathcal{X})$.
- Redefine T_i, so W_{T_i} is "coalescent".
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



Bayesian analysis

This example is from a Bayesian analysis of mixture model using data augmentation.

Monotonicity in the allocation space by enlarging state space to allow datat to be allocated to any number of mixture components (including none).



Plot showing N_1

Initially $W = \mathbb{S}_{s,N}$

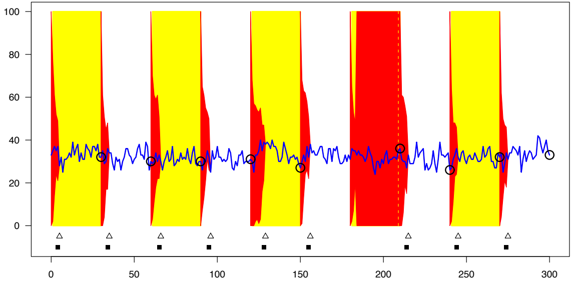
$n = 100$, $r = 3$

$\mu = (0, 1.1, 2.2)$

$m = (\frac{1}{2}, \frac{1}{3}, \frac{1}{3})$

$\sigma^2 = 0.25$

But can do much better (in much shorter time scale and for minimal extra computing cost per iteration) by employing catalysts.



Plot showing N_1

Initially $W = \mathbb{S}_{s,N}$

$n = 100,\ r = 3$

$\mu = (0, 0.5, 1)$

$m = (\frac{1}{2}, \frac{1}{3}, \frac{1}{3})$

$\sigma^2 = 0.25$

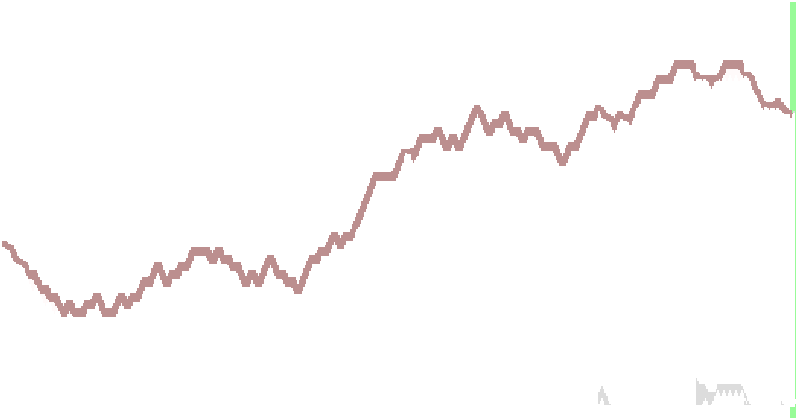
Can use these strategies for perfect Bayesian Monte Carlo inference in realistic moderate sized problems (eg Bayesian analysis of mixtures, hidden Markov models) etc.

But what if we cannot ensure uniform ergodicity?

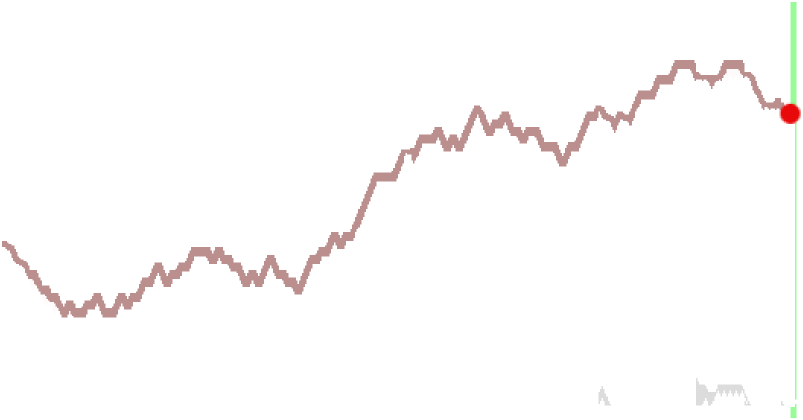
Outside the MCMC context it is not possible to pick and choose our Markov chain mechanism to impose this.

KKC

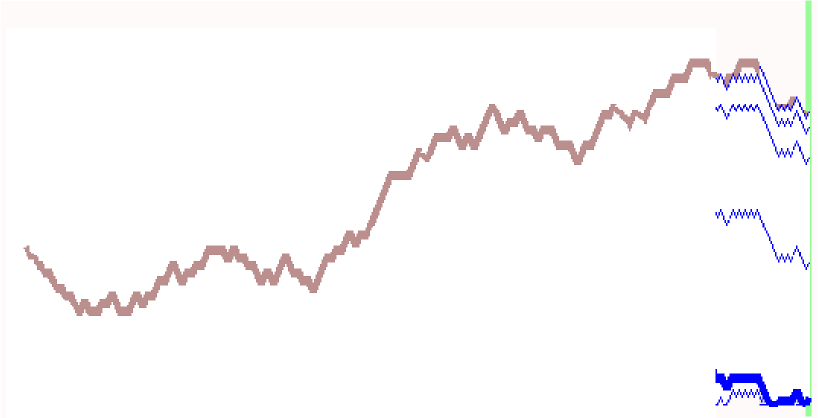
Dominated CFTP (Kendall and Moller 2000)



