Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

Department of mathematics
KTH Royal Institute of Technology
johawes@kth.se

Lecture 8
Markov chain Monte Carlo I
20 April 2017

Plan of today's lecture

Outline

Last time: SIS + ISR

- A simple—but revolutionary!—idea: duplicate/kill particles with large/small weights! (Gordon *et al.*, 1993)
- The most natural approach to such selection is to simply draw new particles $(\tilde{X}^i_{0:n})_{i=1}^N$ among the SIS-produced particles $(X^i_{0:n})_{i=1}^N$ with probabilities given by the normalized importance weights.
- Formally, this amounts to set, for $i \leftarrow 1, 2, \ldots, N$,

$$\tilde{X}^i_{0:n} = X^j_{0:n} \text{ w. pr. } \frac{\omega^j_n}{\sum_{\ell=1}^N \omega^\ell_n}.$$

Last time: SIS + ISR

- After this, the resampled particles $(\tilde{X}^i_{0:n})_{i=1}^N$ are assigned equal weights $\tilde{\omega}^i_n = 1$ and we replace

$$\sum_{i=1}^N \frac{\omega^i_n}{\sum_{\ell=1}^N \omega^\ell_n} \phi(X^i_{0:n}) \quad \text{by} \quad \frac{1}{N} \sum_{i=1}^N \phi(\tilde{X}^i_{0:n}).$$

- Multinomial resampling does not add bias:

Corollary

For all $N \geq 1$ and $n \geq 0$,

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^N \phi(\tilde{X}^i_{0:n})\right) = \mathbb{E}\left(\sum_{i=1}^N \frac{\omega^i_n}{\sum_{\ell=1}^N \omega^\ell_n}\phi(X^i_{0:n})\right).$$

Last time: ... = SIS with resampling (SISR)

- After selection, we proceed with standard SIS and move the selected particles $(\tilde{X}^i_{0:n})_{i=1}^N$ according to $g_n(x_{n+1} \mid x_{0:n})$.
- The full scheme goes as follows. Given $(X^i_{0:n}, \omega^i_n)_{i=1}^N$,
  1. (selection) draw, with replacement, $(\tilde{X}^i_{0:n})_{i=1}^N$ among $(X^i_{0:n})_{i=1}^N$ according to probabilities $(\omega^i_n / \sum_{\ell=1}^N \omega^\ell_n)_{i=1}^N$
  2. (mutation) draw, for all i, $X^i_{n+1} \sim g_n(x_{n+1} \mid \tilde{X}^i_{0:n})$,
  3. set, for all i, $X^i_{0:n+1} = (\tilde{X}^i_{0:n}, X^i_{n+1})$, and
  4. set, for all i,

$$\omega^i_{n+1} = \frac{z_{n+1}(X^i_{0:n+1})}{z_n(X^i_{0:n}) g_n(X^i_{n+1} \mid X^i_{0:n})}.$$

Linear/Gaussian HMM, SISR implementation (cont'd)

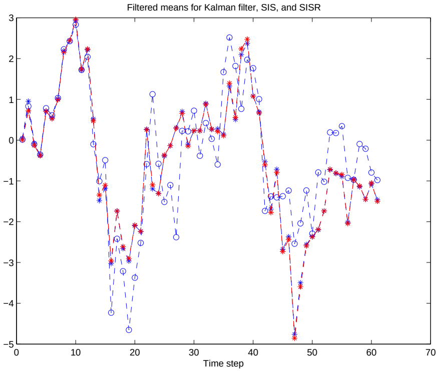

Figure: Comparison of SIS (○) and SISR (∗, blue) with exact values (∗, red) provided by the Kalman filter.

Outline

Outline

Markov Chain Monte Carlo (MCMC)

- Basic idea: To sample from a density f we construct a Markov chain having f as stationary distribution. A law of large numbers for Markov chains guarantees convergence.
- If f is complicated and/or defined on a space of high dimension this is often much easier than transformation-based methods or rejection sampling.
- The samples will however not be independent.

Markov Chain Monte Carlo (MCMC)

- Basic idea: To sample from a density f we construct a Markov chain having f as stationary distribution. A law of large numbers for Markov chains guarantees convergence.
- If f is complicated and/or defined on a space of high dimension this is often much easier than transformation-based methods or rejection sampling.
- The samples will however not be independent.

