Monte Carlo Methods

Spring Semester 2013/14, Department of Applied Mathematics, University of Crete

Instructor: Harmandaris Vagelis, email: vagelis@tem.uoc.gr

Part IV: Markov Chains Monte Carlo

Markov Chains

❑ **Basic Definitions:**

- A Markov chain is a sequence of random variables that can be thought of as evolving over time
- The probability of a transition depends on the particular set the chain is in.
- We define a Markov chain by its transition kernel
  - When $\mathcal{X}$ is *discrete*, the transition kernel simply is a (transition matrix K with elements

$$P_{xy} = P(X_n = y | X_{n-1} = x)\ , \qquad x, y \in \mathcal{X}.$$

  - In the continuous case, the *kernel* also denotes the conditional density $K(x, x')$ $P(X \in A|x) = \int_A K(x, x')dx' = \int_A f(x'|x)dx'$.

Markov Chains

❑ **Essentials of Markov Chains Monte Carlo (MCMC) algorithms:**

- In the setup of MCMC algorithms, Markov chains are constructed from a *transition kernel* K, a conditional probability density

$$X_{n+1} \sim K(X_n, X_{n+1}).$$

- An example is a random walk

$$X_{n+1} = X_n + \epsilon_n$$

where ϵ_n is generated independently of $X_n, X_{n-1}, \ldots$.

- If ϵ_n is symmetric about zero, the sequence is called a *symmetric random walk*

Markov Chains

❑ **Essentials of Markov Chains Monte Carlo (MCMC) algorithms:**

- An irreducible, positive recurrent Markov chain is ergodic, that is, it converges.
- In a simulation setup, a consequence of this convergence property is that the average

$$\frac{1}{N}\sum_{n=1}^{N} h(X_n) \to \mathrm{E}_\pi[h(X)]$$

almost surely.

- Under a slightly stronger assumption a Central Limit Theorem also holds for this average

Metropolis – Hastings Algorithm

❑ **The MCMC principle:**

- It is not necessary to directly simulate from f to calculate

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

- Now we obtain
  - $X_1, \ldots, X_n \sim$ approx f without simulating from f
  - Use an ergodic Markov Chain

Working Principle of MCMC Algorithms

- For an arbitrary starting value $x^{(0)}$, a chain $(X^{(t)})$ is generated using a transition kernel with stationary distribution f
- This ensures the convergence in distribution of $(X^{(t)})$ to a random variable from f
- Given that the chain is ergodic, the starting value $x^{(0)}$ is, in principle, unimportant.

Definition A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose stationary distribution is f.

Metropolis – Hastings Algorithm

❑ **Requirements:**

- The algorithm starts with and target density f
- A candidate density $q(y|x)$
- The ratio

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

must be known up to a constant.

Metropolis – Hastings Algorithm

❑ **The Algorithm:**

- The Metropolis–Hastings algorithm associated with the objective (target) density f and the conditional density q produces a Markov chain $(X^{(t)})$ through the following transition:

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \texttt{with probability } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \texttt{with probability } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

where

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

Metropolis – Hastings Algorithm

❑ **Properties:**

- This algorithm always accepts values y_t such that the ratio $f(y_t)/q(y_t|x^{(t)})$ is increased
- It may accept values y_t such that the ratio is decreased, similar to stochastic optimization
- Like the Accept–Reject method, the Metropolis–Hastings algorithm only depends on the ratios

$$f(y_t)/f(x^{(t)}) \qquad \text{and} \qquad q(x^{(t)}|y_t)/q(y_t|x^{(t)})$$

and is, therefore, independent of normalizing constants

- Under mild conditions, MH is a reversible, ergodic Markov Chain, hence it converges
- ○ The empirical sums $\frac{1}{M}\sum h(X_i)$ converge
  ○ The CLT is satisfied

Random Walk Metropolis

- Take into account the value previously simulated to generate the following value
- This idea is already used in algorithms such as the simulated annealing
  - Since the candidate g in the MH algorithm is allowed to depend on the current state $X^{(t)}$, a first choice to consider is to simulate Y_t according to
  $$Y_t = X^{(t)} + \varepsilon_t,$$
  where ε_t is a random perturbation with distribution g, independent of $X^{(t)}$.
  - $q(y|x)$ is now of the form $g(y - x)$
  - The Markov chain associated with q is a random walk

Random Walk Metropolis

- The choice of a *symmetric function* g (that is, such that $g(-t) = g(t)$), leads to the following random walk MH algorithm

  Given $x^{(t)}$,

  (a) Generate $Y_t \sim g(|y - x^{(t)}|)$.

