Branching process Monte Carlo

Peter Green and Antonietta Mira

Universities of Bristol and Insubria
P.J.Green@bristol.ac.uk
Antonietta.Mira@uninsubria.it

ND in MCMC workshop, Warwick, 23 August 2006

What BPMC is

- aimed at a fixed target
- population Monte Carlo, but not as we know it
- dependent on ergodicity
- different, we hope, but speculative

What BPMC isn't

- aimed at a sequentially evolving target
- much to do with particle filters
- dependent on importance sampling
- proven to be useful

Motivation

Branching processes

- can support antithetic behaviour in a natural way by making offspring negatively correlated
- may assist in navigating past slowly-mixing parts of the state space
- are analytically amenable

Contents

We explore the possible use of branching processes for Monte Carlo simulation, & in particular discuss:

- basic theory of branching processes as could be used for Monte Carlo sampling
- the appropriate analogue of global balance with respect to the target distribution
- evaluation of moments, in particular asymptotic variances

Bienaymé/Galton/Watson branching process

Single-sex individuals give birth to random number (*generically*, Y) of probabilistic replicas of themselves – once created, each individual is independent of everything else.

Individuals may have random lifetimes, but we will be satisfied with counting individuals in *generations*: Z_n in generation n.

$$Z_{n+1} = \sum_{i=1}^{Y} Z_n^{(i)} = \sum_{i=1}^{Z_n} Y^{(i)}$$

(where $^{(i)}$ denote probabilistic replicas)
$\Rightarrow$ analysis via composition of probability generating functions:
$g(s) = E(s^Y), g_n(s) = E(s^{Z_n}|Z_0 = 1),$

$$g_{n+1}(s) = g(g_n(s)) = g_n(g(s))$$

Extinction and criticality

The extinction time T is $\min\{n > 0 : Z_n = 0\}$, so we immediately have $P\{T \leq n\} = P\{Z_n = 0\} = g_n(0)$; with a little more work we find that $P\{T < \infty\} = \zeta$, where $\zeta = \inf\{s \geq 0 : g(s) = s\}$.

It is well known that the process has a threshold behaviour depending on $a = g'(1) = E(Y)$, the *mean family size*.

- if $a < 1$ then $\zeta = 1$ and $E(T) < \infty$: *subcritical*
- if $a = 1$ and $\text{var}(Y) > 0$ then $\zeta = 1$ but $E(T) = \infty$: *critical*
- if $a > 1$ then $\zeta < 1$ and so $P(T = \infty) > 0$: *supercritical*

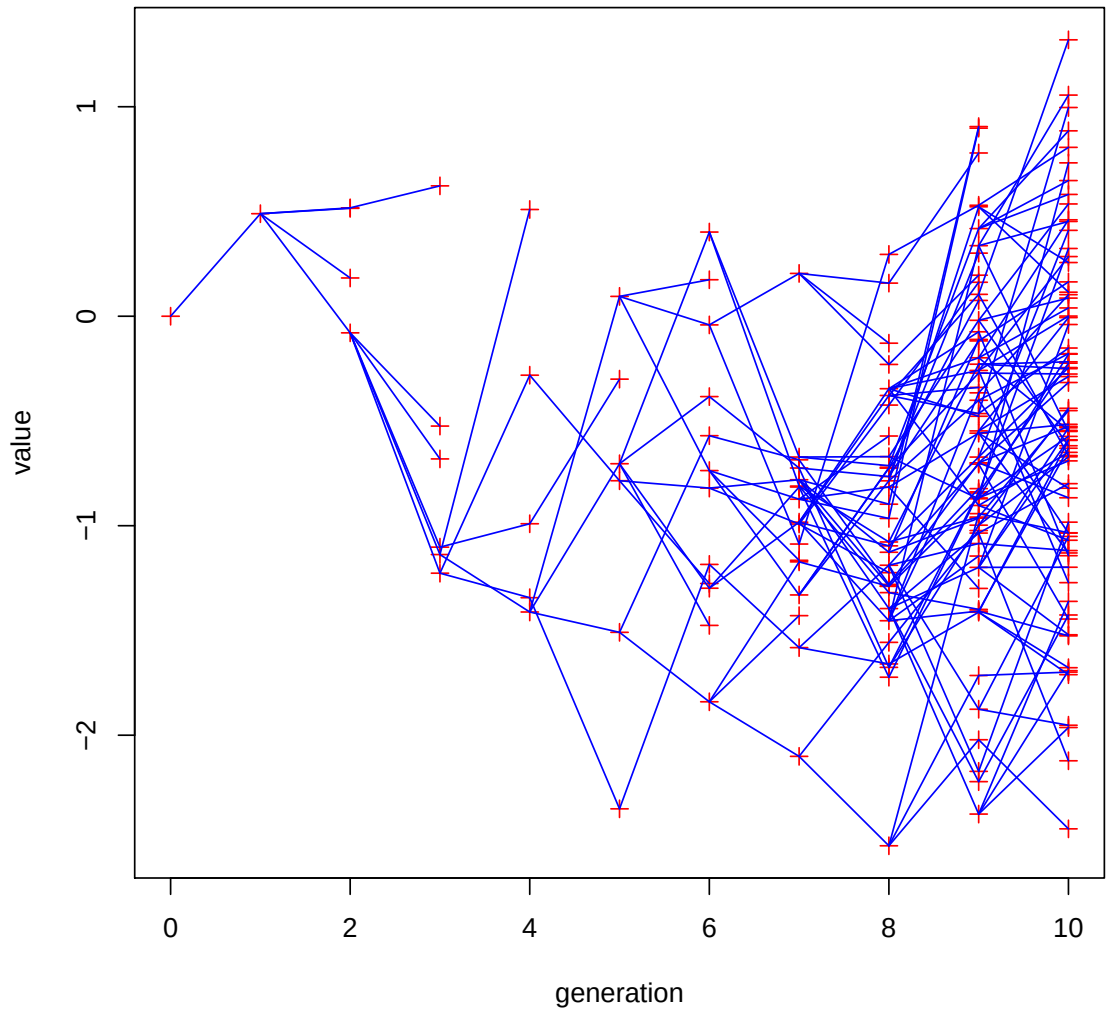
Branching processes and Monte Carlo

For the purposes of Monte Carlo simulation, we associate with each individual in the process a value x in a general state space $\mathcal{X}$.

We are given a *target distribution* π on $\mathcal{X}$, and are interested in efficiently estimating $E_\pi f = \int f(x)\pi(dx)$ for one or more functions $f : \mathcal{X} \to \mathcal{R}$, based on a realisation of our process.