Markov Chain Monte Carlo (MCMC)

- Basic idea: To sample from a density f we construct a Markov chain having f as stationary distribution. A law of large numbers for Markov chains guarantees convergence.
- If f is complicated and/or defined on a space of high dimension this is often much easier than transformation-based methods or rejection sampling.
- The samples will however not be independent.

MCMC (cont.)

- MCMC is currently the most common method for sampling from complicated and/or high dimensional distributions.
- Dates back to the 1950's with two key papers being
  - *Equations of state calculations by fast computing machines* (Metropolis *et al.*, 1953) and
  - *Monte Carlo sampling methods using Markov chains and their applications* (Hastings, 1970).

Outline

Prelude: Markov chains

- Recall that a Markov chain on $\mathsf{X} \subseteq \mathbb{R}^d$ is a stochastic process $(X_k)_{k \geq 0}$ taking values in X such that

$$\mathbb{P}(X_{k+1} \in \mathsf{A} \mid X_0, X_1, \ldots, X_k) = \mathbb{P}(X_{k+1} \in \mathsf{A} \mid X_k)$$

  for all $\mathsf{A} \subseteq \mathsf{X}$. We call the chain time homogeneous if the conditional distribution of X_{k+1} given X_k does not depend on k.
- The distribution of X_{k+1} given $X_k = x$ determines completely the dynamics of the process, and the density q of this distribution is called the transition density of $(X_k)_{k \geq 0}$. Consequently,

$$\mathbb{P}(X_{k+1} \in \mathsf{A} \mid X_k = x_k) = \int_{\mathsf{A}} q(x_{k+1} \mid x_k) \, dx_{k+1}.$$

Prelude: Markov chains

- Recall that a Markov chain on $\mathsf{X} \subseteq \mathbb{R}^d$ is a stochastic process $(X_k)_{k \geq 0}$ taking values in X such that

$$\mathbb{P}(X_{k+1} \in \mathsf{A} \mid X_0, X_1, \ldots, X_k) = \mathbb{P}(X_{k+1} \in \mathsf{A} \mid X_k)$$

  for all $\mathsf{A} \subseteq \mathsf{X}$. We call the chain time homogeneous if the conditional distribution of X_{k+1} given X_k does not depend on k.
- The distribution of X_{k+1} given $X_k = x$ determines completely the dynamics of the process, and the density q of this distribution is called the transition density of $(X_k)_{k \geq 0}$. Consequently,

$$\mathbb{P}(X_{k+1} \in \mathsf{A} \mid X_k = x_k) = \int_{\mathsf{A}} q(x_{k+1} \mid x_k) \, dx_{k+1}.$$

Markov chains (cont.)

- Let $f_n(x_0, x_1, \ldots, x_n)$ be the joint density of $X_0, X_1, \ldots, X_n$.

Theorem

Let $(X_k)_{k \geq 0}$ be Markov with initial distribution χ and transition density q. Then

(i) $$f_n(x_0, x_1, \ldots, x_n) = \chi(x_0) \prod_{k=0}^{n-1} q(x_{k+1} \mid x_k) \quad (n \geq 1),$$

(ii) $$f_n(x_n \mid x_0) = \int \cdots \int \prod_{k=0}^{n-1} q(x_{k+1} \mid x_k) \, dx_1 \cdots dx_{n-1} \quad (n > 1).$$

- Equation (ii) is referred to as the Chapman-Kolmogorov equation.

Stationary Markov chains

- A distribution π on X is said to be stationary if

$$\int q(x \mid z)\pi(z)dz = \pi(x) \quad \text{(global balance)}.$$

- If $\chi = \pi$ it holds that

$$f_1(x_1) = \int q(x_1 \mid x_0)\chi(x_0)dx_0 = \int q(x_1 \mid x_0)\pi(x_0)dx_0 = \pi(x_1)$$

$$\Rightarrow f_2(x_2) = \int q(x_2 \mid x_1)f_1(x_1)dx_1 = \int q(x_2 \mid x_1)\pi(x_1)dx_1 = \pi(x_2)$$

$$\Rightarrow \ldots \Rightarrow f(x_n) = \pi(x_n), \quad \forall n.$$

- Thus, if starting in π, the chain will always stay in π. In this case we call also the chain stationary.

