Monte Carlo and Empirical Methods for Stochastic Inference (MASM11/FMSN50)

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 5
Sequential Monte Carlo methods I
February 5, 2019

Plan of today's lecture

1. Variance reduction reconsidered
2. Sequential MC problems
3. 3 Examples of SMC problems
   - Prelude: Markov chains
   - Example 1: Simulation of extreme events
   - Example 2: Estimation in general HMMs
   - Example 3: Estimation of SAWs

We are here $\longrightarrow$ •

Last time: Variance reduction

Last time we discussed how to reduce the variance of the standard MC sampler by introducing correlation between the variables of the sample. More specifically, we used

1. a control variate Y such that $\mathbb{E}(Y) = m$ is known:

$$Z = \phi(X) + \beta(Y - m),$$

   where β was tuned optimally to $\beta^* = -\mathbb{C}(\phi(X), Y)/\mathbb{V}(Y)$.

2. antithetic variables V and V' such that $\mathbb{E}(V) = \mathbb{E}(V') = \tau$ and $\mathbb{C}(V, V') < 0$:

$$W = \frac{V + V'}{2}.$$

Last time: Variance reduction

The following theorem turned out to be useful when constructing antithetic variables.

Theorem

Let $V = \varphi(U)$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is a monotone function. Moreover, assume that there exists a non-increasing transform $T : \mathbb{R} \to \mathbb{R}$ such that $U \overset{d.}{=} T(U)$. Then $V = \varphi(U)$ and $V' = \varphi(T(U))$ are identically distributed and

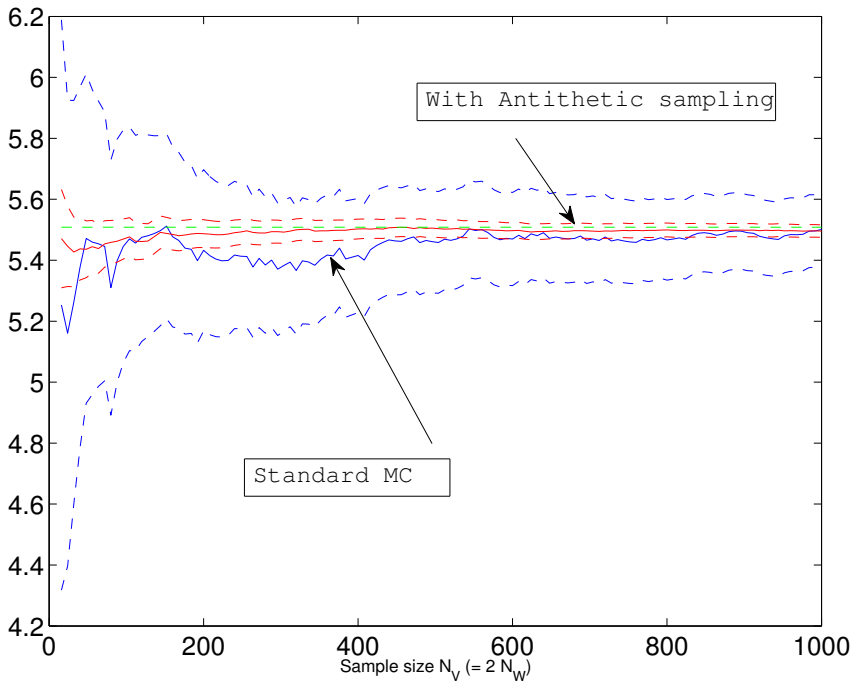
$$\mathbb{C}(V, V') = \mathbb{C}(\varphi(U), \varphi(T(U))) \leq 0.$$

An important application of this theorem is the following: Let F be a distribution function and ϕ a monotone function. Then, letting $U \sim \mathcal{U}(0,1)$, $T(u) = 1 - u$, and $\varphi(u) = \phi(F^{-1}(u))$ yields, for $V = \phi(F^{-1}(U))$ and $V' = \phi(F^{-1}(1-U))$,

$$V \overset{\text{d.}}{=} V' \quad \text{and} \quad \mathbb{C}(V, V') \leq 0.$$

Last time: Variance reduction

$$\tau = 2\int_0^{\pi/2} \exp(\cos^2(x))\,\mathrm{d}x, \qquad \begin{cases} V = 2\frac{\pi}{2}\exp(\cos^2(X)), \\ V' = 2\frac{\pi}{2}\exp(\sin^2(X)), \\ W = \frac{V+V'}{2}. \end{cases}$$



Control variates reconsidered

A problem with the control variate approach is that the optimal β, i.e.

$$\beta^* = -\frac{\mathbb{C}(\phi(X), Y)}{\mathbb{V}(Y)},$$

is generally not known explicitly. Thus, it was suggested to

1. draw $(X_i)_{i=1}^N$,
2. draw $(Y^i)_{i=1}^N$,
3. estimate, via MC, β^* using the drawn samples, and
4. use this to optimally construct $(Z^i)_{i=1}^N$.

