Sampling. MCMC

E. Burnaev, Skoltech

Bayesian Inference

prior $f(\theta)$

data $X^n = (X_1, \ldots, X_n)$

posterior $f(\theta|X^n) = \dfrac{\mathcal{L}(\theta) f(\theta)}{c}$

normalizing constantc $c = \int \mathcal{L}(\theta) f(\theta)\, d\theta$

E.g. posterior mean value

$$\overline{\theta} = \int \theta f(\theta|X^n) d\theta = \frac{\int \theta \mathcal{L}(\theta) f(\theta) d\theta}{c}$$

Bayesian Inference

prior $f(\theta)$

data $X^n = (X_1, \ldots, X_n)$

posterior $f(\theta|X^n) = \dfrac{\mathcal{L}(\theta) f(\theta)}{c}$

normalizing constantc $c = \int \mathcal{L}(\theta) f(\theta)\, d\theta$

E.g. posterior mean value

$$\bar{\theta} = \int \theta f(\theta|X^n) d\theta = \frac{\int \theta \mathcal{L}(\theta) f(\theta) d\theta}{c}$$

Importance Sampling

probability density $f(x)$

integral we want to estimate $I = \int h(x) f(x) dx$

$$I = \int h(x) f(x) dx = \int \frac{h(x) f(x)}{g(x)} g(x) dx = \mathbb{E}_g(Y)$$

with $Y = h(X) f(X) / g(X)$

We simulate $X_1, \ldots, X_N \sim g$

$$\widehat{I} = \frac{1}{N} \sum_i Y_i = \frac{1}{N} \sum_i \frac{h(X_i) f(X_i)}{g(X_i)}$$

Importance Sampling

$$I = \int h(x)f(x)dx = \int \frac{h(x)f(x)}{g(x)} g(x)dx = \mathbb{E}_g(Y)$$

$$X_1, \ldots, X_N \sim g \qquad \widehat{I} = \frac{1}{N}\sum_i Y_i = \frac{1}{N}\sum_i \frac{h(X_i)f(X_i)}{g(X_i)}$$

We denote by $w(x) = h(x)f(x)/g(x)$

$$\mathbb{E}_g(w^2(X)) = \int \left(\frac{h(x)f(x)}{g(x)}\right)^2 g(x)dx = \int \frac{h^2(x)f^2(x)}{g(x)}dx$$

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(s)|f(s)ds}$$ **minimizes variance of** $\widehat{I}$

Importance Sampling

Tail probability $I = \mathbb{P}(Z > 3) = .0013$ **with**
$Z \sim N(0,1)$.

$I = \int h(x) f(x) dx$

$f(x)$ **is N(0,1)**

$h(x) = 1$ **if** $x > 3$ **and 0 otherwise, With N = 100 observ.**

$\mathbb{E}(\widehat{I}) = .0015 \quad \mathbb{V}(\widehat{I}) = .0039$

a lot of data points will be in the middle, not in tails

Importance Sampling

Tail probability $I = \mathbb{P}(Z > 3) = .0013$ **with**
$Z \sim N(0,1)$.