The standard estimator will be the average of the $f(x)$ values of all the $Z_1 + Z_2 + \cdots + Z_n$ individuals in the first n generations of a branching process.

A realisation of BPMC



Branching processes with types

In branching process theory, x corresponds to the 'type' of an individual, a factor introduced into branching process models initially to handle individual characteristics such as sex or age in a biological population, or energy in a nuclear cascade process.

In a branching process with general types ($x \in$ arbitrary $\mathcal{X}$), the population evolves exactly according to the Galton-Watson assumptions made above, but additionally the x values of the individuals propagate through the population, with the x for offspring generated conditionally on their parent's x.

In the language of graphical models, the tree describing the population structure is a directed acyclic graph defining the conditional independence structure of the x process.

Point process notation

In the general-type branching process, the population size Z_n is replaced as the key variable by a finite *point process* on $\mathcal{X}$, giving the types of the individual in the n^{th} generation. Note that several individuals may have the same type, so that this is a point process allowing multiple coincident points. This point process, an integer-valued random measure, will also be denoted by Z_n:

$$Z_n(A) = \#\{\text{individuals in generation } n \text{ with value} \in A\}$$

We suppose the initial individual is of type x, taken as nonrandom but variable for the present, so $Z_0 = \delta_x$ or just x in brief.

The entire distribution of the process is governed by the distribution of Z_1 given $Z_0 = x$.

A convenient representation of this *offspring distribution* uses the *moment generating functional* (MGF) defined for a dummy variable s that is now a nonnegative measurable function from $\mathcal{X}$ to $\mathcal{R}$. The offspring MGF is

$$\Phi^{(x)}(s) = E(e^{-\int s dZ_1} | Z_0 = x).$$

If Z_1 is listed as $\{X^{(i)}, i = 1, 2, \ldots, Y\}$, then we can also write the MGF as

$$\Phi^{(x)}(s) = E(e^{-\sum_{i=1}^{Y} s(X^{(i)})} | Z_0 = x).$$

The basic recurrence obeyed by the process is formally similar to the standard case:

$$Z_{n+1} = \sum_{i=1}^{Y} Z_n^{(i)}$$

but now the $Z_n^{(i)}$ have different initial x-values, so are not i.i.d.

The analogue to the composition of PGFs is also formally similar, but for the effective change of variable from s to $e^{-s(\cdot)}$, and we get the MGF recurrence

$$\Phi_{n+1}^{(x)}(s) = \Phi^{(x)}(-\log \Phi_n^{(\cdot)}(s)).$$

as a generalisation of the earlier PGF recurrence

$$g_{n+1}(s) = g(g_n(s))$$

The recursion

$$\Phi^{(x)}_{n+1}(s) = \Phi^{(x)}(-\log \Phi^{(\cdot)}_n(s)).$$

in principle determines the distribution of Z_n (and can be extended to cover all Z_n jointly). For example, differentiating with respect to (the function) s gives recursions for moments.

In practice, is it usually easier to work from first principles for each moment, using the same condition-on-the-first-family argument.

Moments

We define the *mean kernel* $M(x, B) = E(Z_1(B)|Z_0 = x)$ – the mean number of offspring with values in B born to an individual of value x. Note that this involves integrating over the offspring distribution, but not the value space: in the non-branching case, it is just the transition kernel $P(x, B)$.

Then $E(Z_{n+1}(B)) =$

$$E[E(Z_{n+1}(B)|Z_n)] = E[\int M(\cdot, B)dZ_n] = \int M(\cdot, B)E(dZ_n)$$

So if $E(Z_n(B)) = \mu_n(B)$,

$$\mu_{n+1}(B) = \int \mu_n(dx)M(x, B) = (\mu_n M)(B),$$

say (*we make much use of such operator notation*).

Global balance

Suppose that the process Z_n (whose values are integer-valued measures) is stationary [how? wait and see!]; then $E(Z_n(B)) = \mu(B)$ for all n, and so $E(Z_{n+1}(B)) = (\mu M)(B) = \mu(B)$. Thus

$$\mu M = \mu$$

is the BPMC equivalent of global balance (invariance).

As we will see, averages of $f(x)$ over the population converge to expectations with respect to μ, normalised to be a probability distribution: designing a BPMC sampler means choosing the mean kernel M to satisfy $\mu M = \mu$.

Rules of the game

The objective is to sample from $\pi(dx)$, so as with ordinary MCMC, we can only use π (*up to proportionality*) to create the branching process. For example, the following are valid:

- generate a random number of offspring from a fixed distribution, and draw children's x values i.i.d. from a reversible transition kernel $P(x, B)$ leaving π invariant, e.g. by Metropolis-Hastings
- as above, but P need not be reversible
- as above but children's x's can be dependent draws from $P(x, B)$ (e.g. negatively correlated)
- .. other rules in the same spirit (M constructed implicitly from π) for which μM can be controlled

