Propp-Wilson Algorithm (and sampling the Ising model)

Danny Leshem, Nov 2009

References:

WHAT’S ON?

- Problems with MCMC (e.g. Gibbs sampler)
- Propp-Wilson algorithm
  - Formal description
  - Simple example
  - Proof of correctness
  - Counter-examples for algorithm variations
- Sandwiching technique
  - Simple example
- Ising model
  - Description
  - Exact sampling using Propp-Wilson with Sandwiching
THE PROBLEM WITH MCMC

- We want to sample from distribution $\pi$
  - Construct a (finite, irreducible, aperiodic) Markov chain with (unique) stationary distribution $\pi$
- MCMC (e.g. Gibbs sampler)
  - Starts with some initial state $s$
  - Returns a state $\tilde{s}$ distributed \textit{approximately} $\pi$
  - Must be run for \textit{sufficient} time
- Propp-Wilson
  - $\tilde{s}$ is distributed \textit{exactly} $\pi$ ("perfect simulation")
  - The algorithm \textit{stops automatically}
We want to simulate (on a computer) the running of a Markov chain with $S = \{s_1, \ldots, s_k\}$ and transition matrix $P$.

Assume we can generate $U \sim U[0,1]$.

- Not feasible on a practical computer, but let’s ignore it.

We say $\phi: S \times [0,1] \to S$ is a valid update function for $P$ if

$$\forall s_i, s_j \in S. \mathbb{P}(\phi(s_i, U) = s_j) = P_{i,j}$$

- Of course, update functions are not unique.

Given an initial state $s$ and a sequence $U_0, U_1, \ldots$ we can simulate a run of the chain.
PROPP-WILSON: BACKGROUND & NOTATION

- We want to sample from distribution $\pi$ on $S = \{s_1, \ldots, s_k\}$
- Construct a reversible, irreducible & aperiodic Markov chain with state space $S$ and stationary distribution $\pi$
- Let $P$ be the transition matrix of the chain
- Let $\phi: S \times [0,1] \rightarrow S$ be some update function for $P$
- Let $(N_1, N_2, \ldots)$ be an increasing sequence of positive integers
  - The negative numbers $(-N_1, -N_2, \ldots)$ will be used as starting times for the Markov chain
  - Usually $(N_1, N_2, \ldots) = (1, 2, 4, 8, \ldots)$
- Let $U_0, U_{-1}, \ldots$ be a sequence of i.i.d. random variables $\sim U[0,1]$
1. Set $m = 1$
2. For each $s \in S$, simulate the Markov chain starting at time $-N_m$ in state $s$ and running up to time 0 using the update function $\phi$ with the sequence $(U_{-N_m+1}, U_{-N_m+2}, \ldots, U_{-1}, U_{0})$
3. If all $k$ chains end up in the same state $\tilde{s}$ at time 0, output $\tilde{s}$ and stop
4. Set $m := m + 1$ and go to step 2

- **Note**: The same sequence $U_0, U_{-1}, \ldots$ is used for all $k$ chains.
PROPP-WILSON: SIMPLE EXAMPLE

- [HAG, p. 78]:

---

---

---

---
Is the algorithm guaranteed to stop?
  - Depends on the Markov chain and choice of update function

Example [HAG, problem 10.2]:

- Consider the Markov chain with \( P = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \)

- \( \phi_{good} (s_i, U) = \begin{cases} s_1 & U \in [0,0.5) \\ s_2 & U \in [0.5,1] \end{cases} \)

- \( \phi_{bad} (s_1, U) = \begin{cases} s_1 & U \in [0,0.5) \\ s_2 & U \in [0.5,1] \end{cases} \)

- \( \phi_{bad} (s_2, U) = \begin{cases} s_2 & U \in [0,0.5) \\ s_1 & U \in [0.5,1] \end{cases} \)

- \( \phi_{independent} (s_1, U) = \begin{cases} s_1 & U \in [0,0.25) \cup [0.5,0.75) \\ s_2 & U \in [0.25,0.5) \cup [0.75,1] \end{cases} \)

