Markov Chain Monte-Carlo (MCMC)

What for is it and what does it look like?

A. Favorov, 2003-2013

favorov@sensi.org

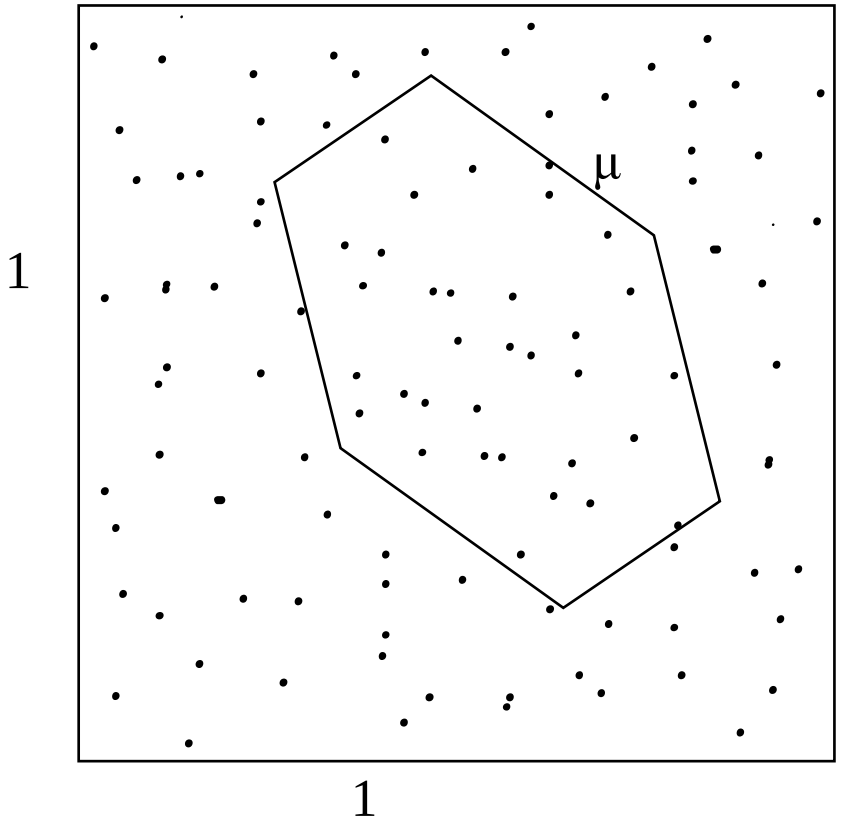
favorov@gmail.com

Monte Carlo method: a figure square

The value μ is unknown.

Let's sample a random value (**r.v.**) ξ:

$$\{x, y\}: \text{i.i.d. as } \text{flat}[0,1]$$
$$\begin{cases} \xi = 1 \Leftrightarrow (x, y) \in \mu \\ \xi = 0 \Leftrightarrow (x, y) \notin \mu \end{cases}$$



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

i.i.d. is "Identically Independently Distributed"

Expectation of ξ: $E\{\xi\} = \overline{\xi} = S(\mu) = S$

Monte Carlo method: efficiency

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

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

Variance $\mathrm{var}\{\xi\} = E\left(\left[\xi - E(\xi)\right]^2\right)$, also notated as σ^2.

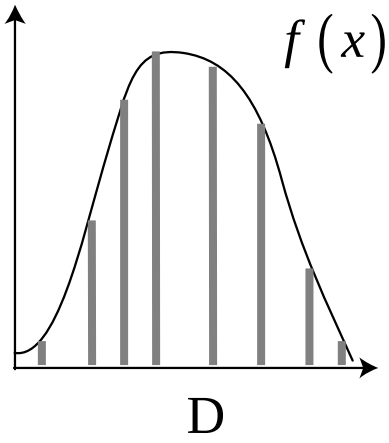
Monte Carlo Integration

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: $E\left[f(x_i)\right] = \frac{1}{|D|}\int_D f(x)\,dx = I$.