Stationary Markov chains

- A distribution π on X is said to be stationary if

$$\int q(x \mid z)\pi(z)dz = \pi(x) \quad \text{(global balance)}.$$

- If $\chi = \pi$ it holds that

$$f_1(x_1) = \int q(x_1 \mid x_0)\chi(x_0)dx_0 = \int q(x_1 \mid x_0)\pi(x_0)dx_0 = \pi(x_1)$$

$$\Rightarrow f_2(x_2) = \int q(x_2 \mid x_1)f_1(x_1)dx_1 = \int q(x_2 \mid x_1)\pi(x_1)dx_1 = \pi(x_2)$$

$$\Rightarrow \ldots \Rightarrow f(x_n) = \pi(x_n), \quad \forall n.$$

- Thus, if starting in π, the chain will always stay in π. In this case we call also the chain stationary.

Detailed balance

- Let $(X_k)_{k\geq 0}$ have transition density q and let λ be a distribution satisfying the detailed balance condition

$$\lambda(x)q(z \mid x) = \lambda(z)q(x \mid z), \quad \forall x, z \in \mathsf{X}.$$

Interpretation:

$$\text{"probability flow" } x \to z = \text{"probability flow" } z \to x.$$

- The following holds:

Theorem

Assume that λ satisfies detailed balance. Then λ is a stationary distribution.

The converse is not true.

Detailed balance

- Let $(X_k)_{k \geq 0}$ have transition density q and let λ be a distribution satisfying the detailed balance condition

  $$\lambda(x)q(z \mid x) = \lambda(z)q(x \mid z), \quad \forall x, z \in \mathsf{X}.$$

  Interpretation:

  $$\text{“probability flow” } x \to z = \text{“probability flow” } z \to x.$$

- The following holds:

Theorem

Assume that λ satisfies detailed balance. Then λ is a stationary distribution.

The converse is not true.

Ergodic Markov chains

- The following definitions will be of importance for the coming developments.

Definition

A Markov chain $(X_n)_{n \geq 0}$ with stationary distribution π is called

(i) ergodic if for all initial distributions χ,

$$\sup_{A \subseteq X} |\mathbb{P}(X_n \in A) - \pi(A)| \to 0, \quad \text{as} \quad n \to \infty.$$

(ii) uniformly ergodic if there is $\rho < 1$ such that for all initial distributions χ,

$$\sup_{A \subseteq X} |\mathbb{P}(X_n \in A) - \pi(A)| \leq \rho^n.$$

Ergodic Markov chains (cont.)

- The following theorem provides geometric ergodicity under the so-called Doeblin condition $(*)$:

Theorem (uniform ergodicity)

Assume that there exists a density μ and a constant $\varepsilon > 0$ such that for all $x, z \in \mathsf{X}$,

$$q(z \mid x) \geq \varepsilon \mu(z). \quad (*)$$

Then the chain $(X_n)_{n \geq 0}$ is uniformly ergodic for

$$\rho = 1 - \varepsilon.$$

Uniformly ergodic Markov chains

- In other words, uniform ergodicity means that the chain forgets its initial distribution geometrically fast.
- The condition $(*)$ is typically satisfied when X is compact (which is e.g. the case when X is finite set); the previous result can however be established under weaker versions of the condition that hold also for non-compact state spaces.
- Uniform ergodicity implies in general that for a large class of objective functions ϕ,

$$|\mathbb{C}(\phi(X_m), \phi(X_n))| \leq C\tilde{\rho}^{|n-m|}$$

for some $\tilde{\rho} < 1$ and some constant $C > 0$ depending on ϕ.

Uniformly ergodic Markov chains

- In other words, uniform ergodicity means that the chain forgets its initial distribution geometrically fast.
- The condition $(*)$ is typically satisfied when X is compact (which is e.g. the case when X is finite set); the previous result can however be established under weaker versions of the condition that hold also for non-compact state spaces.
- Uniform ergodicity implies in general that for a large class of objective functions ϕ,

$$|\mathbb{C}(\phi(X_m), \phi(X_n))| \leq C\tilde{\rho}^{|n-m|}$$

for some $\tilde{\rho} < 1$ and some constant $C > 0$ depending on ϕ.

