Markov Chain Monte-Carlo (MCMC)

What for is it and what does it look like?

A. Favorov, 2003-2018
favorov@sensi.org
favorov@gmail.com
Monte Carlo method: a figure square

The value $\mu$ is unknown.

Let’s sample a random value (r.v.) $\xi$:

\[
\{x, y\}: \text{i.i.d. as flat}[0,1] \\
\begin{align*}
\xi = 1 &\iff (x, y) \in \mu \\
\xi = 0 &\iff (x, y) \notin \mu
\end{align*}
\]

Clever notation: $\xi = I_\mu(x, y)$

i.i.d. is “Identically Independently Distributed”

Expectation of $\xi$: $E\{\xi\} = \bar{\xi} = S(\mu) = S$
Monte Carlo method: efficiency

Large Numbers Law: \( S \approx \hat{S}_m = \frac{1}{m} \sum_{i=1}^{m} \xi_i \)

Central Limit Theorem: \( S - \hat{S}_m \to \frac{1}{\sqrt{m}} \cdot N(0, \text{var} \{\xi\}) \)

Variance \( \text{var} \{\xi\} = E\left(\left[\xi - E(\xi)\right]^2\right) \), also notated as \( \sigma^2 \).
Monte Carlo Integration

We are evaluating $I = \int_D f(x) dx$. $D$ is domain of $f(x)$ or its subset.

We can sample r.v. $x_i \in D$: $x_i$ are i.i.d. uniformly in $D$: $E[f(x_i)] = \frac{1}{|D|} \int_D f(x) dx = I$.

The Monte Carlo estimation: $\hat{I}_m = \frac{|D|}{m} \sum_{i=1}^{m} f(x_i)$,

$I - \hat{I}_m \to \frac{|D|}{\sqrt{m}} \cdot N(0, \text{var}_D \{f(x)\})$

Advantage:

- The multiplier $\sim m^{-\frac{1}{2}}$ does not depend on the space dimension.

Disadvantage:

- a lot of samples are spent in the area where $f(x)$ is small;
- the variation value $\text{var}_D \{f(x)\}$ that determine convergence time can be large.
Monte Carlo importance integration

We are evaluating $I = \int_{D} f(x) \, dx$

Let’s sample $x_i \in D$ from a “trial” distribution $g(x)$ that “looks like” $f(x)$ and $\left| f(x) \right| > 0 \Rightarrow g(x) > 0$. $x_i$ i.i.d. in $D$ as $g(x)$ that “resembles” $f(x)$

Thus $E_g \left( \frac{f(x_i)}{g(x_i)} \right) = \int_{D} \frac{f(x)}{g(x)} g(x) \, dx = \int_{D} f(x) \, dx$

MC evaluation: $\hat{I}_m = \frac{1}{m} \sum_{i=1}^{m} \frac{f(x_i)}{g(x_i)}$; $I - \hat{I}_m \rightarrow \frac{1}{\sqrt{m}} \cdot N \left( 0, \text{var}_{D} \left\{ \frac{f(x)}{g(x)} \right\} \right)$

“More uniform” means “better”.
Another example of importance integration

We are evaluating $\mu = E_{\pi} \{h(x)\} = \int h(x)\pi(x)dx$, where $\pi(x)$ is a distribution, e.g. $\int \pi(x)dx = 1$

- sample $x_i$ from a distribution $g(\cdot)$ so that $\pi(x) > 0 \Rightarrow g(x) > 0$
- Importance weight $w_i = \pi(x_i)/g(x_i)$; $E_g \{w(x)\} = \int \frac{\pi(x)}{g(x)}g(x)dx = 1$

$$\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \frac{\pi(x_i)}{g(x_i)}h(x_i) = \frac{1}{m} \sum_{i=1}^{m} w(x_i)h(x_i) = \frac{\sum_{i=1}^{m} w(x_i)h(x_i)}{\sum_{i=1}^{m} w(x_i)}$$

$$\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} w(x_i)h(x_i) = \frac{\sum_{i=1}^{m} w(x_i)h(x_i)}{\sum_{i=1}^{m} w(x_i)}$$

- Sampling from $\pi(x)$: $\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} h(x_i)$
Rejection sampling (Von Neumann, 1951)

We have a distribution \( \pi(x) \) and we want to sample from it.

We are able to calculate \( f(x) = c \cdot \pi(x) \) for \( \forall x \). Any \( c \).

We are able to sample \( g(x), \exists M : Mg(x) \geq f(x) \).

Thus, we can sample \( \pi(x) \):

- Draw a value \( x \) from \( g(x) \).
- Accept the value \( x \) with the probability \( f(x)/Mg(x) \).

\[
P(\text{accept}) = \int P(\text{accept} | x) P(x) \cdot dx = \int \frac{c \cdot \pi(x)}{Mg(x)} \cdot g(x) \cdot dx = \frac{c}{M}
\]

\[
P(x | \text{accept}) = \frac{P(\text{accept} | x) \cdot P(x)}{P(\text{accept})} = \frac{c \cdot \pi(x)}{Mg(x)} \cdot g(x) \cdot \frac{M}{c} = \pi(x)
\]
We want to be able to draw \( x^{(i)} \) from a distribution \( \pi(x) \). We know how to compute the value of a function \( f(x) \) so that \( f(x) \sim \pi(x) \) at each point and we are able to draw \( x \) from \( T(x | y) \) (instrumental distribution, transition kernel).

Let’s denote the \( i \)-th step result as \( x^{(i)} \).