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

$$I - \widehat{I_m} \to \frac{|D|}{\sqrt{m}} \cdot N\left(0, \mathrm{var}_D\left\{f(x)\right\}\right)$$



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 $\mathrm{var}_D\left\{f(x)\right\}$ that determine convergence time can be large.

Monte Carlo importance sampling

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 $|f(x)| > 0 \Rightarrow g(x) > 0$. 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: $\widehat{I_m} = \frac{1}{m}\sum_{i=1}^{m} \frac{f(x_i)}{g(x_i)}$; $I - \widehat{I_m} \to \frac{1}{\sqrt{m}} \cdot N\left(0, \mathrm{var}_D\left\{\frac{f(x)}{g(x)}\right\}\right)$

"More uniform" means "better".

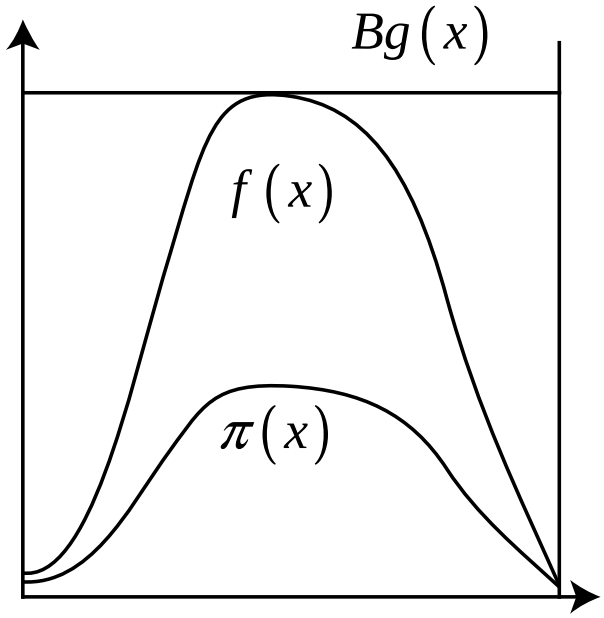
Another example of importance sampling

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 $g(.)$ so that $\pi(x) > 0 \Rightarrow g(x) > 0$
- Importance weight $w_i = \pi(x_i)/g(x_i)$; $E_g\{w(x)\} = \int \frac{\pi(x)}{g(x)} g(x)dx = 1$
- $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} \frac{\pi(x_i)}{g(x_i)} h(x_i) = \frac{1}{m}\sum_{i=1}^{m} w(x_i)h(x_i) = \sum_{i=1}^{m} w_i h(x_i) \Big/ \sum_{i=1}^{m} w_i$
- $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} w(x_i)h(x_i) = \sum_{i=1}^{m} w_i h(x_i) \Big/ \sum_{i=1}^{m} w_i$
- 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 $g(x), \exists B : Bg(x) \geq f(x)$.

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

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

$$P(accept) = \int P(accept \mid x) P(x) \cdot dx = \int \frac{c \cdot \pi(x)}{Bg(x)} \cdot g(x) \cdot dx = \frac{c}{B}$$

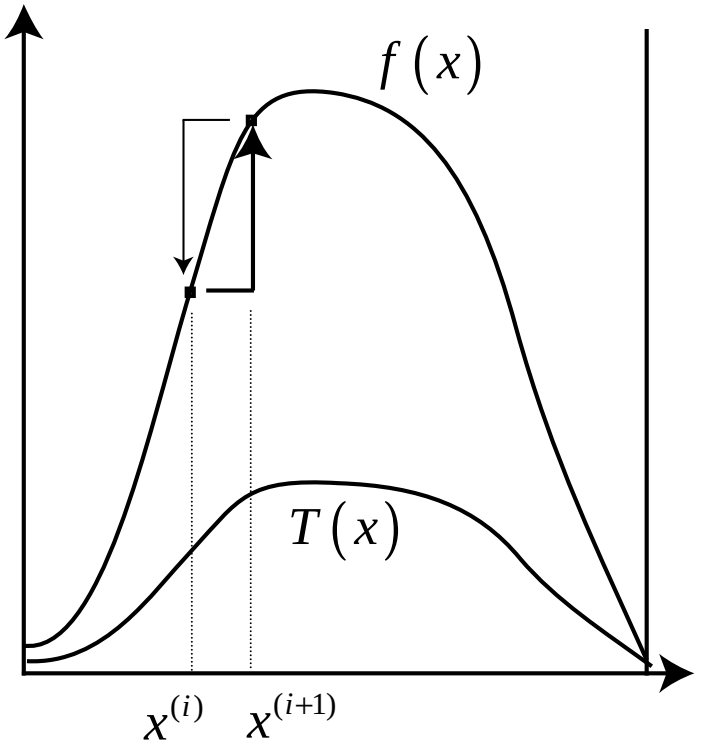
$$P(x \mid accept) = \frac{P(accept \mid x) \cdot P(x)}{P(accept)} = \frac{c \cdot \pi(x)}{Bg(x)} \cdot g(x) \cdot \frac{B}{c} = \pi(x)$$