Uniformly ergodic Markov chains

- In other words, uniform ergodicity means that the chain forgets its initial distribution geometrically fast.
- The condition $(*)$ is typically satisfied when X is compact (which is e.g. the case when X is finite set); the previous result can however be established under weaker versions of the condition that hold also for non-compact state spaces.
- Uniform ergodicity implies in general that for a large class of objective functions ϕ,

$$|\mathbb{C}(\phi(X_m), \phi(X_n))| \leq C\tilde{\rho}^{|n-m|}$$

for some $\tilde{\rho} < 1$ and some constant $C > 0$ depending on ϕ.

A coupling-based proof

- Define the transition density

$$\tilde{q}(x_{k+1} \mid x_k) = \frac{q(x_{k+1} \mid x_k) - \varepsilon\mu(x_{k+1})}{1 - \varepsilon} \quad (\geq 0 \text{ by } (*))$$

  and let χ and χ' be two initial distributions.
- Define two new Markov chains $(X_k)_{k\geq 0}$ and $(X'_k)_{k\geq 0}$ as follows:
  - Draw $X_0 \sim \chi$ and $X'_0 \sim \chi'$.
  - given X_k and X'_k, toss an ε-coin. If
    - (i) head (w. pr. ε), draw $X_{k+1} \sim \mu(x_{k+1})$ and set $X'_{k+1} = X_{k+1}$ ($\Rightarrow$ *coupling*).
    - (ii) tail (w. pr. $1 - \varepsilon$), draw $X_{k+1} \sim \tilde{q}(x_{k+1} \mid X_k)$. In addition, draw independently $X'_{k+1} \sim \tilde{q}(x_{k+1} \mid X'_k)$; however, if the chains have coupled earlier, keep $X'_{k+1} = X_{k+1}$.

A coupling-based proof

- Define the transition density

$$\tilde{q}(x_{k+1} \mid x_k) = \frac{q(x_{k+1} \mid x_k) - \varepsilon\mu(x_{k+1})}{1 - \varepsilon} \quad (\geq 0 \text{ by } (*))$$

  and let χ and χ' be two initial distributions.
- Define two new Markov chains $(X_k)_{k \geq 0}$ and $(X'_k)_{k \geq 0}$ as follows:
  - Draw $X_0 \sim \chi$ and $X'_0 \sim \chi'$.
  - given X_k and X'_k, toss an ε-coin. If
    - (i) head (w. pr. ε), draw $X_{k+1} \sim \mu(x_{k+1})$ and set $X'_{k+1} = X_{k+1}$ ($\Rightarrow$ *coupling*).
    - (ii) tail (w. pr. $1 - \varepsilon$), draw $X_{k+1} \sim \tilde{q}(x_{k+1} \mid X_k)$. In addition, draw independently $X'_{k+1} \sim \tilde{q}(x_{k+1} \mid X'_k)$; however, if the chains have coupled earlier, keep $X'_{k+1} = X_{k+1}$.

Example: a chain on a discrete set

- Let $X = \{1, 2, 3\}$ and

$$\begin{pmatrix} q(1|1) = 0.4 & q(2|1) = 0.4 & q(3|1) = 0.2 \\ q(1|2) = 0 & q(2|2) = 0.7 & q(3|2) = 0.3 \\ q(1|3) = 0 & q(2|3) = 0.1 & q(3|3) = 0.9 \end{pmatrix}.$$

- This chain has $\pi = (0, 0.25, 0.75)$ as stationary distribution (check global balance).
- Moreover, the chain satisfies $(*)$ with

$$\varepsilon = 0.2 \quad \text{and} \quad \mu = (0, 0.5, 0.5).$$

It is thus uniformly ergodic.

Example: a chain on a discrete set (cont.)