  (b) Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \min\left\{1, \dfrac{f(Y_t)}{f(x^{(t)})}\right\} \\ x^{(t)} & \text{otherwise.} \end{cases}$$

Random Walk Metropolis

- Hastings (1970) considers the generation of the normal distribution $\mathcal{N}(0,1)$ based on the uniform distribution on $[-\delta, \delta]$
- The algorithm: At time t

(a) Generate $Y = X_t + U$

(b)

$$\rho = \min\left\{e^{-.5(Y^2 - X_t^2)}, 1\right\}$$

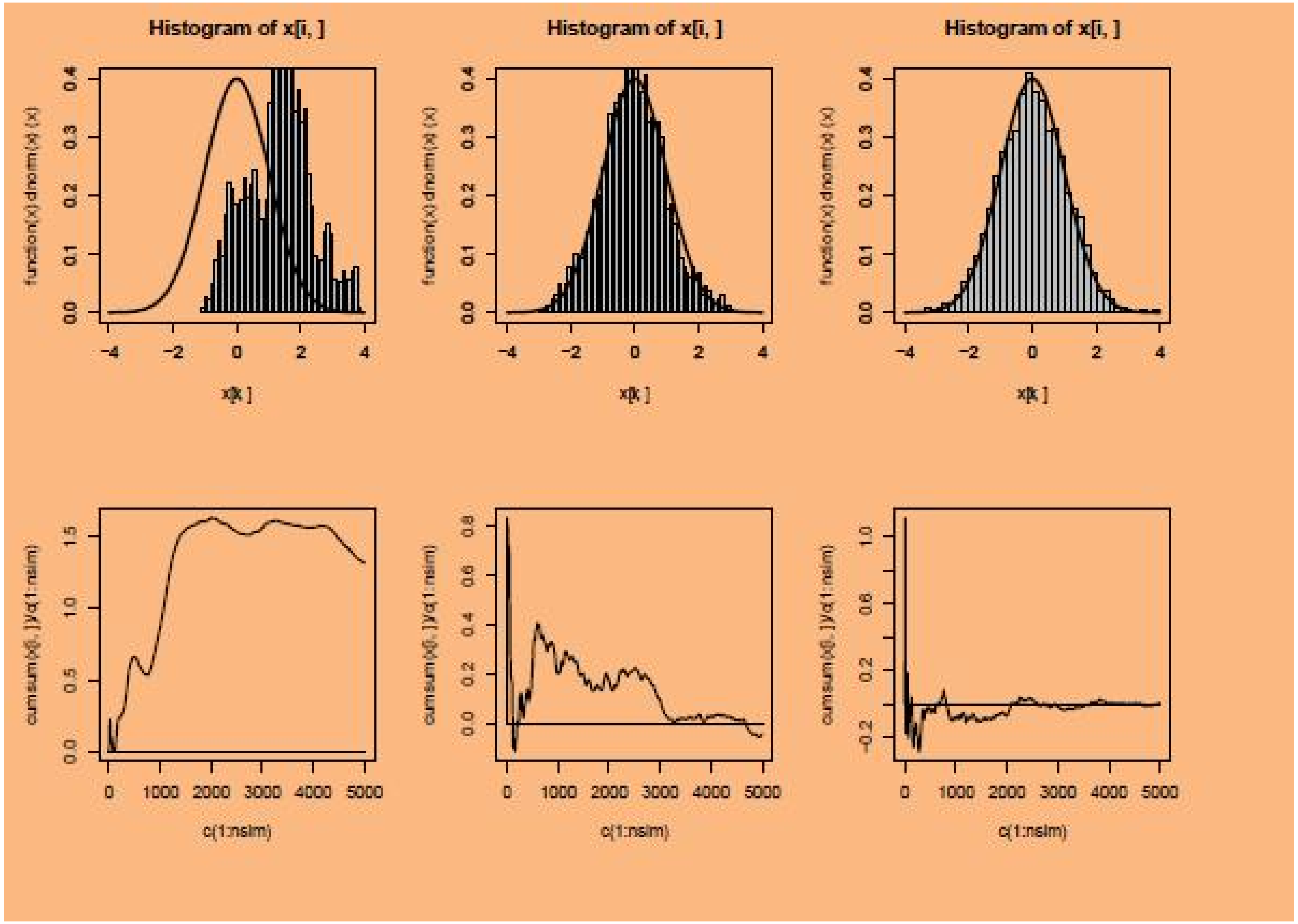
(c)

