Coupling from the Past

Yanheng Wang | 31 July

Monte Carlo v.s. Las Vegas

One runs the algorithm for a *bounded* time, and there is a small chance of error after the run.

MCMC

$\|\mu P^t - \pi\| < \epsilon$ when halting at $t = O(p(n, 1/\epsilon))$.

Monte Carlo v.s. Las Vegas

One runs the algorithm for a *bounded* time, and there is a small chance of error after the run.

One runs the algorithm *indefinitely*, until he gets the correct answer. The expected time is bounded, however.

MCMC

$\|\mu P^t - \pi\| < \epsilon$ when halting at $t = O(p(n, 1/\epsilon))$.

CFTP

$\|\mu P^T - \pi\| = 0$ when halting; $\mathbb{E}[T] = O(p(n))$.

A Different Description of MC

It's convenient to think of Markov chains from a different point of view.

Definition. Suppose S is a random variable over the space $\mathcal{S}$. Fix a function $f : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$. We say that S together with f *induce* a transition matrix $P(x,y) := \Pr[f(x,S) = y]$ where $x, y \in \mathcal{X}$.

A Different Description of MC

It's convenient to think of Markov chains from a different point of view.

Definition. Suppose S is a random variable over the space $\mathcal{S}$. Fix a function $f : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$. We say that S together with f *induce* a transition matrix $P(x, y) := \Pr[f(x, S) = y]$ where $x, y \in \mathcal{X}$.

Theorem. Suppose S and f induce P. If $\{S_t\}$ is an i.i.d. sequence in the same distribution as S, then the sequence generated by $X_t := f(X_{t-1}, S_t)$ forms a Markov chain with transition matrix P.

Definition. Suppose S is a random variable over the space $\mathcal{S}$. Fix a function $f : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$. We say that S together with f *induce* a transition matrix $P(x, y) := \Pr[f(x, S) = y]$ where $x, y \in \mathcal{X}$.

Theorem. Suppose S and f induce P. If $\{S_t\}$ is an i.i.d. sequence in the same distribution as S, then the sequence generated by $X_t := f(X_{t-1}, S_t)$ forms a Markov chain with transition matrix P.

e.g. Metropolis chain of q-colouring can be generated by

- $\mathcal{X} := \{$All colourings on graph $G\}$
- $\mathcal{S} := \{1, 2, \ldots, n\} \times \{1, 2, \ldots, q\}$
- S is uniformly distributed on $\mathcal{S}$
- $f(x, s) :=$ "Decode the tuple $s = (i, c)$; Colour the i-th position of x as c and return the new colouring."

An Introductory Scenario

Suppose you are told (by an idiot) to generate a random binary string $\beta \in \{0,1\}^n$ *via a Markov chain*.

Here's a natural construction. In each step,

1. Select a random position $1 \leq i \leq n$;
2. Flip a fair coin. If it lands heads up, set the i-th bit 0; otherwise, set the i-th bit 1.

Problem. Define $\mathcal{S}$, S, and $f(x,s)$.

An Introductory Scenario

Suppose you are told (by an idiot) to generate a random binary string $\beta \in \{0,1\}^n$ *via a Markov chain*.

Here's a natural construction. In each step,

1. Select a random position $1 \leq i \leq n$;
2. Flip a fair coin. If it lands heads up, set the i-th bit 0; otherwise, set the i-th bit 1.

Problem. Define $\mathcal{S}$, S, and $f(x, s)$.

What's the stationary distribution π of this chain?

1. Select a random position $1 \le i \le n$;
2. Flip a fair coin. If it lands heads up, set the i-th bit 0; otherwise, set the i-th bit 1.

Now denote $\{X_t^x\}$ as the Markov chain started from the initial binary string $x \in \{0,1\}^n$.

We couple all these 2^n chains together by "sharing the random source $\{S_t\}$":

1. Select a random position $1 \le i \le n$;
2. Flip a fair coin. If it lands heads up, set the i-th bits in *all* chains 0; otherwise, set the i-th bits in *all* chains 1.