- \( \phi_{independent} (s_2, U) = \begin{cases} s_2 & U \in [0,0.5) \\ s_1 & U \in [0.5,1] \end{cases} \)
“0-1 law” for termination of the Propp-Wilson algorithm
- If every two initial states have a positive probability to meet after $R$ steps, they will (almost surely) meet by the Borel-Cantelli lemma (meeting time follows geometric distribution)
- Special case: if each of the $k$ Markov chains is (ergodic and) independent of the others, the algorithm will (almost surely) stop
  - But the right dependence can speed things up considerably
[HAG] **Theorem 10.1**: Let $S, P, \phi, \pi, (N_1, N_2, \ldots)$ as before. Suppose the Propp-Wilson algorithm terminates (almost surely), and let $Y$ be its output. Then for every $i \in \{1, \ldots, k\}$ we have:

$$\mathbb{P}(Y = s_i) = \pi_i$$

**Proof**: Let $s_i$ be any state and let $\varepsilon > 0$. Enough to show

$$|\mathbb{P}(Y = s_i) - \pi_i| \leq \varepsilon$$

The algorithm terminates (almost surely), so we can pick $M$ large enough so that

$$\mathbb{P}(\text{the algorithm ran at most } M \text{ steps}) \geq 1 - \varepsilon$$
PROPP-WILSON: PROOF OF CORRECTNESS 2/2

- We now use a coupling argument (“coupling from the past”):
  imagine running two chains from time \( -N_M \) to time 0 (using
  same update function and random variables)
  \[ (X_{-N_M}, ..., X_0) \] with initial state set to, say, \( s_1 \)
  \[ (\tilde{X}_{-N_M}, ..., \tilde{X}_0) \] with initial state chosen according to \( \pi \)
- Since \( \pi \) is stationary, \( \tilde{X}_0 \) is also distributed according to \( \pi \)
- The punch: for our large \( M \), we have \( \mathbb{P}(X_0 = \tilde{X}_0) \geq 1 - \varepsilon \)
  \[ \text{And by the coupling lemma: } X_0 \text{ is distributed like } \pi \text{ (up to}
  \text{ variation distance } \varepsilon) \]
- To complete the proof, we note that \( Y = X_0 \) (for “\( M = \infty \)”)?
  \[ \text{We showed that for every } \varepsilon > 0, \text{ } Y \text{ is distributed like } \pi \text{ (up to}
  \text{ variation distance } \varepsilon), \text{ meaning } Y \text{ is distributed exactly } \pi \]
VARIATION #1: “COUPLING TO THE FUTURE”

- Why bother with running the chains further and further into the past? Let’s start all chains at time 0, run them forward until the first time they “merge” and output that value!
- [HAG, counter-example 10.1]:
  - Consider the Markov chain with \( P = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix} \)
  - It is reversible with stationary distribution
    \[
    \pi = (\pi_1, \pi_2) = \left(\frac{2}{3}, \frac{1}{3}\right)
    \]
  - Consider running two chains starting in time 0, with initial states \( s_1, s_2 \). Say they merge for the first time in time \( N \). By definition, at time \( N - 1 \) they don’t have the same value, so one must be in \( s_2 \). Surely, then, that chain will be in \( s_1 \) in time \( N \)... and we sampled from the distribution \((1,0)\).
VARIATION #2: INDEPENDENT ITERATIONS 1/2

- Must we really save and reuse the random variables $(U_{-N_{m+1}}, U_{-N_{m+2}}, \ldots, U_{-1}, U_0)$ when restarting the chain at time $-N_{m+1}$? Let’s generate a new sequence in each iteration!

- [HAG, counter-example 10.2]:
  - Consider the Markov chain with $P = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix}$
  - Run Propp-Wilson with a new sequence of random variables every iteration, and take $(N_1, N_2, \ldots) = (1, 2, 4, 8, \ldots)$
  - Let $Y$ be the output of this algorithm, and let $M$ be the iteration in which the algorithm stopped.
VARIATION #2: INDEPENDENT ITERATIONS 2/2

- We have:

\[
P(Y = s_1) = \sum_{m=1}^{\infty} P(Y = s_1 | M = m) \geq P(Y = s_1, M = 1) + P(Y = s_1, M = 2) =
\]

\[
P(M = 1) \cdot P(Y = s_1 | M = 1) + P(M = 2) \cdot P(Y = s_1 | M = 2) = \frac{1}{2} \cdot 1 + \frac{3}{8} \cdot \frac{2}{3} = \frac{3}{4} > \frac{1}{3}
\]

- The calculation is justified [HAG, problem 10.3]:

\[
P(M = 2) = P(M \neq 1) \cdot P(M = 2 | M \neq 1) = \frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}
\]
THE SANDWICHING TECHNIQUE