This yields a so-called batch estimator of β^*. However, this procedure is computationally somewhat complex.

An online approach to optimal control variates

The estimators

$$C_N \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\phi(X_i)(Y^i - m)$$

$$V_N \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}(Y^i - m)^2$$

of $\mathbb{C}(\phi(X), Y)$ and $\mathbb{V}(Y)$, respectively, can be implemented recursively according to

$$C_{\ell+1} = \frac{\ell}{\ell+1}C_\ell + \frac{1}{\ell+1}\phi(X_{\ell+1})(Y^{\ell+1} - m)$$

and

$$V_{\ell+1} = \frac{\ell}{\ell+1}V_\ell + \frac{1}{\ell+1}(Y^{\ell+1} - m)^2.$$

with $C_0 = V_0 = 0$.

An online approach to optimal control variates (cont.)

Inspired by this we set for $\ell = 0, 1, 2, \ldots, N-1$,

$$\begin{aligned} Z_{\ell+1} &= \phi(X_{\ell+1}) + \beta_\ell (Y^{\ell+1} - m), \\ \tau_{\ell+1} &= \frac{\ell}{\ell+1}\tau_\ell + \frac{1}{\ell+1} Z_{\ell+1}, \end{aligned} \qquad (*)$$

where $\beta_0 \overset{\text{def}}{=} 1$, $\beta_\ell \overset{\text{def}}{=} -C_\ell / V_\ell$ for $\ell > 0$, and $\tau_0 \overset{\text{def}}{=} 0$ yielding an online estimator. One may then establish the following convergence results.

Theorem

Let τ_N be obtained through $(*)$. Then, as $N \to \infty$,

(i) $\tau_N \to \tau$ *(a.s.),*

(ii) $\sqrt{N}(\tau_N - \tau) \xrightarrow{d.} \mathcal{N}(0, \sigma_*^2)$,

where $\sigma_*^2 \overset{\text{def}}{=} \mathbb{V}(\phi(X))\{1 - \rho(\phi(X), Y)^2\}$ is the optimal variance.

Example: the tricky integral again

We estimate

$$\tau = \int_{-\pi/2}^{\pi/2} \exp(\cos^2(x))\,\mathrm{d}x \overset{sym}{=} 2\int_0^{\pi/2} \underbrace{\frac{\pi}{2}\exp(\cos^2(x))}_{=\phi(x)} \underbrace{\frac{2}{\pi}}_{=f(x)}\,\mathrm{d}x$$
$$= \mathbb{E}_f(\phi(X))$$

using

$$Z = \phi(X) + \beta^*(Y - m),$$

where $Y = \cos^2(X)$ is a control variate with

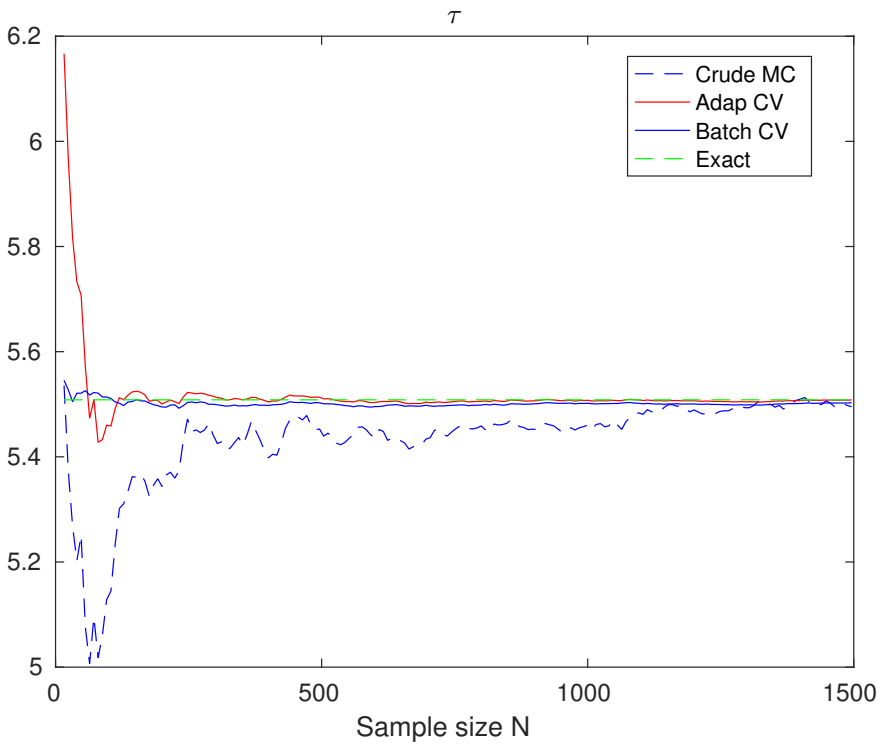
$$m = \mathbb{E}(Y) = \int_0^{\pi/2} \cos^2(x)\frac{2}{\pi}\,\mathrm{d}x = \{\text{use integration by parts}\} = \frac{1}{2}.$$