1. Select a random position $1 \leq i \leq n$;
2. Flip a fair coin. If it lands heads up, set the i-th bits in *all* chains 0; otherwise, set the i-th bits in *all* chains 1.

e.g. $\quad n = 2$

$X_t^{00} = 01 \qquad X_t^{01} = 01 \qquad X_t^{10} = 00 \qquad X_t^{11} = 11$

We select randomly $i = 2$, and the coin lands heads up.

$X_{t+1}^{00} = 00 \qquad X_{t+1}^{01} = 00 \qquad X_{t+1}^{10} = 00 \qquad X_{t+1}^{11} = 10$

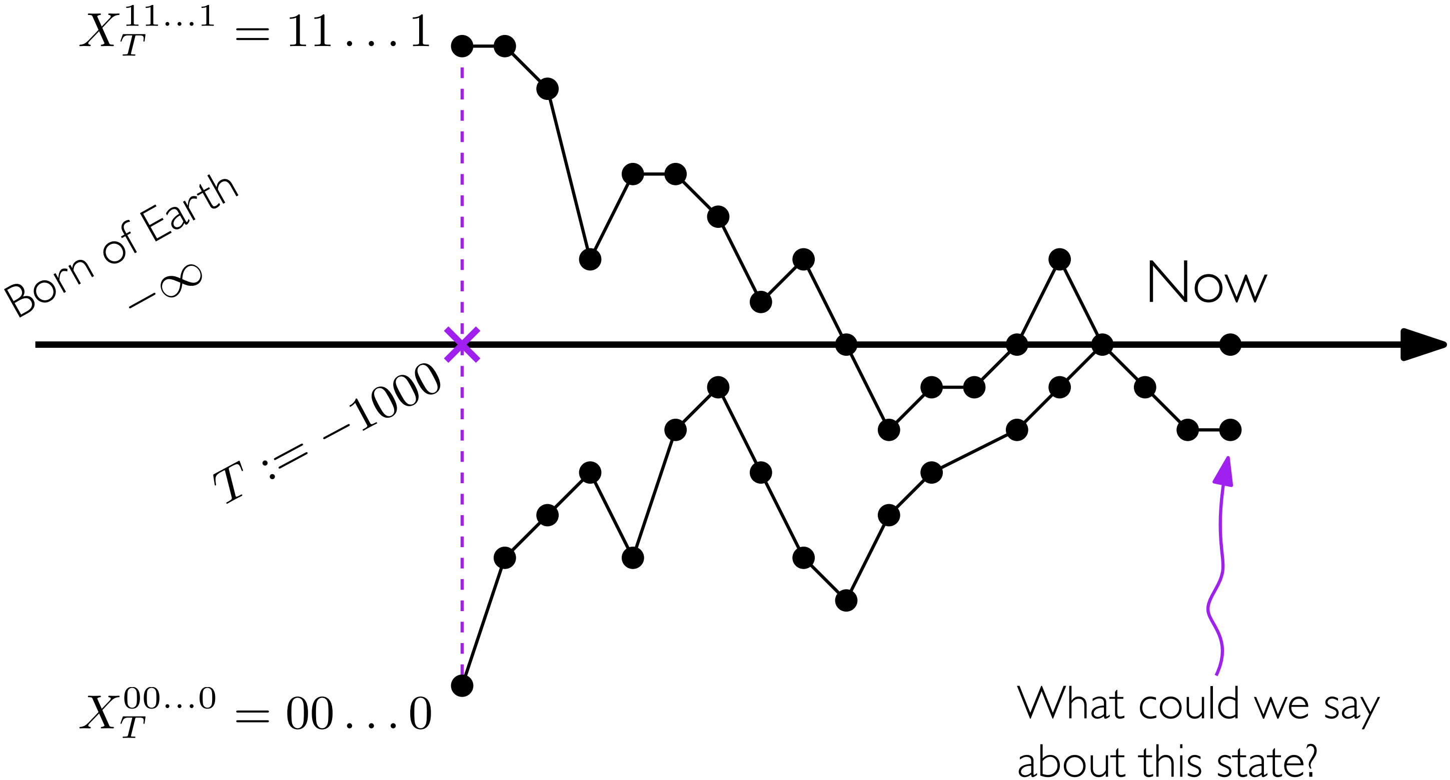
1. Select a random position $1 \leq i \leq n$;
2. Flip a fair coin. If it lands heads up, set the i-th bits in *all* chains 0; otherwise, set the i-th bits in *all* chains 1.

