Coupling from the Past

This note provides a guided tour through two proofs of the coupling from the past algorithm (J. G. Propp and Wilson 1996; J. Propp and Wilson 1997) to perfectly sample from the stationary distribution of an irreducible Markov chain. It also explains why two related ideas (a forward scheme and sampling fresh random functions at each iteration) both fail at perfectly sampling from π.

Random mapping representation for $P(\cdot, \cdot)$

Let $\mathcal{X}$ be a finite set and let $P(\cdot, \cdot)$ be an irreducible and aperiodic transition matrix on $\mathcal{X}$, with stationary distribution π. Throughout, a random function $f : \mathcal{X} \to \mathcal{X}$ is given such that for any $x, y \in \mathcal{X}$ we have

$$\mathbf{P}(f(x) = y) = P(x, y).$$

Let $(f_t)_{t=0,1,2,\ldots}$ be a countable sequence of iid copies of f that the practitioner may use for sampling. For any $t \geq 0$ we thus have $\mathbf{P}(f_t(x) = y) = P(x, y)$ for all deterministic $x, y \in \mathcal{X}$.

The goal is the perfectly sample from the unique stationary distribution π of $P(\cdot, \cdot)$. It is perhaps natural to define random variables on $\mathcal{X}$ using that the events

$$\{f_\tau \circ f_{T-1} \circ \ldots \circ f_1 \text{ is constant } \} \text{ or } \{f_1 \circ f_2 \circ \ldots \circ f_\tau \text{ is constant } \},$$

where τ is a possibly random integer, as in both cases we can use the unique value of the corresponding constant function to canonically define a random variable in $\mathcal{X}$.

Coalescence and zero-one law

The first question is whether such random integer τ actually exists. If $\mathbf{P}(f_t \circ f_{t-1} \circ ... f_1) = 0$ for all deterministic $t \geq 1$ then no such random τ exists by sigma-additivity. On the other hand, if $q_t = \mathbf{P}(f_t \circ f_{t-1} \circ ... f_1) > 0$ for some $t \geq 1$, then

$$\begin{aligned}\mathbf{P}\Big(f_{kt} \circ f_{kt-1} \circ ... f_1 \text{ is not constant}\Big) &\leq \mathbf{P}\Big(\cap_{0=1}^{k} \Big\{ f_{it} \circ f_{it-1} \circ ... f_{(i-1)t+1} \text{ is not constant} \Big\}\Big) \\ &= (1-q_t)^k\end{aligned}$$

which converges to 0 as $k \to +\infty$, so that by the monotonge convergence theoren,

$$\tau = \min\{t \geq 1 : f_\tau \circ ... \circ f_1 \text{ is constant } \}$$

is finite with probability one. In summary $\mathbf{P}(\tau = +\infty) \in \{0, 1\}$, which is an example of a zero-one law.

Finite coalesence

There are random functions f such that $\mathbf{P}(\tau = +\infty) = 1$, even for irreducible and aperiodic chains, for instance by coupling the values of f to synchronously move on the cycle $\mathcal{X} = \{0, ..., n-1\}$ as in

$$\mathbf{P}(f(x) = x + \Delta \mod n) = 1/3, \qquad \text{for all } \Delta \in \{-1, 0, 1\}.$$

This gives a first negative answer: We cannot always assume that $\mathbf{P}(\tau < +\infty) = 1$.

However, for a given aperiodic and irreducible transition matrix $P(\cdot, \cdot)$ we can always construct a random mapping representation f such that $\mathbf{P}(\tau < +\infty) = 1$ by choosing f such that $(f(x))_{x \in \mathcal{X}}$ are mutually independent. Since there exists an integer $k \geq 1$ such that $P^k(x, y) > 0$ for all $x, y \in \mathcal{X}$, taking any fixed y_0 we find that

$$\mathbf{P}(\tau \leq k) \geq \mathbf{P}(\cap_{x \in \mathcal{X}} \{y_0 = f_k \circ ... \circ f_1(x)\}) = \prod_{x \in \mathcal{X}} P^k(x, y_0) > 0.$$

By the previous zero-one law, in this case $\mathbf{P}(\tau < +\infty) = 1$.

From now on, we assume that the random mapping representation is such that $\mathbf{P}(\tau < +\infty) = 1$.