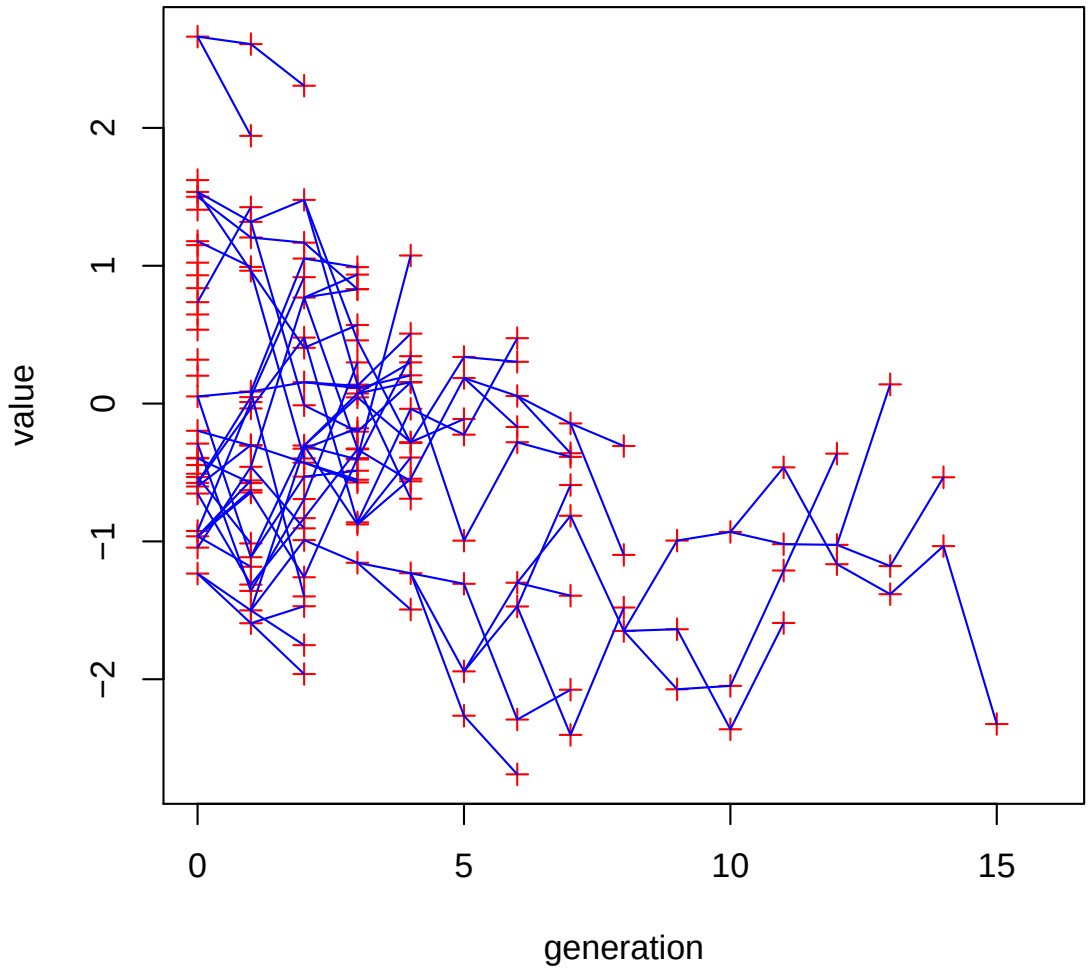
Reconciling branching and stability

Can we follow the branching paradigm strictly (individuals give birth to random number of probabilistic replicas of themselves – once created, each individual is independent of everything else) and still obtain a process that is useful for Monte Carlo simulation?

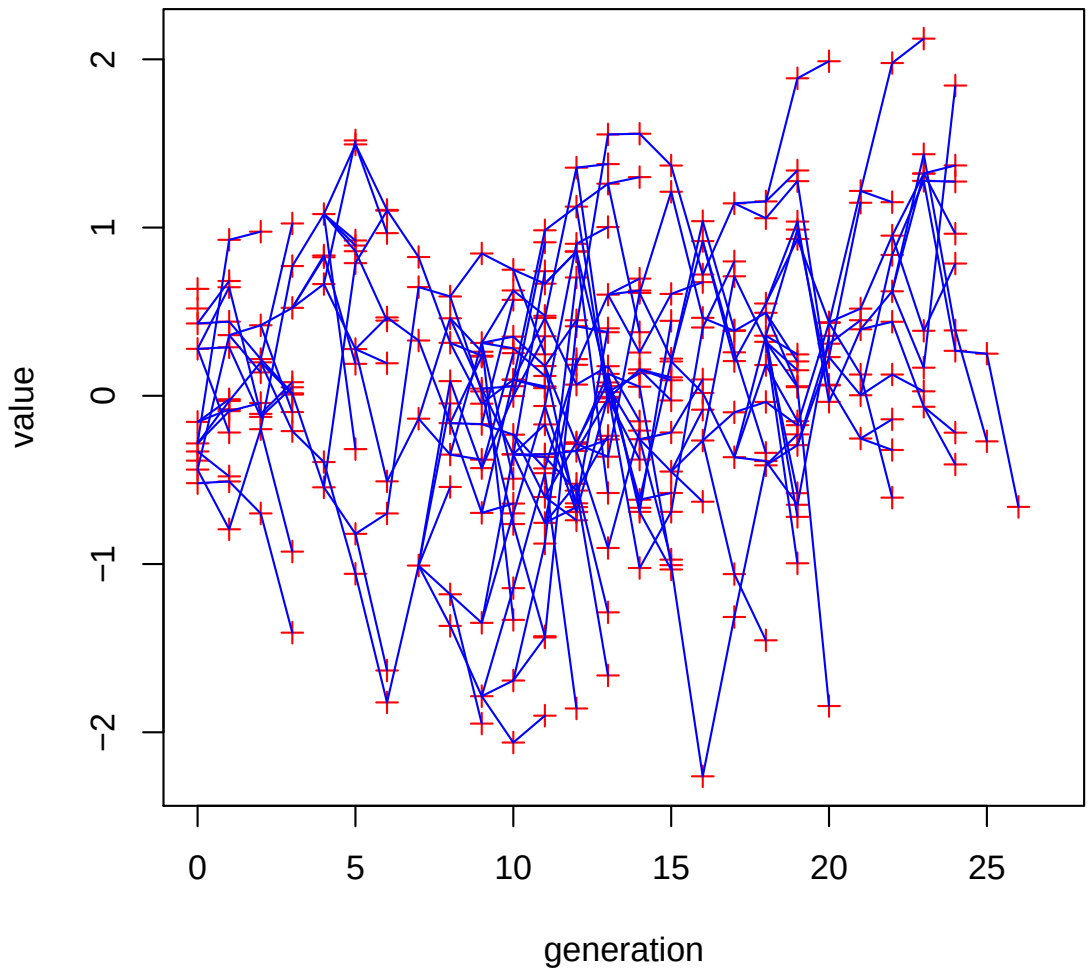
The threshold theorem seems to tell us that a branching process can never be stable – the population size goes to 0 or ∞.

(We stated this result for the single-type process, but it holds in the multi-type case, in the absence of degeneracy).