Metropolis algorithm (1953)

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 flat distribution.

.

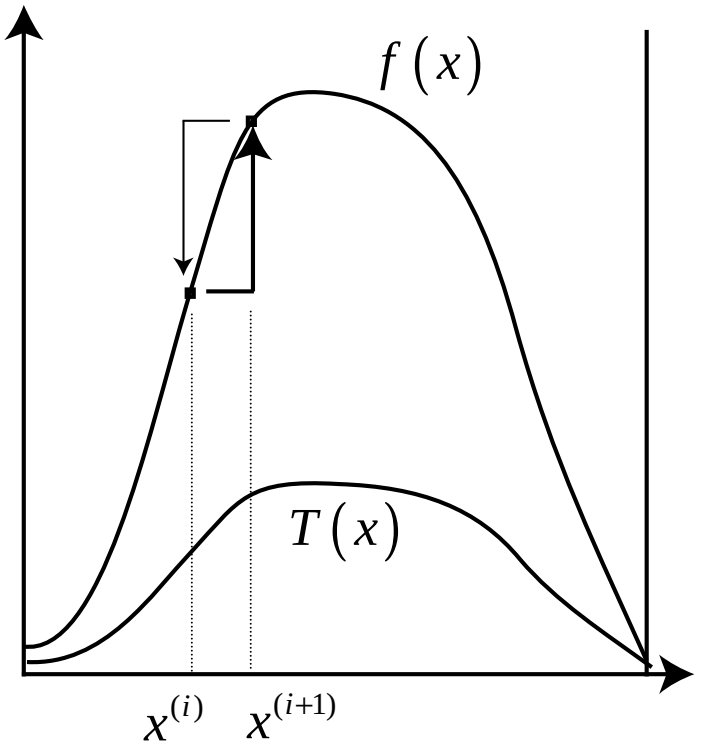
Let's denote the i-th step result as $x^{(i)}$.



- Draw $y^{(i)}$ from flat. It is an analog of g in importance sampling.
- Transition probability $\varphi\left(y^{(i)} \mid x^{(i)}\right) = \min\left(1, \frac{f\left(y^{(i)}\right)}{f\left(x^{(i)}\right)}\right)$.
- The new value is accepted $x^{(i+1)} = y^{(i)}$ with probability $\varphi\left(y^{(i)} \mid x^{(i)}\right)$. Otherwise, it is rejected and $x^{(i+1)} = x^{(i)}$.

Metropolis-Hastings algorithm (1953,1970)

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 $T(x \mid y)$ (instrumental distribution, transition kernel).
Let's denote the i-th step result as $x^{(i)}$.



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

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

- The new value is accepted $x^{(i+1)} = y^{(i)}$ with probability $\varphi\left(y^{(i)} \mid x^{(i)}\right)$. 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 $\pi(\cdot) \sim f(\cdot)$, then the MH algorithm keeps the distribution.