- Propp-Wilson dictates that we follow $k$ different Markov chains (one for each state in $S$). For large $k$, this is unfeasible.
- Do we really need to run all chains?
  - some Markov chains obey certain monotonicity properties (with respect to some ordering of the state space)
  - allows us to run only a small subset of the $k$ chains (say, 2)
Consider the “ladder walk” Markov chain with
\[
P = \begin{bmatrix}
0.5 & 0.5 & 0 & 0 & 0 \\
0.5 & 0 & 0.5 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 \\
0 & 0 & 0.5 & 0 & 0.5 \\
0 & 0 & 0 & 0.5 & 0.5
\end{bmatrix}
\]

“Try to take a step up or down the ladder with probability 0.5. If you can’t, stay where you are”

The stationary distribution is uniform: \( \pi = \left( \frac{1}{5}, \ldots, \frac{1}{5} \right) \)

Define the following update function:
\[
\phi(s, U) = \begin{cases} 
\max(1, s - 1) & U \in [0,0.5) \\
\min(5, s + 1) & U \in (0.5,1] 
\end{cases}
\]

It preserves ordering between the states.
\[
x \in [0,1], \ i \leq j \in \{1, \ldots, 5\} \Rightarrow \phi(i, x) \leq \phi(j, x)
\]
All chains are bounded between the top and bottom chain. When the two meet, all others join as well!
- It’s enough to run just these two chains ($2 \ll k$)
Let $G = (V, E)$ be a graph

The Ising model is a way to pick a random $\xi \in \{-1,1\}^V$

- vertices are atoms in a ferromagnetic material
- we assign a spin orientation to each vertex

Let $\beta \geq 0$ be the fixed *inverse temperature* of the model

Define the *energy* of a spin configuration as
$$H(\xi) = -\sum_{\langle x, y \rangle \in E} \xi(x) \cdot \xi(y)$$

- “low energy = high agreement”

Define the probability measure $\pi_{G,\beta}$ as
$$\pi_{G,\beta}(\xi) = \frac{1}{Z_{G,\beta}} e^{-\beta \cdot H(\xi)}$$

- $Z_{G,\beta}$ is a just normalizing constant
The case $\beta = 0$ (“infinite temperature”)

- Each configuration is equally probable, meaning each vertex independently takes a value $\{-1,1\}$ w.p. $\frac{1}{2}$ each

The case $\beta \to \infty$ (“zero temperature”)

- The probability mass is equally divided between the “all plus” and “all minus” configurations

Phase transition phenomena

- The model depends quantitatively on whether $\beta$ is above or below a certain threshold value $\beta_c$

  - e.g. for $m \times m$ square lattice, $\beta_c = \frac{1}{2} \log(1 + \sqrt{2}) \approx 0.441$

For $\beta < \beta_c$, the proportion of +1 spins tends to $\frac{1}{2}$ as $m \to \infty$

For $\beta > \beta_c$, one of the spins “takes over” as $m \to \infty$
ISING MODEL: EXAMPLES

\( \beta = 0 \)

\( \beta = 0.15 \)

\( \beta = 0.3 \)

\( \beta = 0.5 \)
ISING MODEL: GIBBS SAMPLER 1/3

- Given $\beta, G = (V, E)$ define a Markov chain
  - $S = \{-1,1\}^V$
  - Given $X_n$ we obtain $X_{n+1}$ by picking a vertex $x \in V$ at random and then picking $X_{n+1}(x)$ according to the conditional distribution (under $\pi_{G,\beta}$) given $X_n(V \setminus \{x\})$. Other vertices remain unchanged.

- Notation
  - For $x \in V, \xi \in \{-1,1\}^V \setminus \{x\}$ define
    - $\xi^+ \in \{-1,1\}^V$ the expanded state s.t. $\xi^+(x) = +1$
    - $\xi^- \in \{-1,1\}^V$ the expanded state s.t. $\xi^-(x) = -1$
    - $k_+(x, \xi)$ the number of $+1$ neighbors of $x$ (in $\xi$)
    - $k_-(x, \xi)$ the number of $-1$ neighbors of $x$ (in $\xi$)
\textbf{Lemma} [HAG, problem 11.3]:
\[
\frac{\pi_{G,\beta}(\xi^+)}{\pi_{G,\beta}(\xi^-)} = e^{2\beta(k_+(x,\xi) - k_-(x,\xi))}
\]