$$X_{t+1} = \begin{cases} Y & \texttt{with probability } \rho \\ X_t & \texttt{otherwise} \end{cases}$$

- Three samples of $20,000$ points produced by this method for $\delta = 0.1, 0.5$, and 1.

Random Walk Metropolis

- Histograms - Estimators:



Gibbs Sampler

❑ **Basic Idea:**

Consider a probability distribution with density $f(x_1, ..., x_p)$, for some $p > 1$.

Algorithm 4.1: (Systematic scan) Gibbs sampler

Starting with $(X_1^{(0)}, \dots, X_p^{(0)})$ iterate for $t = 1, 2, \dots$

1. Draw $X_1^{(t)} \sim f_{X_1|X_{-1}}(\cdot|X_2^{(t-1)}, \dots, X_p^{(t-1)})$.

...

j. Draw $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t)}, \dots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \dots, X_p^{(t-1)})$.

...

p. Draw $X_p^{(t)} \sim f_{X_p|X_{-p}}(\cdot|X_1^{(t)}, \dots, X_{p-1}^{(t)})$.

Gibbs Sampler

- **Random Gibbs Sampler:**

Algorithm 4.2: Random scan Gibbs sampler

Starting with $(X_1^{(0)}, \ldots, X_p^{(0)})$ iterate for $t = 1, 2, \ldots$

1. Draw an index j from a distribution on $\{1, \ldots, p\}$ (e.g. uniform)
2. Draw $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})$, and set $X_\iota^{(t)} := X_\iota^{(t-1)}$ for all $\iota \neq j$.

Gibbs Sampler

- **Important Aspects:**

- Only the so-called *full-conditional* distributions $X_i|X_{-i}$ are used in the Gibbs sampler.
  - Do the full conditionals fully specify the joint distribution?
- The sequence $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots)$ is a Markov chain.
  - Is the target distribution $f(x_1, \ldots, x_p)$ the invariant distribution of this Markov chain?
  - Will the Markov chain converge to this distribution?
  - If so, what can we use for inference: the whole chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(T)})$ or only the last value $\mathbf{X}^{(T)}$?

Gibbs Sampler

❑ **General Comments:**

❖ One special case of Metropolis-Hastings is very popular and does not require any choice of step size.

❖ Gibbs sampling is the composition of a sequence of M-H transition operators, each of which acts upon a single component of the state space.

❖ By themselves, these operators are not ergodic, but in aggregate they typically are.

❖ Most commonly, the proposal distribution is taken to be the conditional distribution, given the rest of the state. This causes the acceptance ratio to always be one and is often easy because it is low-dimensional.

Bibliography

- *Monte Carlo Strategies in Scientific Computing*, J. Liu, Springer, New York, 2001.
- *Monte Carlo Statistical Methods*, C. Robert, G. Casella, Springer, New York, 2004.
- *Stochastic Methods: A Handbook for the Natural and Social Sciences,* C. Gardiner, Springer, New York, 2009.