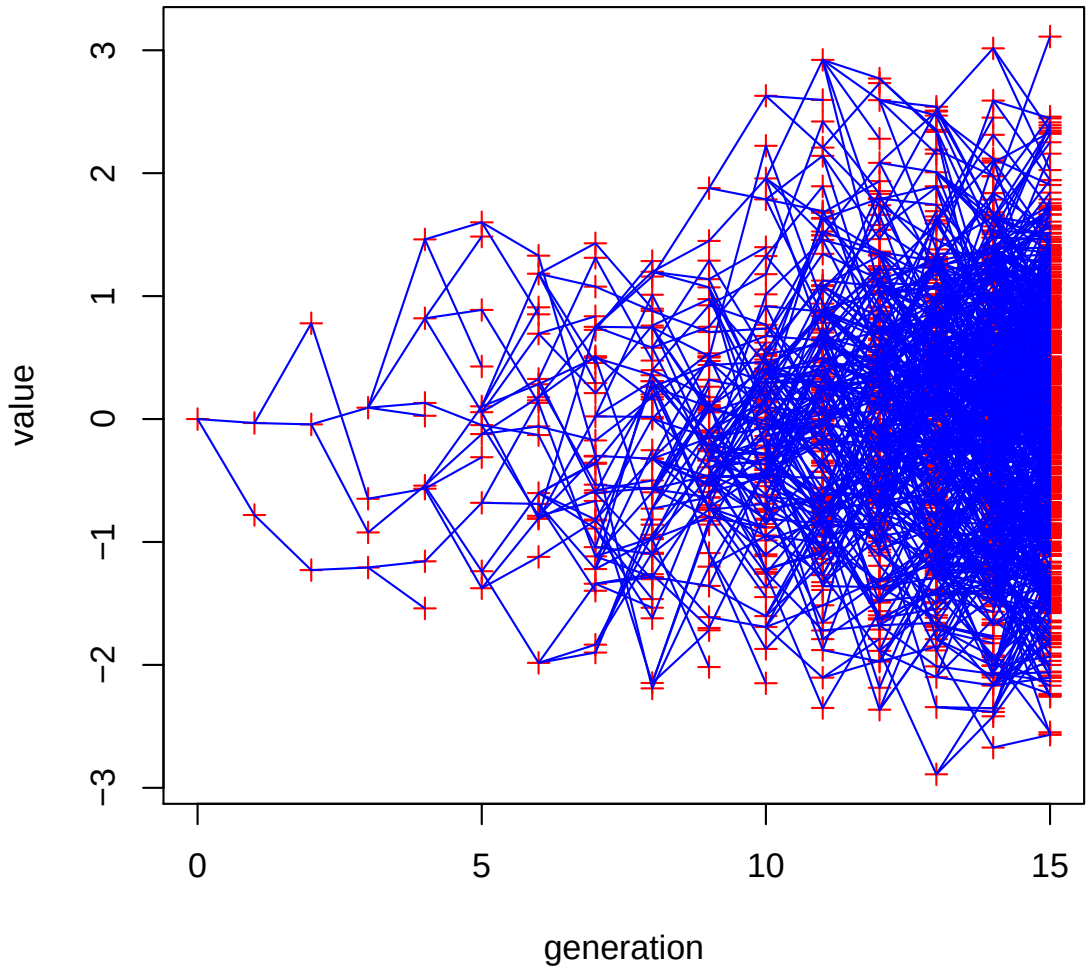
A subcritical BP



A critical BP



A supercritical BP



So we have some unattractive alternatives:

- a subcritical or critical process dies out, generating an a.s. finite total number of individuals: if the average $f(x)$ isn't accurate enough, all you can do is independently re-start.
- a supercritical process, if it doesn't die out, grows exponentially, so you are putting a lot of computing effort into 'genealogies' that are short, relatively – convergence?
- introducing immigration stabilises a (sub)critical process, but means constantly refreshing population with x values out of equilibrium

Some of these may be worth exploring, but our preference is construct a degenerate process that circumvents the problem.

Multi-category branching processes

For convenience and practicality in formulating such processes, it is best to extend the 'type' of each individual to encode more than just the 'value' x. We will illustrate this by adding a discrete 'category' k, so that the type is now (k, x).

We can then easily devise useful branching process MC methods where k (but not x) is used to regulate the reproduction process, while x values remain distributed asymptotically as $\pi(dx)$.

It is convenient to write

$$Z_n(\{k\}\times B) = Z_n^k(B) \quad \text{and} \quad M((k,x),\{k'\}\times B) = M^{kk'}(x,B),$$

etc., thus regarding Z_n as a (dependent) collection (indexed by k) of point processes on $\mathcal{X}$, rather than a single point process on a richer space.

The crown process (a.k.a. 'kings and bastards')

Individuals are of two categories (0=king, 1=bastard):

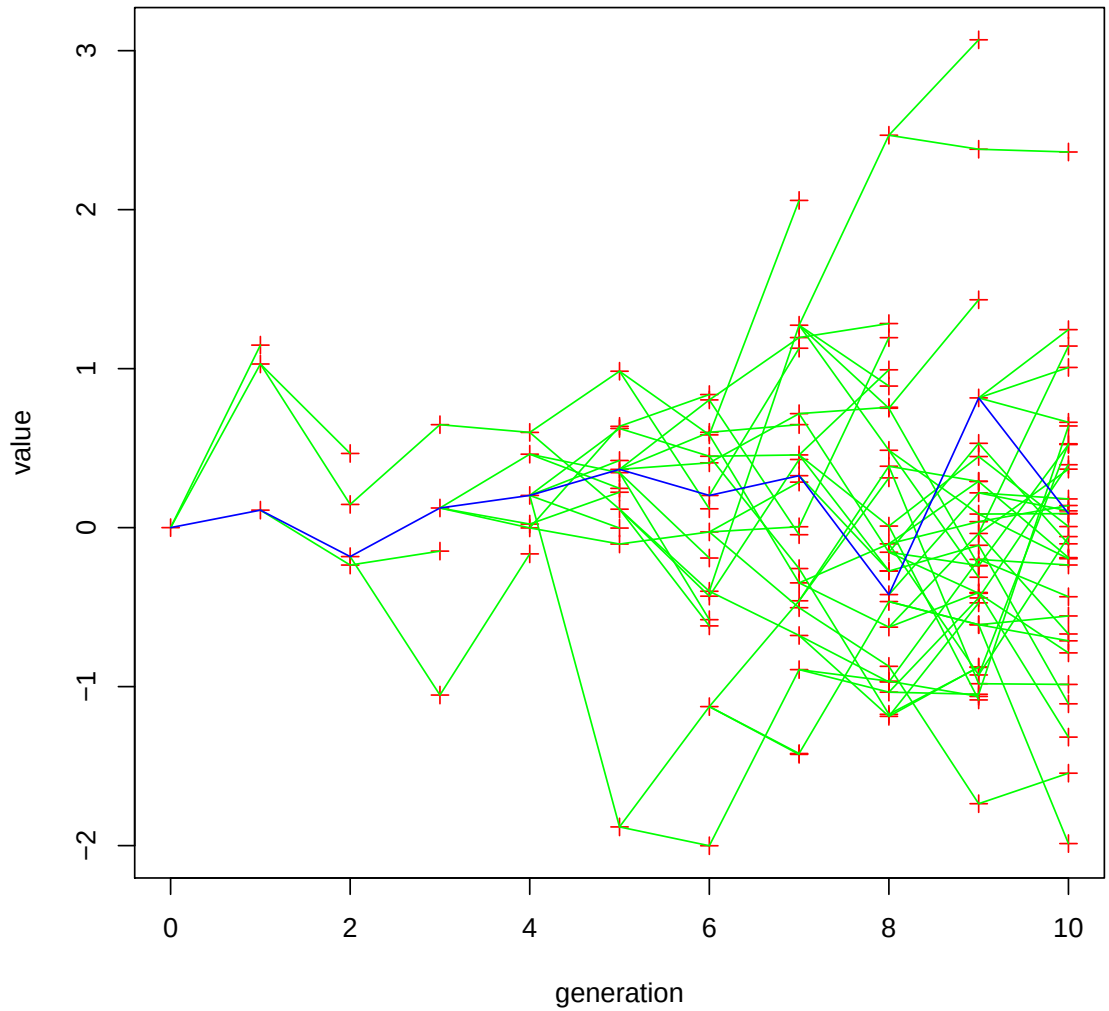
- a king has one king-type child, and a random number of bastards
- each bastard has no king-type children, and a (mean < 1) number of bastards