$I = \int h(x) f(x) dx$

$f(x)$ **is N(0,1)**

$h(x) = 1$ **if** $x > 3$ **and 0 otherwise, With N = 100 observ.**

g ~ $\mathrm{Normal}(4,1)$

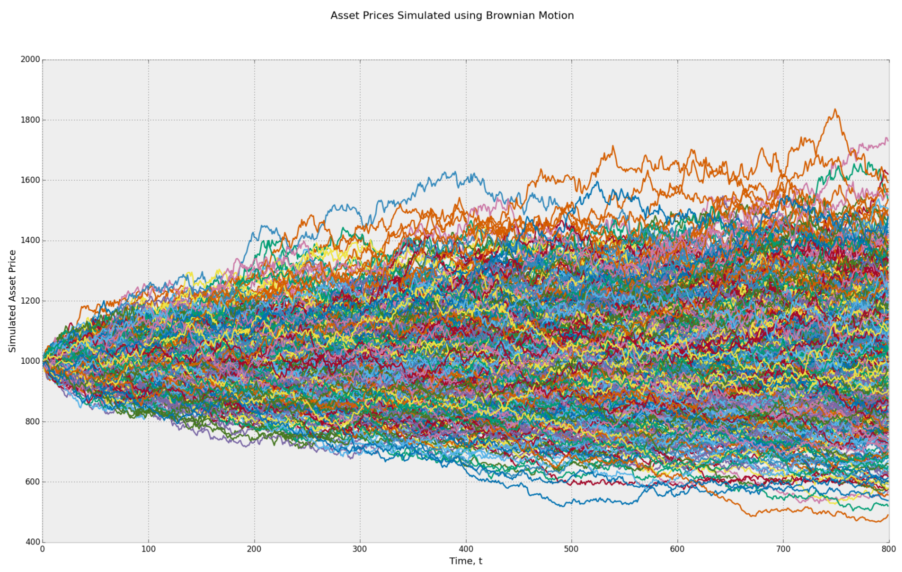
$\mathbb{E}(\widehat{I}) = .0011$ **(0.0015 before)**

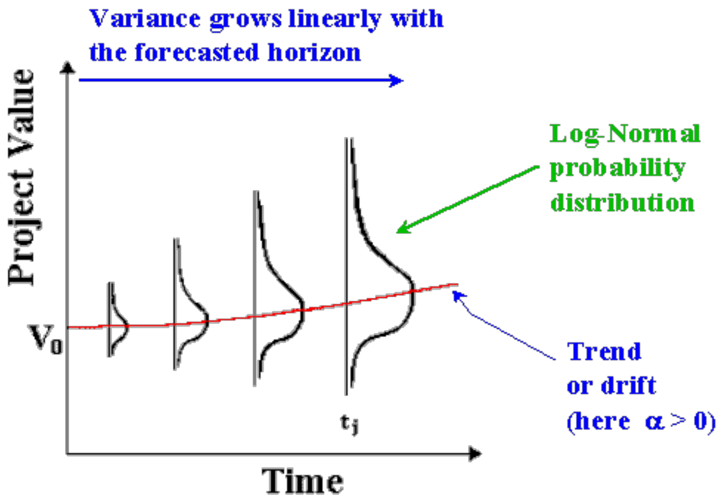
$\mathbb{V}(\widehat{I}) = .0002$ **(0.0039 before)**

Markov Processes

X(***t***) **is a random process**

$\left\{X_i = X\left(t_i\right)\right\}_{i=1}^{n}$ **is a finite-dimensional set of cross-sections**





Markov Processes

X(***t***) **is a random process**

$\left\{X_i = X\left(t_i\right)\right\}_{i=1}^{n}$ **is a finite-dimensional set of cross-sections**

In a general case:

$$F_{X_n}(u_n, t_n \,|\, (X_1, \ldots; X_{n-1}) \in B^{(n-1)}) \equiv P\{X(t_n) < u_n \,|\, (X_1, \ldots; X_{n-1}) \in B^{(n-1)}\}$$

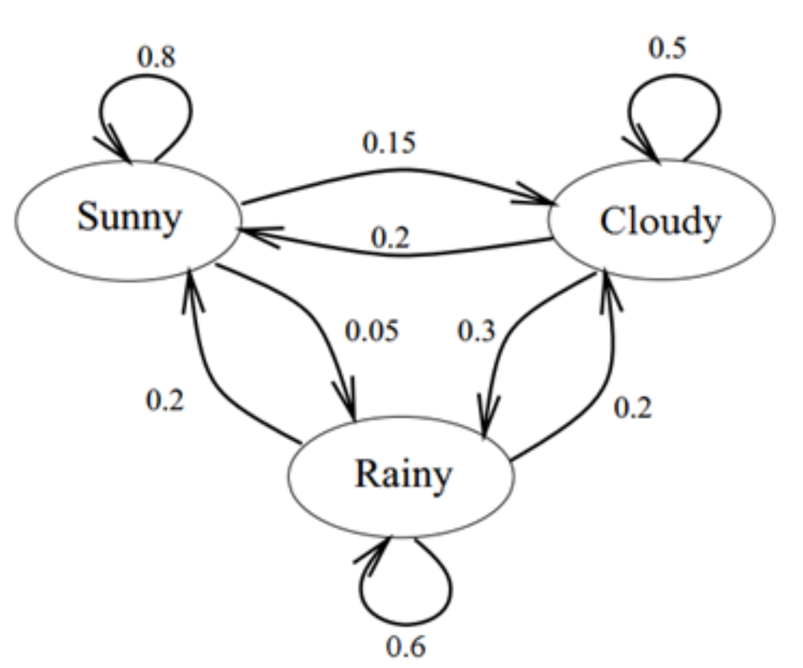
Definition: ***X***(***t***) **is a Markov Process iff** $\forall n,\ t_1 < t_2 < \ldots < t_n,\ x_{n-1}, B^{(n-2)}$