Local balance condition for two points x and y : $flux(x \to y) = flux(y \to x)$

Let's check it:

$$flux(x \to y) = f(x) \cdot T(y \mid x) \cdot \varphi(y \mid x);\ flux(y \to x) = f(y) \cdot T(x \mid y) \cdot \varphi(x \mid y)$$

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

$$= \min\left(T(y \mid x) \cdot f(x), T(x \mid y) \cdot f(y)\right) = f(y) \cdot T(x \mid y) \cdot \varphi(x \mid y) = flux(y \to x)$$

Why does it work: the local balance stability

Let's suppose a deviation from the $f(x)$ distribution: $f_{real}(x) = f(x) + \Delta$.

What happen with the fluxes?

$$flux_{new}(y \to x) = f(y) \cdot T(x \mid y) \cdot \varphi(x \mid y) = flux(y \to x)$$

$$flux_{new}(x \to y) = f_{new}(x) \cdot T(y \mid x) \cdot \varphi(y \mid x)$$

$$= f_{real}(x) \cdot T(y \mid x) \cdot \varphi(y \mid x)$$

$$= flux(y \to x) + \Delta \cdot T(y \mid x) \cdot \varphi(y \mid x)$$

$$= flux_{new}(y \to x) + \Delta \cdot T(y \mid x) \cdot \varphi(y \mid x)$$

The change in flux compensate the deviation. The balance is stable.

$f(x)$ distribution is a stable distribution for the MH Markov chain.

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 $\varphi\left(x^{(i+1)} \mid x^{(i)}\right) \cdot T\left(x^{(i+1)} \mid x^{(i)}\right)$. 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). $f(x) = \exp\left(-\frac{C(x) - C_{\min}}{t}\right)$.

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 prior and Bayesian paradigm

$$P(M \mid D) = \frac{P(D \mid M) \cdot P(M)}{P(D)} \propto P(D \mid M) \cdot P(M)$$ here, evidence

posterior $\propto$ likelihood $\cdot$ prior

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 $\left(x_1, x_2 .... x_k\right)$.

Let's denote $x_{-m} = \left(x_1, x_2 .., x_{m-1}, x_{m+1}, ... x_k\right), 1 \le m \le k$

On each step of the Markov Chain we choose the "current coordinate" m_i .

Then, we calculate the distribution $f\left(x_{m_i} \mid x^{(i)}_{-m_i}\right)$ and draw the next value $y^{(i)}_{m_i}$ from the distribution.

All other coords are the same as on the previous step, $y^{(i)}_{-m_i} = x^{(i)}_{-m_i}$.

For such a transition kernel,

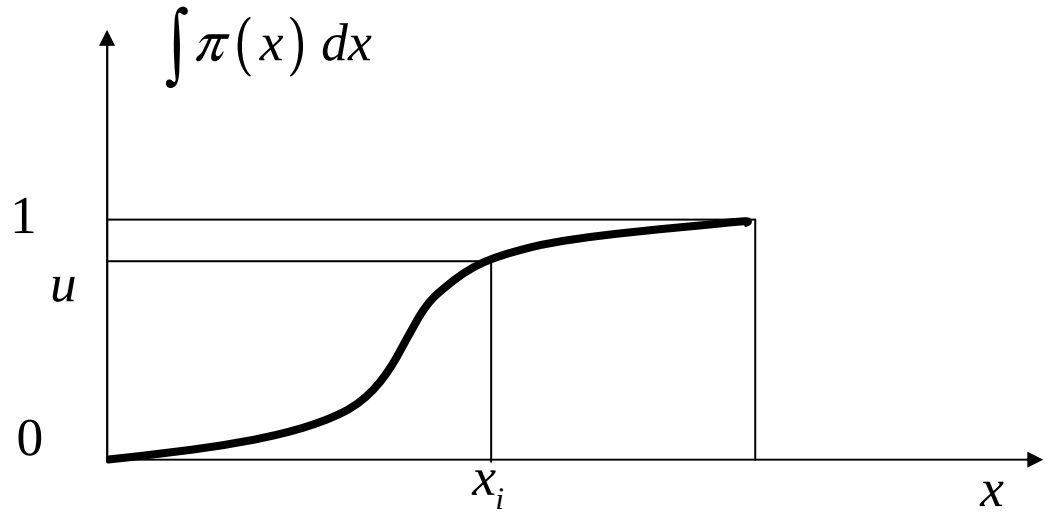
$$\varphi\left(y^{(i)} \mid x^{(i)}\right) = \min\left(1, \frac{T\left(x^{(i)} \mid y^{(i)}\right) \cdot f\left(y^{(i)}\right)}{T\left(y^{(i)} \mid x^{(i)}\right) \cdot f\left(x^{(i)}\right)}\right) = 1 .$$