\textbf{Proof:}
By definition:
\[
\pi_{G,\beta}(\xi) = \frac{1}{Z_{G,\beta}} e^{-\beta \cdot H(\xi)} = \frac{1}{Z_{G,\beta}} e^{\beta \cdot \sum_{(y,z) \in E} \xi(y) \cdot \xi(z)}
\]
So, ignoring identical elements (edges not connected to \(x\)):
\[
\frac{\pi_{G,\beta}(\xi^+)}{\pi_{G,\beta}(\xi^-)} = \frac{e^{\beta(k_+(x,\xi) - k_-(x,\xi))}}{e^{\beta(k_-(x,\xi) - k_+(x,\xi))}} = e^{2\beta(k_+(x,\xi) - k_-(x,\xi))}
\]
Corollary:

\[ \pi_{G,\beta}(X(x) = +1|X(V \setminus \{x\}) = \xi) = \frac{e^{2\beta (k_+(x,\xi) - k_-(x,\xi))}}{e^{2\beta (k_-(x,\xi) - k_+(x,\xi))} + 1} \]

Proof:

\[ \pi_{G,\beta}(X(x) = +1|X(V \setminus \{x\}) = \xi) = \]

\[ \frac{\pi_{G,\beta}(X(x) = +1 \text{ and } X(V \setminus \{x\}) = \xi)}{\pi_{G,\beta}(X(V \setminus \{x\}) = \xi)} = \]

\[ \frac{\pi_{G,\beta}(\xi^+)}{\pi_{G,\beta}(\xi^+) + \pi_{G,\beta}(\xi^-)} = \frac{\pi_{G,\beta}(\xi^+)}{\pi_{G,\beta}(\xi^+) + \pi_{G,\beta}(\xi^-)} \]

And then apply the lemma.
Define a (partial) ordering on the state space

For two configurations $\xi, \eta \in \{-1,1\}^v$ we write $\eta \preceq \xi$ if

$$\forall x \in V. \eta(x) \leq \xi(x)$$

We have unique $\xi_{\text{max}}$ ("all +1") and $\xi_{\text{min}}$ ("all -1")

By the corollary, our previous update function can be based on

$$X_{n+1}(x, \xi) = \begin{cases} +1 & U_{n+1} < \frac{e^{2\beta(k_+(x,\xi)-k_-(x,\xi))}}{e^{2\beta(k_-(x,\xi)-k_+(x,\xi))} + 1} \\ -1 & \text{otherwise} \end{cases}$$
Claim: Given any vertex $x$ and $\eta \leq \xi$, define

$$
\xi_2(y) = \begin{cases} 
\xi(y) & y \neq x \\
+1 & y = x, U < \frac{y \neq x}{e^{2\beta(k_+(x,\xi) - k_-(x,\xi))} + 1} \\
-1 & \text{otherwise}
\end{cases}
$$

And similarly define $\eta_2$. Then $\eta_2 \leq \xi_2$.

Proof: For $y \neq x$ this is a direct result of $\eta \leq \xi$.

For $x$, we note that $k_+(x,\xi) \geq k_+(x,\eta)$, $k_-(x,\xi) \leq k_-(x,\eta)$.

So $2\beta(k_+(x,\xi) - k_-(x,\xi)) \geq 2\beta(k_+(x,\eta) - k_-(x,\eta))$.

Finally, $\frac{e^{2\beta(k_+(x,\xi) - k_-(x,\xi))}}{e^{2\beta(k_-(x,\eta) - k_+(x,\eta))} + 1} \geq \frac{e^{2\beta(k_+(x,\eta) - k_-(x,\eta))}}{e^{2\beta(k_-(x,\eta) - k_+(x,\eta))} + 1}$ since $\frac{z}{z+1}$ is strictly increasing for $z \geq 0$. This proves the claim.
We showed it’s enough to run two chains only in order to sample a configuration from the Ising model.

But even so, do these chains “join” within a reasonable time?
  - Depends on the graph and on the temperature...
  - For the $m \times m$ square lattice
    - if $\beta < \beta_c$ then this (expected) time grows like a low-degree polynomial in $m$.
    - if $\beta > \beta_c$ it grows exponentially in $m$.
  - Propp & Wilson showed one can still use their algorithm (with another twist) to obtain samples in both cases.