- Draw \( y^{(i)} \) from \( T(y | x^{(i)}) \). \( T(y | x^{(i)}) \) is flat in pure Metropolis. It is an analog of \( g \) in importance sampling.
- Transition probability

\[
\varphi(y^{(i)} | x^{(i)}) = \min\left(1, \frac{T(x^{(i)} | y^{(i)}) \cdot f(y^{(i)})}{T(y^{(i)} | x^{(i)}) \cdot f(x^{(i)})}\right).
\]

- The new value is accepted \( x^{(i+1)} = y^{(i)} \) with probability \( \varphi(y^{(i)} | x^{(i)}) \). Otherwise, it is rejected \( x^{(i+1)} = x^{(i)} \).
Why does it work: the local balance

Let’s show that if $x$ is already distributed as $\pi(x) \sim f(x)$, then the MH algorithm keeps the distribution.

Local balance condition for two points $x$ and $y$:

$$f(x) \cdot T(y|x) \cdot \varphi(y|x) = f(y) \cdot T(x|y) \cdot \varphi(x|y).$$

Let’s check it:

$$f(x) \cdot T(y|x) \cdot \varphi(y|x) = f(x) \cdot T(y|x) \cdot \min\left(1, \frac{T(x|y) \cdot f(y)}{T(y|x) \cdot f(x)}\right) =$$

$$= \min(T(y|x) \cdot f(x), T(x|y) \cdot f(y)) = f(y) \cdot T(x|y) \cdot \varphi(x|y)$$

The balance is stable:

$f(x) \cdot T(y|x) \cdot \varphi(y|x)$ is the flow from $x$ to $y$ and

$f(y) \cdot T(x|y) \cdot \varphi(x|y)$ is the flow from $y$ to $x$.

The stable local balance is enough (BTW, it is not a necessary condition).
Markov chains, Maximization, Simulated Annealing

$x_i$ created as described above is a Markov chain (MC) with transition kernel $\varphi(x^{(i+1)} | x^{(i)}) \cdot T(x^{(i+1)} | x^{(i)})$. The fact that the chain has a stationary distribution and the convergence of the chain to the distribution can be proved by the MC theory methods.

**Minimization.** $C(x)$ is a cost (a fine). $f(x) = \exp \left( - \frac{C(x) - C_{\text{min}}}{t} \right)$.

We can characterize the transition kernel with a temperature. Then we can decrease the temperature step-by-step (simulated annealing). MCMC and SA are very effective for optimization since gradient methods use to be locked is a local maximum while pure MC is extremely ineffective.
MCMC and its variations are often used for the best model search.

Let’s can formulate some requirements for the algorithm and thus for the transition kernel:

- We want it not to depend on the current data.
- We want to minimize the rejection rate.

So, an effective transition kernel is so that the prior $P(M)$ is its stationary distribution.
Terminology: names of relative algorithms

- MCMC, Metropolis, Metropolis-Hastings, hybrid Metropolis, configurational bias Monte-Carlo, exchange Monte-Carlo, multigrid Monte-Carlo (MGMC), slice sampling, RJMCMC (samples the dimensionality of the space), Multiple-Try Metropolis, Hybrid Monte-Carlo…..

- Simulated annealing, Monte-Carlo annealing, statistical cooling, umbrella sampling, probabilistic hill climbing, probabilistic exchange algorithm, parallel tempering, stochastic relaxation…..

- Gibbs algorithm, successive over-relaxation….
Gibbs Sampler (Geman and Geman, 1984)

Now, \( x \) is a \( k \)-dimensional variable \((x_1, x_2, \ldots, x_k)\).

Let’s denote \( x_{-m} = (x_1, x_2, \ldots, x_{m-1}, x_{m+1}, \ldots, x_k) \), \( 1 \leq m \leq k \)

On each step of the Markov Chain we choose the “current coordinate” \( m_i \).

Then, we calculate the distribution \( f(x_{m_i} | x_{-m_i}) \) and draw the next value \( y_{m_i}^{(i)} \) from the distribution.

All other coords are the same as on the previous step, \( y_{-m_i}^{(i)} = x_{-m_i}^{(i)} \).

For such a transition kernel,

\[
\varphi(y^{(i)} | x^{(i)}) = \min \left( 1, \frac{T(x^{(i)} | y^{(i)}) \cdot f(y^{(i)})}{T(y^{(i)} | x^{(i)}) \cdot f(x^{(i)})} \right) = 1.
\]

- We have no rejects, so the procedure is very effective.
- The “temperature” decreases rather fast.
Inverse transform sampling (well-known)

We want to sample from the density $\pi(x)$. We know how to calculate the inverse function for the cumulative distribution.

- Generate a random number from the $[0,1]$ uniform distribution; call this $u_i$.
- Compute the value $x_i$ such that $\int_{-\infty}^{x_i} \pi(x) \, dx = u_i$
- $x_i$ is the random number that is drawn from the distribution described by $\pi(x)$.

\[ [x, x + \Delta x] \leftrightarrow [u, u + \Delta u] \]
\[ p(x) \Delta x = \text{uniform}(u) \times \Delta u \]
\[ p(x) = \text{uniform} \times \frac{\Delta u}{\Delta x} = \pi(x) \]
Slice sampling (Neal, 2003)

Sampling of $x$ from $f(x)$ is equivalent to sampling of $(x, y)$ pairs from they area.

So, we introduce an auxiliary variable $y$ and iterate as follows:

- for a sample $x_t$ we choose $y_t$ uniformly from the interval $[0, f(x_t)]$
- given $y_t$ we choose $x_{t+1}$ uniformly at random from $\{x : f(x) > y_t\}$

the sample of $x$ distributed as $f(x)$ is obtained by ignoring the $y$ values.
Literature


Sheldon Ross. A First Course in Probability

Sometimes, it works 😊