$$F_{X_n}(u_n, t_n \,|\, x_{n-1}, t_{n-1}, (x_1, \ldots, x_{n-2}) \in B^{(n-2)}) \equiv F_{X_n}(u_n, t_n \,|\, x_{n-1}, t_{n-1}).$$

Discrete Markov Chain

S **is a set of states, e.g.**

$S = (-k, \ldots, -1, 0, 1, \ldots, k),$ $\qquad S = (0, 1, 2, \ldots, k)$



Discrete Markov Chain

***S* is a set of states, e.g.**

$S = (-k, \ldots, -1, 0, 1, \ldots, k),$ $\qquad$ $S = (0, 1, 2, \ldots, k)$

$$\forall m_0 < m_1 < \ldots < m_{n-2} < m < n,\ i,\ j,\ i_0, \ldots i_{n-2}$$

$$P\{X(n) = j \mid X(m) = i,\ X(m_{n-2}) = i_{n-2}, \ldots, X(m_0) = i_0\} =$$

$$= P\{X(n) = j \mid X(m) = i\} = p_{ij}(m, n)$$

Kolmogorov-Chapman equation

$$p_{ij}(0, n) = \sum_{k \in S} p_{ik}(0, m)\, p_{kj}(m, n)$$

Discrete Markov Chain

Kolmogorov-Chapman equation

$$p_{ij}(0,n) = \sum_{k\in S} p_{ik}(0,m)\, p_{kj}(m,n)$$

$$\vec{\mathbf{p}}(m) = \| p_j(m) \| = \| P\{X(m) = j\} \|$$

$$\vec{\mathbf{p}}(n) = \boldsymbol{P}^{\mathrm{T}}(m,n)\, \vec{\mathbf{p}}(m)$$

In homogeneous case

$$\boldsymbol{P}^{(n)} = \boldsymbol{P}(\mathrm{n}-1,n) = \| p_{ij}^{(n)} \| = \| p_{ij} \|$$

$$\forall i, j, m, o < m < n;\ \ p_{ij}(n) = \sum_{k\in S} p_{ik}(n-m)\, p_{kj}(m)$$