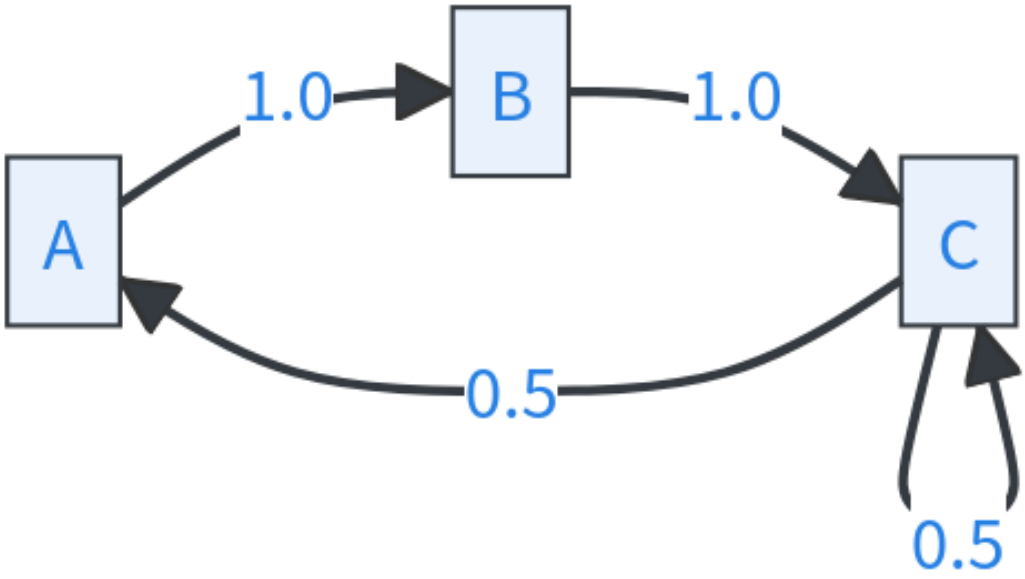
Forward

The first idea is to apply the random functions $f_1, f_2, ...$ forward, as one would naturally proceed to smaple a Markov chain. Start from an initialization $x_0 \in \mathcal{X}$, apply f_1 to obtain $f_1(x_0)$ as the first state of the Markov chain, apply f_2 to obtain $f_2 \circ f_1(x_0)$ as the second state, apply f_3 to obtain $f_3 \circ f_2 \circ f_1(x_0)$, etc. Consider the random variable

$$f_\tau \circ f_{\tau-1} \circ ... \circ f_1(x_0),$$

that is, the value of the first constant function of the form $f_t \circ f_{t-1} \circ ... \circ f_1$. With the following example,

```
graph LR
    A -->|1.0| B
    B -->|1.0| C
    C -->|0.5| C
    C -->|0.5| A
```



any random mapping representation is such that $\mathbf{P}(f_\tau \circ ... \circ f_1(x_0) = C) = 1$ so this random variable cannot be distributed according to the stationary distribution $\pi > 0$ which has positive mass on each of $\{A, B, C\}$.

Backward - Coupling from the Past

The key idea in coupling from the past (J. G. Propp and Wilson 1996) is to consider a "backward" scheme instead, where the random functions are appplied backward instead of the previous forward scheme. Define the random integer $F \geq 1$ as

$$F = \min\{t \geq 1 : f_1 \circ f_2 \circ ... \circ f_t \text{ is constant } \}$$

and denote by $\hat{X}$ the value of the corresponding constant function, that is,

$$\hat{X} = f_1 \circ f_2 \circ ... \circ f_F(x_0)$$

for a choice of some deterministic $x_0 \in \mathcal{X}$ that does not matter.

Now consider a sequence $(g_t)_{t\geq 1}$ of iid functions with the same distribution as f and the $(f_t)_t$. Define, in exactly the same way,

$$G = \min\{t \geq 1 : g_1 \circ g_2, \circ ... \circ g_t \text{ is constant } \}, \quad \hat{Y} = g_1 \circ g_2 \circ ... \circ g_G(x_0).$$

By construction $(G, \hat{Y})$ is equal in distribution to $(F, \hat{X})$ as the first uses $(g_t)_{t\geq 1}$, the second uses $(f_t)_{t\geq 1}$, and both sequences are assumed iid with the same distribution as f.

Now define the $(g_t)_{t\geq 1}$ by $g_t = f_{t-1}$, so that $(g_t)_{t\geq 1}$ is indeed a sequence of iid functions with the same distribution as f. Then

$$\begin{aligned} f_0(\hat{X}) &= f_0 \circ f_1 \circ ... \circ f_F(x_0) \\ &= g_1 \circ \underbrace{g_2 \circ ... \circ g_{F+1}}_{\text{constant}}(x_0). \end{aligned}$$

Since $f_1 \circ ... ... \circ f_F = g_2 \circ ... \circ g_{F+1}$ is constant by definition of F, necessarily $g_1 \circ g_2 \circ ... \circ g_{F+1}$ is constant too because composing a constant function with another produces a constant function. This implies $G \leq F + 1$ and $f_0(\hat{X}) = \hat{Y}$ with probability one. Now f_0 is independent of $\hat{X}$ since $(F, \hat{X})$ is defined as a function of $(f_t)_{t\geq 1}$ only, excluding f_0. That $\hat{X}$ is independent of f_0 and the equality in distribution $f_0(\hat{X}) =^d \hat{X}$ proves that $\hat{X}$ is distributed according to π. This argument has likely appeared before, though I am yet to find a reference.