For the moment, we just sum $f(x)$ over individuals of both categories.

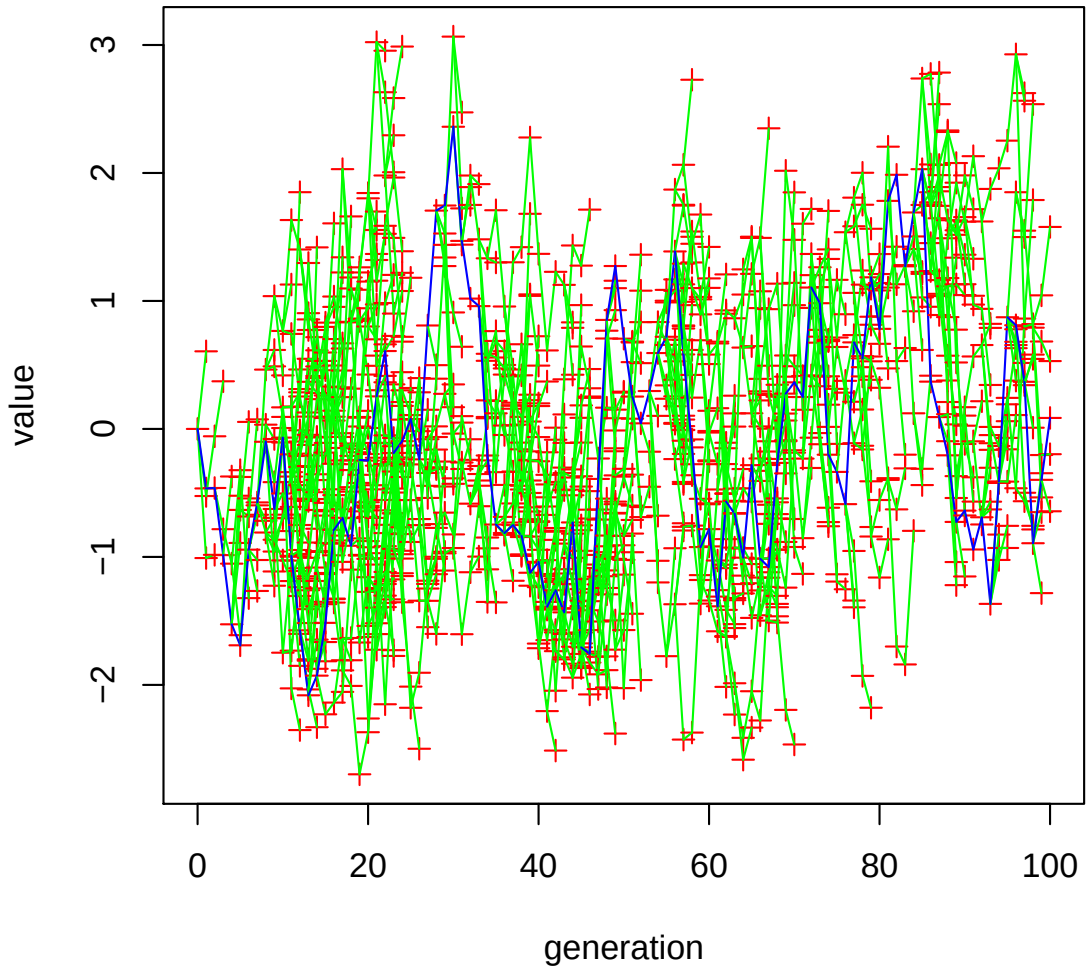
M^{00} is an ergodic probability transition kernel, M^{01} arbitrary, $M^{10} \equiv 0$, M^{11} is subcritical. This makes Z_n ergodic.

If M^{01} and M^{11} preserve π (the invariant distribution of M^{00}) up to proportionality, we have unbiased sampling.