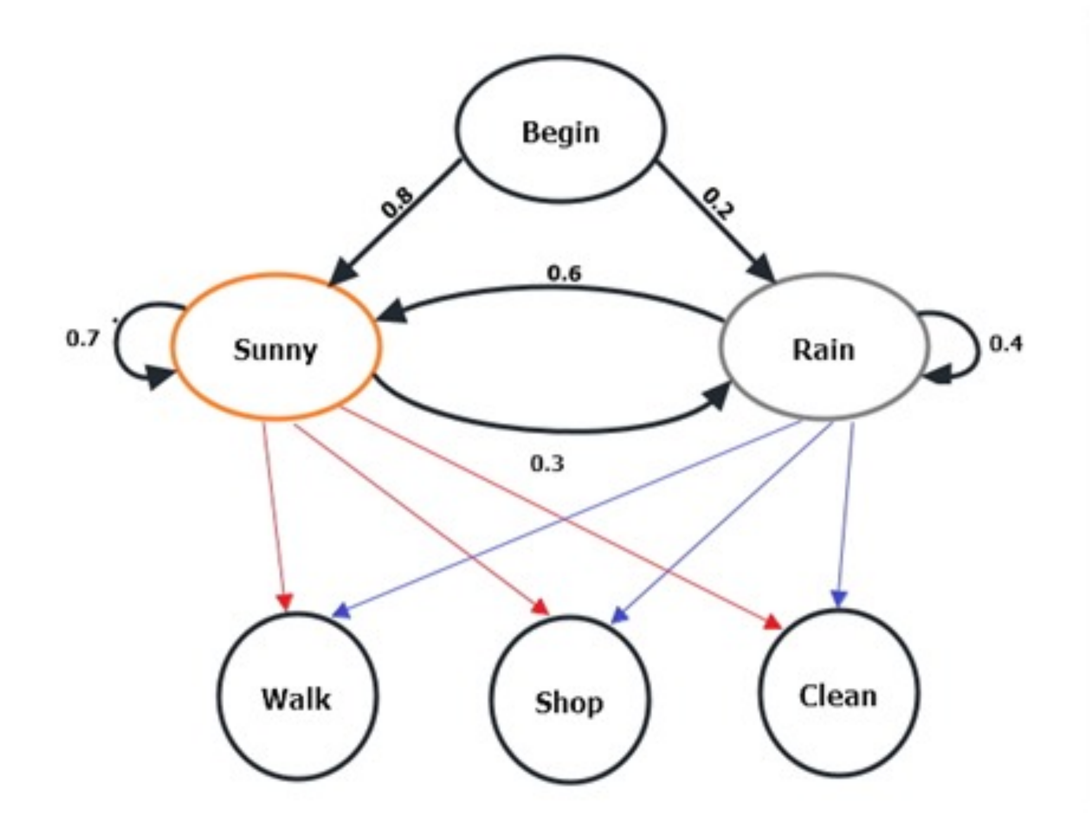
$$\boldsymbol{P}(n) = \boldsymbol{P}(n-m)\boldsymbol{P}(m) = \boldsymbol{P}(n-1)\boldsymbol{P} = \boldsymbol{P}^n$$

Discrete Markov Chain

$\{\bar{p}_j\}_{j \in S}$ is a stationary distribution if

$$p_j(n+1) = \sum_{i \in S} p_i(n) p_{ij},\ j = 1,2,\ldots$$

Hidden Markov Chain



MCMC

probability density $f(x)$

integral we want to estimate $I = \int h(x) f(x) dx$

We generate $X_1, X_2, \ldots,$ **with a stationary distribution** $f(x)$

$$\frac{1}{N} \sum_{i=1}^{N} h(X_i) \xrightarrow{\mathrm{P}} \mathbb{E}_f(h(X)) = I. \quad \textbf{(*)}$$

Metropolis-Hastings algorithm

We've already generated $X_0, X_1, \ldots, X_i$

We want to generate X_{i+1}

Step 1. Generate a candidate $Y \sim q(y|X_i)$.

Step 2. Calculate $r \equiv r(X_i, Y)$

$$r(x,y) = \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)},\ 1 \right\}$$

Step 3. $$X_{i+1} = \begin{cases} Y & \text{with probability } r \\ X_i & \text{with probability } 1-r. \end{cases}$$

***X_i* defined in such a way is obviously a Markov Chain**

Metropolis-Hastings algorithm

Cauchy $$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

$$r(x,y) = \min\left\{ \frac{f(y)}{f(x)},\ 1 \right\} = \min\left\{ \frac{1+x^2}{1+y^2},\ 1 \right\}.$$