Backward (another proof of perfect sampling)

Another simple proof observes that for any $\epsilon > 0$, there exists $t \geq 1$ such that

$$\mathbf{P}(f_1 \circ ... \circ f_t \text{ is constant }) \geq 1 - \epsilon$$

by the assumption $\mathbf{P}(\tau < +\infty) = 1$. If X_π is distributed according to π independently of $(f_t)_{t\geq 1}$, then $f_1 \circ f_2 \circ ... \circ f_t(X_\pi)$ also has distribution π since π is stationary. If ν is the

distribution of $\hat{X}$, by definition of $\hat{X}$ we have $\mathbf{P}(\hat{X} = X_\pi) \geq 1 - \epsilon$. Since the total variation distance is the infimum of $\mathbf{P}(X \neq Y)$ over all couplings (X, Y) of ν and π,

$$\|\nu - \pi\|_{TV} \leq \mathbf{P}(\hat{X} \neq X_\pi) \leq \epsilon.$$

This holds for all $\epsilon > 0$ hence $\hat{X} \sim \pi$. This argument can be found in (Häggström 2002, Theorem 10.1).

Part III (sampling t new, fresh random functions)

Consider now the following algorithm:

Algorithm 2:

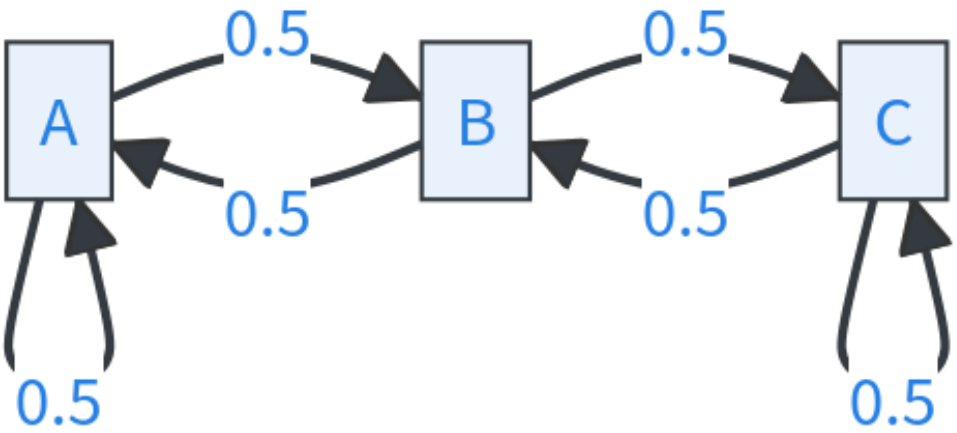
a. Set $t = 1$.
b. Generate $f_1^{(t)}, f_2^{(t)}, f_3^{(t)}, ..., f_t^{(t)}$ iid copies of the random function f independently of all previous iterations of the algorithm
c.
- If $f_1^{(t)} \circ ... \circ f_t^{(t)}$ is a constant function, then output its unique value $\hat{Z}$ and stop the algorithm.
- Otherwise, throw away $f_1^{(t)}, f_2^{(t)}, f_3^{(t)}, ..., f_t^{(t)}$, increase t by one, i.e., set $t := t + 1$ and go to step b.

It was known early on with the works J. G. Propp and Wilson (1996); J. Propp and Wilson (1997) that this, perhaps natural, idea fails to perfectly sample from π because of the following example. This means that reusing the same randomness in coupling from the past is key to perfectly sample from π.

To see that Algorithm 2 fails, let $\mathcal{X} = \{A, B, C\}$ with the random function f defined by

$$\mathbf{P}\Big(f(A) = B, f(B) = C, f(C) = C\Big) = 1/2, \quad \mathbf{P}\Big(f(A) = A, f(B) = A, f(C) = B\Big) = 1/2.$$

```
graph LR
    A -->|0.5| A
    A -->|0.5| B
    B -->|0.5| C
    B -->|0.5| A
    C -->|0.5| B
    C -->|0.5| C
```



This Algorithm 2 fails to sample from π because, if T denotes the random time at which the algorithm terminates,

$$\mathbf{P}\Big(T = 2, \hat{Z} \in \{A, C\}\Big) + \mathbf{P}\Big(T = 3, \hat{Z} \in \{A, C\}\Big) > \pi(A) + \pi(C).$$

That is, only looking at the events that the algorithm terminates within three iterationms, the algorithm has already oversampled from $\{A, C\}$.

References

Häggström, Olle. 2002. *Finite Markov Chains and Algorithmic Applications.* Vol. 52. Cambridge University Press.

Propp, James Gary, and David Bruce Wilson. 1996. "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics." *Random Structures & Algorithms* 9 (1-2): 223–52.

Propp, James, and David Wilson. 1997. "Coupling from the Past: A User's Guide." *Microsurveys in Discrete Probability* 41: 181–92.