Definition. We define a partial order $\leq$ on the space $\{0,1\}^n$ as follows:
$b_1 b_2 \ldots b_n \leq b'_1 b'_2 \ldots b'_n \iff b_i \leq b'_i$ for all i.

e.g. $0011010 \leq 0111011$

Observation. If $X_t^x \leq X_t^y$, then $X_{t+1}^x \leq X_{t+1}^y$.
Observation. $X_t^x \leq X_t^{11\ldots 1}$, for all x and t. Similarly, $X_t^{00\ldots 0} \leq X_t^x$ for all x and t.

Observation. $X_t^x \le X_t^{11\ldots1}$, for all x and t. Similarly, $X_t^{00\ldots0} \le X_t^x$ for all x and t.

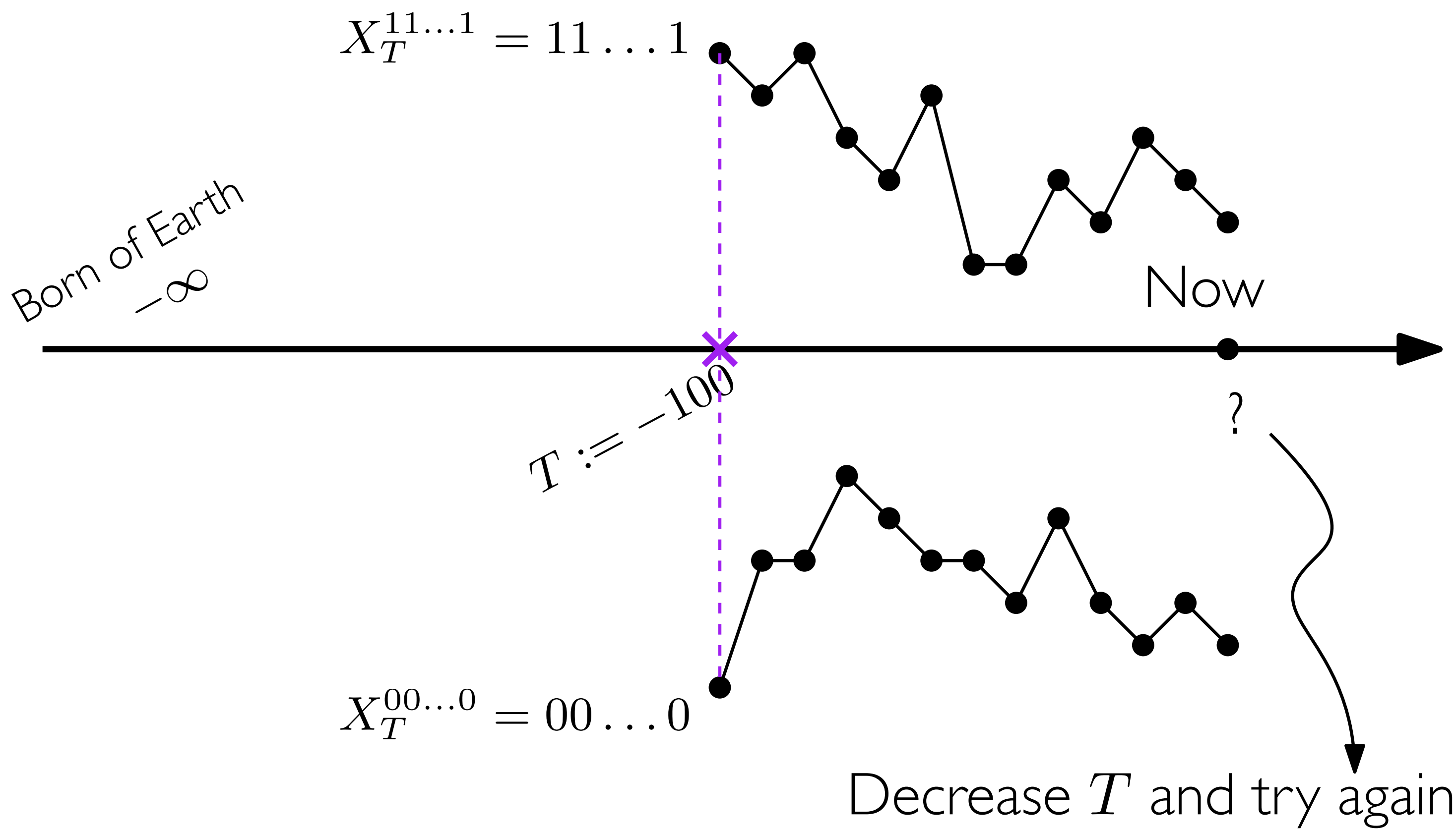


Claim. Started at time $T < 0$, if the simulation for $\{X_t^{11\ldots1}\}$ and $\{X_t^{00\ldots0}\}$ meets at state X at time 0, i.e. $X_0^{11\ldots1} = X_0^{00\ldots0} = X$, then $X \sim \pi$.

Remark. Strictly speaking, we must clarify why the simulation result X is random. We'll be back on a more formal version soon. My point here is to give you a taste on how the argument should proceed.

"**Proof.**"
By our observation, X_0^x is bounded between$X_0^{00\ldots0}$ and $X_0^{11\ldots1}$ for all x. Since the upper bound coincide with the lower bound, we must conclude that everything collapses to the single point X.
Now imagine a fictional chain started at the born of Earth. It runs long enough so it must have converged to stationary π now. But let us recall that when it enters the zone $[T, 0]$, it must be bounded as well. Therefore, its current state also equals X. So, $X \sim \pi$.