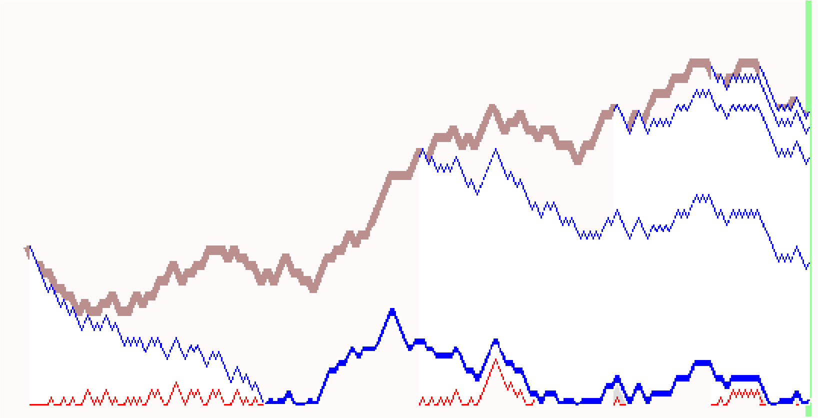
Dominated CFTP (Kendall and Moller 2000)



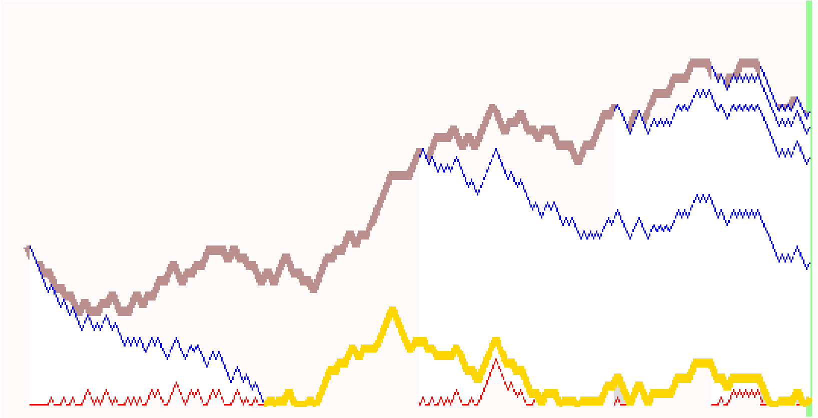
Dominated CFTP (Kendall and Moller 2000)



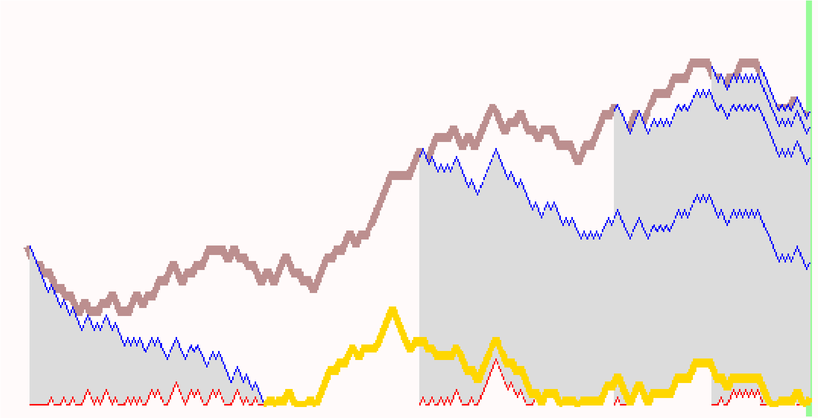
Dominated CFTP (Kendall and Moller 2000)



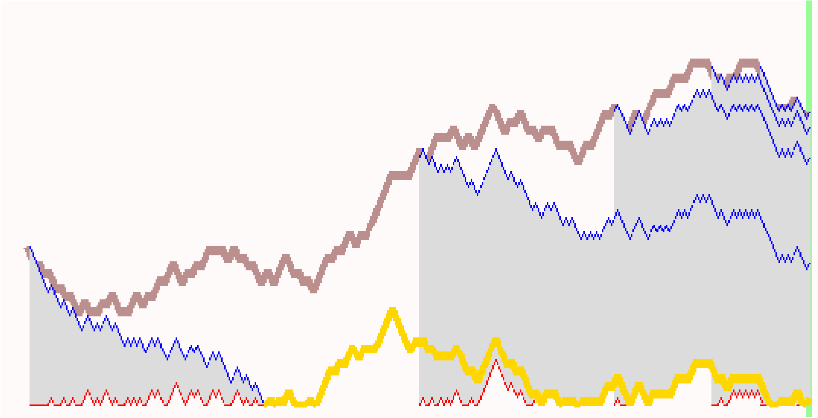
Dominated CFTP (Kendall and Moller 2000)



Dominated CFTP (Kendall and Moller 2000)



Dominated CFTP (Kendall and Moller 2000)



Most successful area for CFTP by far has been in stochastic geometry.

KKCs include: first perfect simulation algorithms for important processes such as the area-interaction point process, cluster point processes, and other interacting point process models.

Also KKCs in CFTP for queues, links to geometric ergodicity, and fundamental questions about small sets.

The future for CFTP

Despite some isolated successes, CFTP has not been found to be practical in the hardest Bayesian computation problems.

Small set constructions usually suffer from the curse of dimensionality in continuous state spaces.

So these methods usually require monotonicity or other structure to be applicable.

Some more recent non-reversible MCM algorithms are explicitly built around practical regenerations and therefore make CFTP readily applicable. See McKimm, Pollock,. Roberts and R (2023)