Proposal density $$q(y|x) = N(x, b^2)$$

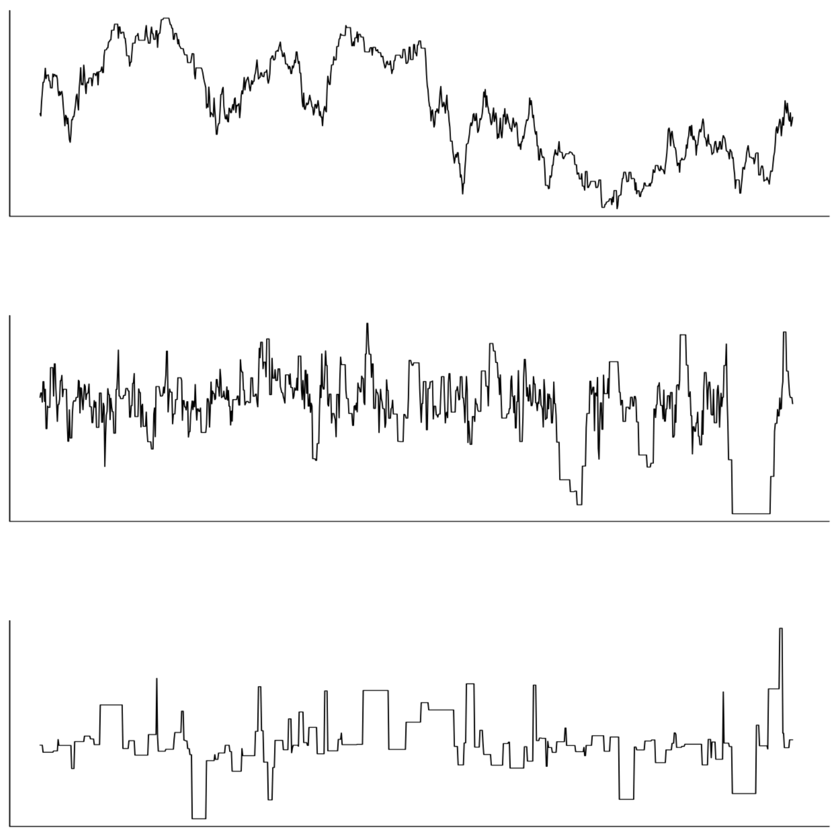
$$X_{i+1} = \begin{cases} Y & \text{with probability } r(X_i, Y) \\ X_i & \text{with probability } 1 - r(X_i, Y). \end{cases}$$

$$Y \sim N(X_i, b^2)$$

Metropolis-Hastings algorithm

$Y \sim N(X_i, b^2)$

$N = 1,000$ using $b = .1$, $b = 1$ and $b = 10$.



Metropolis-Hastings algorithm

In order to prove that the convergence (*) holds we

- ✓ **should prove that *f(x)* is a stationary distribution of the Markov Chain, defined by the Metropolis-Hastings algorithm**

- ✓ **imposing some additional requirements on *q(x|y)* and *f(x)* using a general theory we can get that the distribution of this Markov Chain converges to the stationary one (i.e. to *f(x)*), so (*) holds**

Metropolis-Hastings algorithm

Let us denote by $p(x, y)$ **a probability to jump from *x* to *y*, i.e. this is a transition density with *x* as a starting point**

f(x) **is stationary if**

$$f(x) = \int f(y)p(y,x)dy$$

We can prove that the following condition is the same as stationarity

$$f(x)p(x,y) = f(y)p(y,x).$$

In fact

$$\int f(y)p(y,x)dy = \int f(x)p(x,y)dy = f(x)\int p(x,y)dy = f(x)$$

Metropolis-Hastings algorithm

Without loss of generality we can assume that

$$f(x)q(y|x) > f(y)q(x|y).$$

$$r(x,y) = \frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}$$

$$p(x,y) = q(y|x)r(x,y) = q(y|x)\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)} = \frac{f(y)}{f(x)}q(x|y)$$

Thus

$$f(x)p(x,y) = f(y)q(x|y).$$

Another case is proved in the same way

Metropolis-Hastings algorithm

Thus

$$f(x)p(x,y) = f(y)q(x|y).$$

On the other hand $p(y,x)$ **is a probability to jump from *y* to *x*, i.e. this is a transition density with *y* as a starting point**

This requires two things: (i) the proposal distribution must generate x, and (ii) you must accept *x*

This occurs with probability

$$p(y,x) = q(x|y)r(y,x) = q(x|y).$$

Thus

$$f(y)p(y,x) = f(y)q(x|y).$$

Another case is proved in the same way