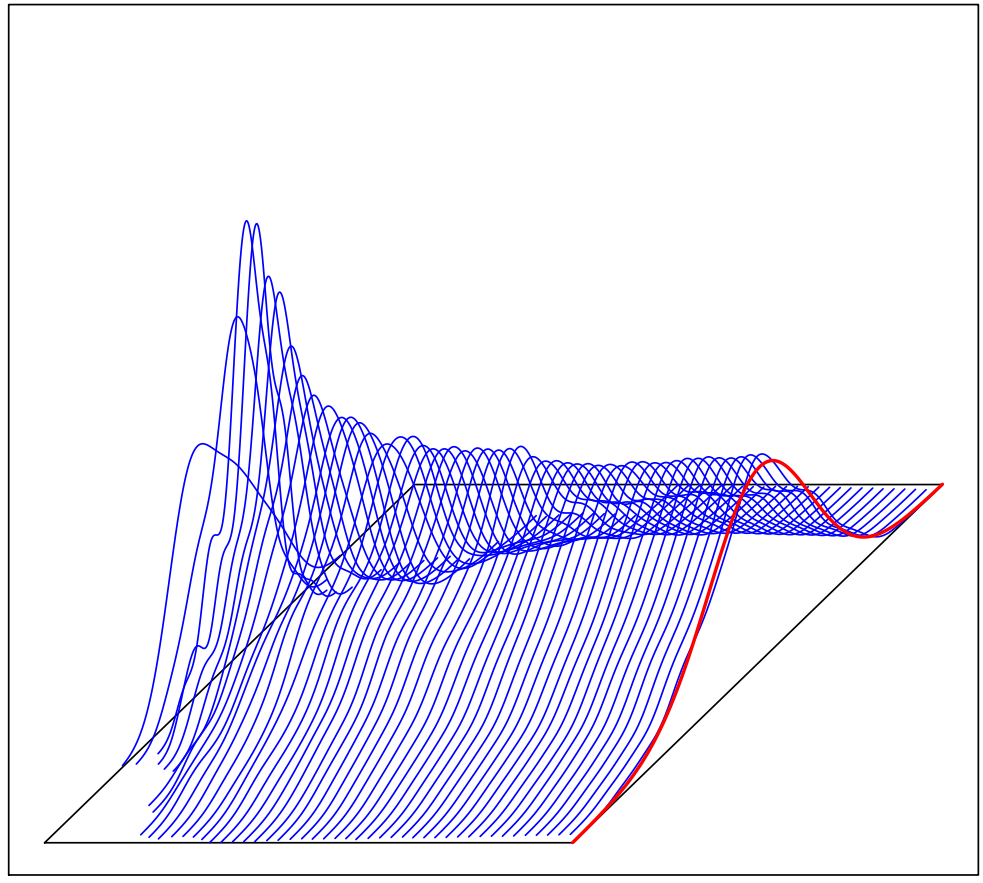
A realisation of a crown process



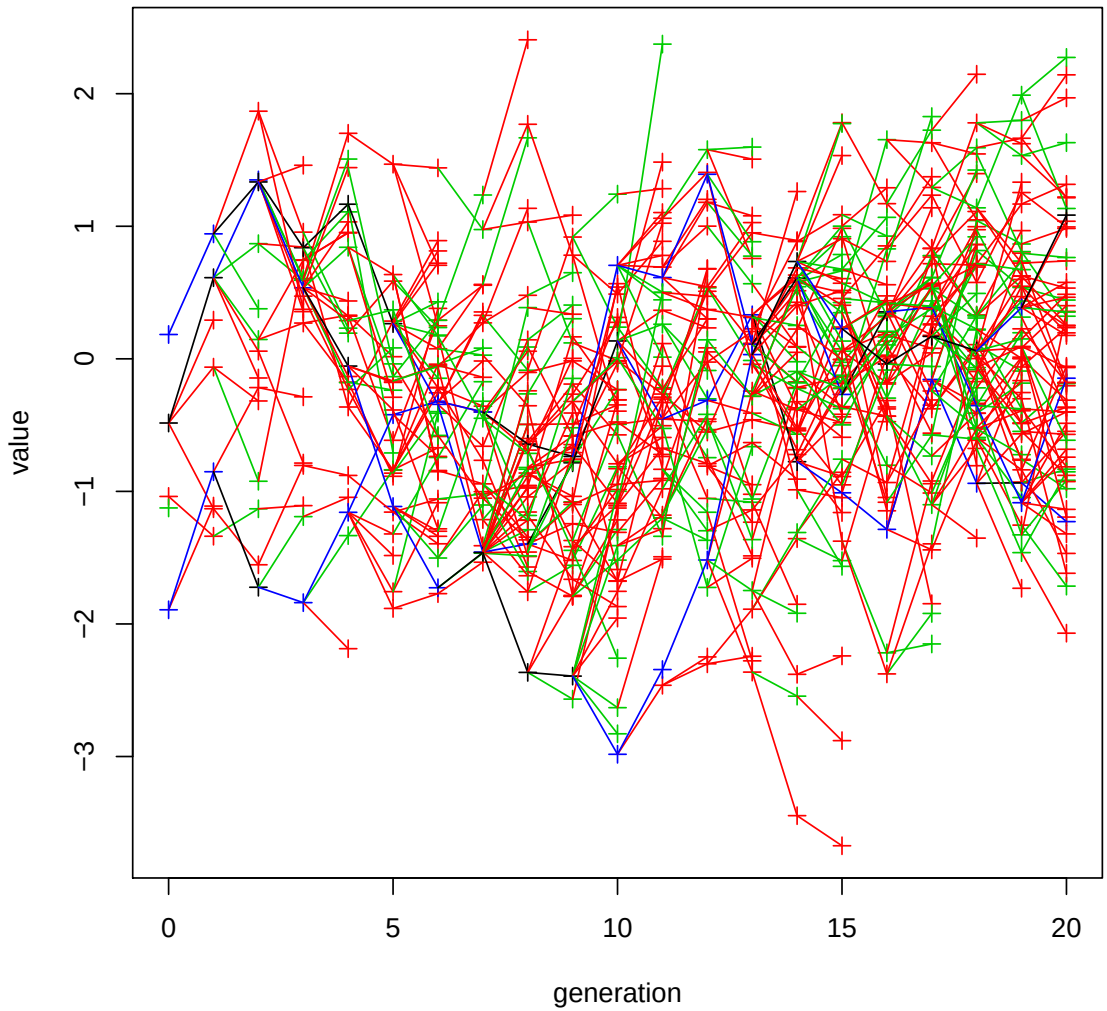
A longer run



Convergence of cumulative population



A more complicated variation



Ergodic averages

Our BPMC estimator of the expectation of f is S_n/N_n, where S_n is the total of $f(x)$ for all individuals in generations $0, 1, \ldots, n-1$:

$$S_n = \sum_{j=0}^{n-1} \int f dZ_j$$

and N_n is the number of individuals involved, the same expression with f replaced by $\mathbf{1}$, the function that is identically 1.

We can adapt the earlier argument to obtain a recursive expression for the joint MGF of (S_n, N_n):

$$E(\exp(-t_1 S_n - t_2 N_n)|Z_0 = x) = \Psi_n^{(x)}(s) \quad \text{where} \quad s(\cdot) = t_1 f(\cdot) + t_2$$

and

$$\Psi_{n+1}^{(x)}(s) = e^{-s(x)} \Phi^{(x)}(-\log \Psi_n^{(\cdot)}(s))$$

Now under stationarity,

$$E(S_n) = \sum_{j=0}^{n-1} \int f E(dZ_j) = \sum_{j=0}^{n-1} \int f d\mu = n \int f d\mu = n\mu' f$$

say, and $E(N_n) = n\mu'\mathbf{1}$.

Under appropriate conditions, we will have a law of large numbers, and

$$\frac{S_n}{N_n} \overset{\text{a.s.}}{\to} \frac{\mu' f}{\mu'\mathbf{1}} = E_\pi(f)$$

where $\pi = \mu/\mu'\mathbf{1}$ is μ normalised to be a probability measure.