- 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 $\pi(x)$.



$$[x, x+\Delta x] \leftrightarrow [u, u+\Delta u]$$

$$p(x)\Delta x = uniform(u) \times \Delta u$$

$$p(x) = uniform \times \frac{\Delta u}{\Delta x} = \pi(x)$$

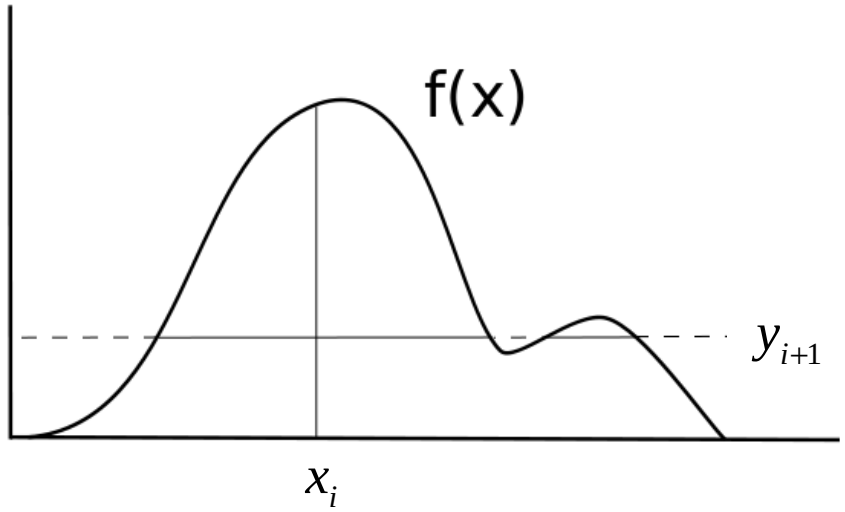
Slice sampling (Neal, 2003)

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 $\left[0, f(x_t)\right]$
- given y_t we choose x_{i+1} uniformly at random from $\left\{x : f(x) > y_t\right\}$

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



Literature

Liu, J.S. (2002) *Monte Carlo Strategies in Scientific Computing.* Springer-Verlag, NY, Berlin, Heidelberg.

Robert, C.P. (1998) *Discretization and MCMC Convergence Assessment*, Springer-Verlag.

Laarhoven, van, P.M.J. and Aarts, E.H.L (1988) *Simulated Annealing: Theory and Applications.* Kluwer Academic Publishers.

Geman, S and Geman, D (1984). *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images.* IEEE Transactions on Pattern Analysis and Machine Intelligence. **6**, 621-641.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1996) *Bayesian computation and Stochastic Sytems.* Statistical Science, **10**, 1, 3-66.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993). *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.* Science **262**, 208-214.

Sivia, D.S. (1996) *Data Analysis. A Bayesian tutorial.* Clarendon Press, Oxford.

Neal, Radford M. (2003) *Slice Sampling*. The Annals of Statistics 31(3):705-767.

http://civs.ucla.edu/MCMC/MCMC_tutorial.htm

Sheldon Ross. A First Course in Probability

Соболь И.М. Метод Монте-Карло

Sometimes, it works ☺

Favorov, A.V., Andreewski, T.V., Sudomoina, M.A., Favorova O.O., Parmigiani, G. Ochs, M.F. (2005). *A Markov chain Monte Carlo technique for identification of combinations of allelic variants underlying complex diseases in humans.* Genetics **171**(4): 2113-21.

Favorov, A.V., Gelfand, M.S., Gerasimova, A.V. Ravcheev, D.A., Mironov, A.A., Makeev, V. J. (2005). *A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length.* Bioinformatics **21**(10): 2240-2245.