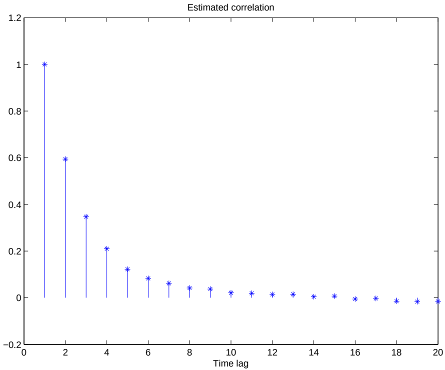

Figure: Estimated correlation obtained by simulating the chain 1000 time steps.

A law of large numbers for Markov chains

- In the case where the states of (X_k) are only weakly dependent there is, just like in the case of independent variables, an LLN:

Theorem (law of large numbers for Markov chains)

Let $(X_n)_{n \geq 0}$ be a stationary Markov chain (with stationary distribution π) and ϕ a function s.t.

$$\mathbb{C}(\phi(X_0), \phi(X_n)) \to 0 \quad \text{as } n \to \infty.$$

Then

$$\frac{1}{n} \sum_{k=1}^{n} \phi(X_k) \overset{\mathbb{P}}{\to} \int \phi(x)\pi(x)\, dx \quad \text{as} \quad n \to \infty.$$

A law of large numbers for Markov chains (cont.)

- Note that $\int \phi(x)\pi(x)\, dx$ is the mean of $\phi(X_n)$ under π.
- In particular, uniformly ergodic Markov chains satisfy the condition of the LLN.
- The assumption that the chain is initialized in the stationary distribution can, by assuming ergodicity, be removed straightforwardly.
- There are stronger versions of the previous LLN, e.g. for convergence with probability one ("almost sure convergence").

Example: a chain on a discrete set reconsidered

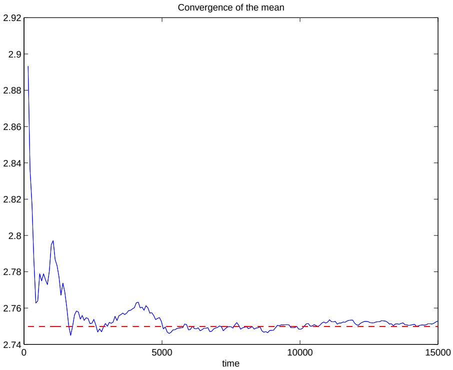

Figure: Plot of means $\frac{1}{n}\sum_{k=1}^{n} X_k$ with increasing n. Here the mean of the stationary distribution is $1 \cdot 0 + 2 \cdot 0.25 + 3 \cdot 0.75 = 2.75$ (red line).

Outline

Next week

- Now we have gained enough understanding of Markov chains to be able to understand MCMC in some detail.
- Thus, tomorrow and next week we will deal with the main objective of MCMC, namely how to, given a density f, construct a Markov chain $(X_k)_{k \geq 0}$ having f as stationary distribution.
- Focus will be set on
  - the Metropolis-Hastings algorithm and
  - the Gibbs sampler.
- We will also work out a full example of an implementation.

Next week

- Now we have gained enough understanding of Markov chains to be able to understand MCMC in some detail.
- Thus, tomorrow and next week we will deal with the main objective of MCMC, namely how to, given a density f, construct a Markov chain $(X_k)_{k \geq 0}$ having f as stationary distribution.
- Focus will be set on
  - the Metropolis-Hastings algorithm and
  - the Gibbs sampler.
- We will also work out a full example of an implementation.

Next week

- Now we have gained enough understanding of Markov chains to be able to understand MCMC in some detail.
- Thus, tomorrow and next week we will deal with the main objective of MCMC, namely how to, given a density f, construct a Markov chain $(X_k)_{k \geq 0}$ having f as stationary distribution.
- Focus will be set on
  - the Metropolis-Hastings algorithm and
  - the Gibbs sampler.
- We will also work out a full example of an implementation.

Next week

- Now we have gained enough understanding of Markov chains to be able to understand MCMC in some detail.
- Thus, tomorrow and next week we will deal with the main objective of MCMC, namely how to, given a density f, construct a Markov chain $(X_k)_{k \geq 0}$ having f as stationary distribution.
- Focus will be set on
  - the Metropolis-Hastings algorithm and
  - the Gibbs sampler.
- We will also work out a full example of an implementation.