Observation. $X_t^x \leq X_t^{11\ldots1}$, for all x and t. Similarly, $X_t^{00\ldots0} \leq X_t^x$ for all x and t.



CFTP in General

We have a chain generated by S and $f : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$. In addition, we have a partial order $\leq$ defined on $\mathcal{X}$.

Theorem. If $\forall s \in \mathcal{S}, x_1 \leq x_2 \Rightarrow f(x_1, s) \leq f(x_2, s)$, then the following algorithm returns a random sample over $\mathcal{X}$ with stationary distribution of the chain.

$T := -1/2$
repeat
  $T := 2T$
  $X_T^\top := \top$; $X_T^\perp := \perp$
  for $t := T + 1 \ldots 0$ **do**
    Choose independently $S_t \sim S$
    $X_t^\top := f(X_{t-1}^\top, S_t)$
    $X_t^\perp := f(X_{t-1}^\perp, S_t)$
until $X_0^\top = X_0^\perp$;
return $X_0^\top$

Remark. In subsequent rounds, we do *not* pick fresh variables if they had been chosen in former rounds!

```
T := -1/2
repeat
    T := 2T
    X_T^⊤ := ⊤; X_T^⊥ := ⊥
    for t := T + 1 ... 0 do
        Choose independently S_t ~ S
        X_t^⊤ := f(X_{t-1}^⊤, S_t)
        X_t^⊥ := f(X_{t-1}^⊥, S_t)
until X_0^⊤ = X_0^⊥;
return X_0^⊤
```

Proof. Consider the moment before we return.
We have in essence picked T independent random variables $S_{T+1}, S_{T+2}, \ldots, S_0$ and used them to drive two chains, namely, $\{X_t^\top\}$ and $\{X_t^\perp\}$.