Variances

The second-order structure of the process (again, integrating over the branching, not the value space) is controlled by the *covariance kernel* $G(x, B, C) = \mathsf{cov}[Z_1(B), Z_1(C)|Z_0 = x]$.

Suppose $\gamma_n(B, C) = \mathsf{cov}[Z_n(B), Z_n(C)]$, then

$$\gamma_{n+1}(B, C) = E[\mathsf{cov}(Z_{n+1}(B), Z_{n+1}(C)|Z_n)]$$

$$+\mathsf{cov}[E(Z_{n+1}(B)|Z_n), E(Z_{n+1}(C)|Z_n)]$$

$$= \int G(z, B, C)\mu_n(dz) + \int\int M(z, B)M(z', C)\gamma_n(dz, dz')$$

$$= ((\mu_n \odot G) + M'\gamma_n M)(B, C), \text{say.}$$

Under stationarity, in brief

$$\gamma = (\mu \odot G) + M'\gamma M$$

Asymptotic variance 1

For the variance of S_n/N_n, we need variances and covariances of S_n and N_n.

To find the variance of S_n, we need the equilibrium autocovariances of the sequence $(\int f dZ_n)$. We have

$$\text{cov}(\int f dZ_0, \int f dZ_n) = f'\gamma M^n f$$

for $n \geq 0$.

We are writing

$$f'\gamma M^n f = \int\int\int f(x_1)\gamma(dx_1, dx')M^n(x', dx_2)f(x_2)$$

where

$$M^n(x, B) = \int M(x, dx')M^{n-1}(x', B).$$

Asymptotic variance 2

Thus

$$
\begin{aligned}
n^{-1}\mathsf{var}(S_n) &\to \sum_{j=-\infty}^{\infty} \mathsf{cov}\left(\int f dZ_0, \int f dZ_j\right) \\
&= \lim_{n\to\infty} \sum_{j=-n+1}^{n-1} \mathsf{cov}\left(\int f dZ_0, \int f dZ_j\right) \\
&= \lim_{n\to\infty} f' \left(\sum_{j=1}^{n-1} M'^j \gamma + \gamma + \gamma \sum_{j=1}^{n-1} M^j\right) f \\
&= \lim_{n\to\infty} f' C_n f, \quad \text{say.}
\end{aligned}
$$

Asymptotic variance 3

But $M^n \to M^\infty$, where $MM^\infty = M^\infty M = M^\infty$ and $\gamma M^\infty = 0$, whence we find

$$(I - M + M^\infty)' C_n (I - M + M^\infty) = \gamma - M'\gamma M = \mu \odot G$$

Thus

$$C_n \to C = ((I - M + M^\infty)')^{-1} (\mu \odot G)(I - M + M^\infty)^{-1}$$

and

$$\text{var}(S_n) \sim n f' C f.$$

By the same argument,

$$\text{cov}(S_n, N_n) \sim n f' C \mathbf{1} \qquad \text{and} \qquad \text{var}(N_n) \sim n \mathbf{1}' C \mathbf{1}.$$

Asymptotic variance 4

By the delta method, the variance of our branching Monte Carlo estimator S_n/N_n is

$$\mathrm{var}\left(\frac{S_n}{N_n}\right) = \frac{v^*(f)}{E(N_n)} + O(E(N_n)^{-2})$$

as $n \to \infty$, where the variance factor $v^*(f)$ characterising the performance of the estimator is

$$\lim_{n\to\infty} \frac{1}{E(N_n)} \left\{ \mathrm{var}(S_n) - 2\frac{E(S_n)}{E(N_n)}\mathrm{cov}(S_n, N_n) + \frac{E(S_n)^2}{E(N_n)}\mathrm{var}(N_n) \right\}$$

$$= \frac{1}{\mu'\mathbf{1}} \left\{ f'Cf - 2\left(\frac{\mu' f}{\mu'\mathbf{1}}\right) f'C\mathbf{1} + \left(\frac{\mu' f}{\mu'\mathbf{1}}\right)^2 \mathbf{1}'C\mathbf{1} \right\}.$$

Asymptotic variance 5

One use for this is to try to understand how to design the process to reduce variance. The key is

$$C = ((I - M + M^{\infty})')^{-1}(\mu \odot G)(I - M + M^{\infty})^{-1}$$

where

$$(\mu \odot G)(B, C) = \int G(z, B, C)\mu(dz)$$

Loosely, negatively associated offspring reduces $f'Cf$.

Unhelpful toy example

Univariate x, target $\mathrm{N}(0,1)$, using "AR(1)" dynamics:
$x_{\mathrm{child}} = \rho x_{\mathrm{parent}} + \sqrt{1-\rho^2}\mathrm{N}(0,1)$, with $\rho = 0.8$.