Since $f(x, s)$ is monotone with respect to x, we know for sure that X_0^x is bounded between $X_0^\top$ and $X_0^\perp$, for all $x \in \mathcal{X}$.
Let's imagine a fictional chain started at the born of Earth. We cut it off at time T, and drive it using our random variables $S_{T+1}, \ldots, S_0$ from then on. Since it runs long enough, it must have converged to stationary π at time 0. But since its current state is bounded by $X_0^\top$ from above, and $X_0^\perp$ from below, and because $X_0^\top = X_0^\perp$, we must confess that they coincide.
Thus, $X_0^\top \sim \pi$. ∎

$T := -1/2$
repeat
  $T := 2T$
  $X_T^\top := \top$; $X_T^\perp := \perp$
  for $t := T+1 \ldots 0$ **do**
    Choose independently $S_t \sim S$
    $X_t^\top := f(X_{t-1}^\top, S_t)$
    $X_t^\perp := f(X_{t-1}^\perp, S_t)$
until $X_0^\top = X_0^\perp$;
return $X_0^\top$

Proof.(Concise and formal version)
To avoid clutter, denote $f_t(x) := f(x, S_t)$. (It gives a random variable parameterised by x.) Further denote $g_{t_1}^{t_2} := f_{t_2} \circ \cdots \circ f_{t_1}$. Let $Y := g_{-\infty}^{T}(x)$, and $Z := g_{-\infty}^{0}(x)$. Clearly,
(1) $Z \sim \pi$;
(2) $Z = g_{T+1}^{0} \circ g_{-\infty}^{T}(x) = g_{T+1}^{0}(Y)$.
But we know that the function g_{T+1}^{0} collapses everything to a single point $X_0^\top$ when the procedure terminates, we must conclude that $Z = X_0^\top$. Thus $X_0^\top \sim \pi$. ∎

Coupling to the Future?

Coupling to the Future?

Problem. Try to devise a "coupling to the future" method, and see where the proof breaks down.

Problem. Why must we reuse variables that had been chosen in previous rounds? Can we use fresh random variables in each round? (*Hint: Consider the algorithm as a whole. What is the distribution of* $\{S_t\}$ *if you observe from outside?*)

Analysing the Expected Time

Definition. Let T^* be the value of $-T$ when the procedure exits.

We wish to bound $\mathbb{E}[T^*]$ by a polynomial.
A lower bound is immediate: $\mathbb{E}[T^*] \geq t_{\text{mix}}/4$.

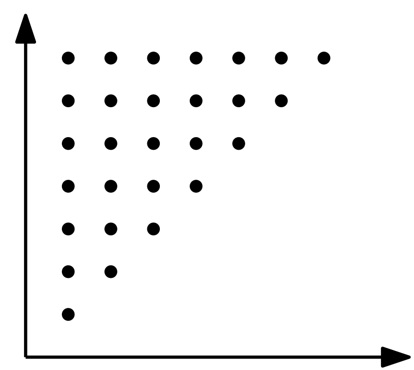
Analysing the Expected Time

Definition. Let T^* be the value of $-T$ when the procedure exits.

We wish to bound $\mathbb{E}[T^*]$ by a polynomial.
A lower bound is immediate: $\mathbb{E}[T^*] \geq t_{\mathsf{mix}}/4$.

What about the upper bound? We write the expectation as

$$\mathbb{E}[T^*] = \sum_{i=0}^{\infty} \Pr[T^* > i]$$



$$\boxed{\mathbb{E}[T^*] = \sum_{i=0}^{\infty} \text{Pr}[T^* > i]}$$

defined on $\mathcal{X}$ defined on $\mathbb{N}$

not $\mathcal{X}$!

Let $\phi : \mathcal{X} \to \mathbb{N}$ be a function satisfying $x < y \Rightarrow \phi(x) < \phi(y)$.

$$\begin{aligned}
\text{Pr}[T^* > i] &= \text{Pr}[X^{\top}_{0;-i} > X^{\perp}_{0;-i}] \\
&\leq \text{Pr}[\phi(X^{\top}_{0;-i}) - \phi(X^{\perp}_{0;-i}) > 0] \\
&\leq \mathbb{E}[\phi(X^{\top}_{0;-i}) - \phi(X^{\perp}_{0;-i})] \\
&= \mathbb{E}[\phi(X^{\top}_{0;-i})] - \mathbb{E}[\phi(X^{\perp}_{0;-i})] \\
&= \sum_{x \in \mathcal{X}} \phi(x)(\text{Pr}[X^{\top}_{0;-i} = x] - \text{Pr}[X^{\perp}_{0;-i} = x]) \\
&\leq h \cdot d(i)
\end{aligned}$$

where $h := \max_{x \in \mathcal{X}} \phi(x)$, and $d(i) := \max_{\mu,\nu} \|\mu P^i - \nu P^i\|$.

$$\boxed{\mathbb{E}[T^*] = \sum_{i=0}^{\infty} \Pr[T^* > i]}$$

$$\boxed{\Pr[T^* > i] \le \ell \cdot d(i)}$$

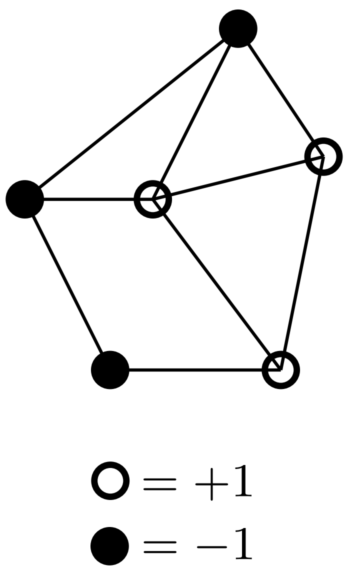
$\boxed{\text{And it's well-known that } d(k \cdot t) \le d(t)^k.}$

We will do the summation in *blocks* of size m.

$$\begin{aligned}
\mathbb{E}[T^*] &= \sum_{b=0}^{\infty} \sum_{r=0}^{m-1} \Pr[T^* > bm + r] \\
&\le \sum_{b=0}^{\infty} m \cdot \Pr[T^* > bm] \\
&\le \sum_{b=0}^{\infty} m \cdot h \cdot d(bm) \\
&\le \sum_{b=0}^{\infty} m \cdot h \cdot d(m)^b
\end{aligned}$$

Take, say, $m := t_{\text{mix}}$. ■

Realistic Example: The Ising Model



Definition. A *spin configuration* of a graph is an assignment $x : V \to \{+1, -1\}$.

Definition. The *energy* of a spin configuration σ is defined as $H(x) := -\sum_{\{u,v\}\in E} x(u)x(v)$.

Remark. The energy increases when the contention strengthens between neighbours. (Note that we consider only adjacent vertices, an approximation of the real world!)

In physical world, the probability that configuration σ occurs is given by $\pi(x) := \frac{1}{Z_\beta} e^{-\beta H(x)}$, with $\beta > 0$ a constant and Z_β the normalizing factor. This is what we call "the principle of minimum energy".

In physical world, the probability that configuration σ occurs is given by $\pi(x) := \frac{1}{Z_\beta} e^{-\beta H(x)}$, with $\beta > 0$ a constant and Z_β the normalizing factor. This is what we call "the principle of minimum energy".

Design of Markov chain:

- Space $\mathcal{X} := \{+1, -1\}^V$.
- Space $\mathcal{S} := V \times [0, 1]$.
- S is uniformly distributed on $\mathcal{S}$.
- $f(x, s)$ is defined as

  1. Unpack $s =: (v, r)$;
  2. Let x_+ and x_- be the configurations yielded from x_{t-1} by mapping the vertex v to $+1$ and -1, respectively;
  3. If $r < \frac{\pi(x_+)}{\pi(x_+)+\pi(x_-)}$, return x_+; otherwise, return x_-.

And we couple the chains by sharing randomness, as usual.

Definition. We define a partial order $\leq$ on space $\mathcal{X}$, much the same way as before:
$x \leq x' \iff \forall v \in V : x(v) \leq x'(v)$.

And we also have $X_{t-1}^{x} \leq X_{t-1}^{y} \Rightarrow X_{t}^{x} \leq X_{t}^{y}$ in our coupling.

Definition. We define a partial order $\leq$ on space $\mathcal{X}$, much the same way as before:
$x \leq x' \iff \forall v \in V : x(v) \leq x'(v)$.

And we also have $X_{t-1}^{x} \leq X_{t-1}^{y} \Rightarrow X_t^x \leq X_t^y$ in our coupling.

Finally, we have top element $\top =$ "all ones" and bottom element $\bot =$ "all minus ones". The height of the partial order, h, is of course $|V|$.

So we are done! ∎