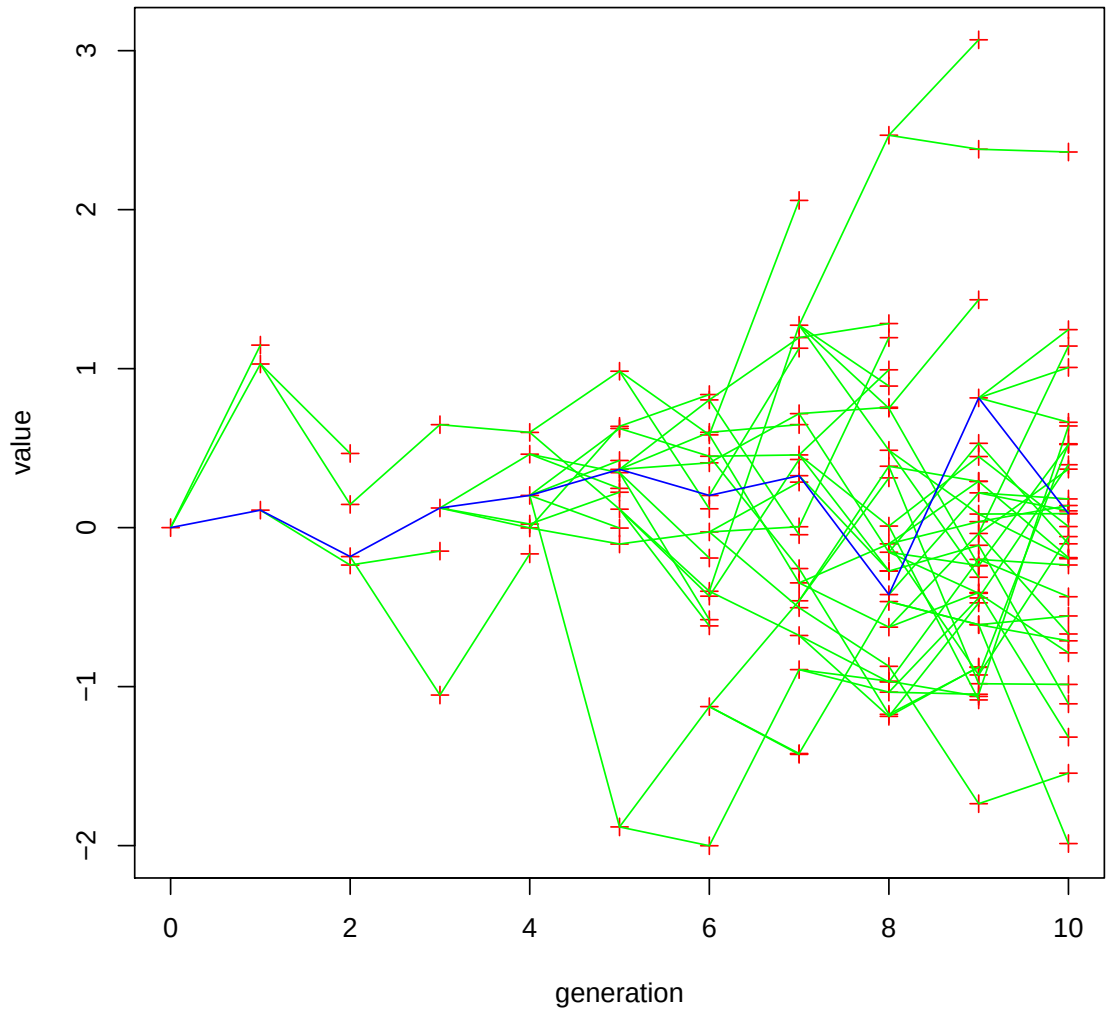
Crown process, with offspring bastard distribution $(0.0, 0.0, 1.0)$ for kings and $(0.51, 0.0, 0.49)$ for bastards, yields asymptotic variance $\approx 29.7/N$. ($N =$ total number generated; variance estimated from 1000 independent replicates.)

Antithetic version, using maximally antithetic normal increments, yields $\approx 20.6/N$ (30% better).

However, ordinary MCMC asymptotic variance is $N^{-1}(1+\rho)/(1-\rho) = 9/N$ with $\rho = 0.8$.

Can potential advantages of branching overcome the disadvantage that pairs of individuals are on average closer together, so more correlated?

A realisation of a crown process



Unimpressive toy example

For finite state space chains, we can use the asymptotic variance results (or indeed compute variances for finite runs if desired).

It is difficult to discern a general pattern from our experiments, but our examples include cases (with the same offspring distributions as above) where

- branching without antithetics is 21% worse than regular MCMC and two kinds of antithetic modification are 10% or 52% better than regular MCMC
- branching without antithetics is 1% worse than regular MCMC and two kinds of antithetic modification are 33% or 97% better than regular MCMC

Ergodic weighted averages

Consider the weighted BPMC estimator$S_n^{\mathrm{w}}/N_n^{\mathrm{w}}$, where S_n^{w} is the total of $w(x)f(x)$ for all individuals in generations $0, 1, \ldots, n-1$:

$$S_n^{\mathrm{w}} = \sum_{j=0}^{n-1} \int wf dZ_j$$

and N_n^{w} is the total of the $w(x)$ involved, the same expression with f replaced by $\mathbf{1}$. Now under stationarity,

$$E(S_n^{\mathrm{w}}) = \sum_{j=0}^{n-1} \int wf E(dZ_j) = \sum_{j=0}^{n-1} \int wf d\mu = n \int wf d\mu = n\mu'(wf)$$

say, and $E(N_n^{\mathrm{w}}) = n\mu' w$.

Under appropriate conditions, we will have a law of large numbers, and

$$\frac{S_n^{\mathrm{w}}}{N_n^{\mathrm{w}}} \overset{\text{a.s.}}{\to} \frac{\mu'(wf)}{\mu' w} = E_\pi(f)$$

where π is $w\mu$ normalised to be a probability measure, i.e.

$$\pi(dx) \propto w(x)\mu(dx).$$

Category-specific invariant measures

Suppose we have multiple categories, indexed by $k \in \mathcal{K}$. Recall the notation

$$Z_n(\{k\}\times B) = Z_n^k(B) \quad \text{and} \quad M((k,x),\{k'\}\times B) = M^{kk'}(x,B).$$

Suppose the process is ergodic, then $E(Z_n^k(B)) = \mu^k(B)$ does not depend on n, but may be different for different k; the global balance (invariance) equations are

$$\mu^k = \sum_{j\in\mathcal{K}} \mu^j M^{jk}$$

If we use category-specific weights $w^k(x)$ in accumulating $f(x)$, then

$$\frac{S_n^{\mathrm{w}}}{N_n^{\mathrm{w}}} \overset{\text{a.s.}}{\to} \frac{\mu'(wf)}{\mu' w} = E_\pi(f)$$

where

$$\pi(dx) \propto \sum_{k \in \mathcal{K}} w^k(x) \mu^k(dx).$$

So we can have different invariant measures in different categories, and either weight out ($w^k(x) = 0$) those where $\mu^k \not\propto \pi$, or use weights to adjust appropriately. You could have one or more categories where μ^k is deliberately over-dispersed (tempered) to assist mixing, but which are not used in ergodic averaging.

Fractional individuals

Freeing Z_n from being integer-valued (but remaining discrete) allows us to represent *weighted* individuals $\{(x_i, \theta_i)\}$,

$$Z_n(A) = \sum_{\text{individuals in generation } n} \theta_i I[x_i \in A]$$

Individuals can 'decide for themselves' whether to

- pass on their weight θ to their offspring
- resample/replicate: generate random number (with mean θ, arbitrary distribution) of offspring and give them weight 1
- (etc)

. . . still have $\mu_{n+1} = \mu_n M$ where $\mu_n(B) = E(Z_n(B))$.

Blessing or curse?

In spite of these flexibilities, we have not found a convincing demonstration of effectiveness (yet?).

Is this the inevitable downside of the very lack of interaction between individuals, once created, that is key to the analysis?

Answers to:

P.J.Green@bristol.ac.uk

Antonietta.Mira@uninsubria.it