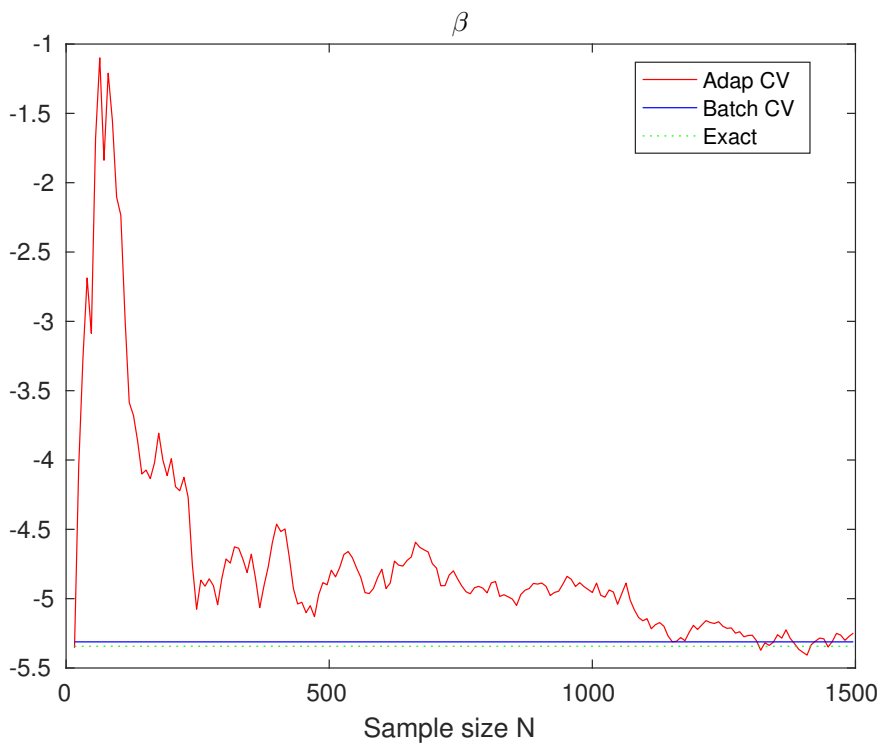
However, the optimal coefficient β^* is in general not known explicitly (tedious calculations give $\beta^* = -4\,\mathrm{e}^{\frac{1}{2}}\pi\,\mathrm{I}_1\left(\frac{1}{2}\right) \approx -5.3432$) .

Example: the tricky integral again

```
cos2 = @(x) cos(x).^2;
phi = @(x) 2(pi/2)*exp(cos2(x));
m = 1/2;
X = (pi/2)*rand;
Y = cos2(X);
c = phi(X)*(Y - m);
v = (Y - m)^2;
tau_CV = phi(X) + (Y - m);
beta = - c/v;
for k = 2:N,
    X = (pi/2)*rand;
    Y = cos2(X);
    Z = phi(X) + beta*(Y - m);
    tau_CV = (k - 1)*tau_CV/k + Z/k;
    c = (k - 1)*c/k + phi(X)*(Y - m)/k;
    v = (k - 1)*v/k + (Y - m)^2/k;
    beta = - c/v;
end
```

Example: the tricky integral again





We are here $\longrightarrow \bullet$

Sequential MC problems

We will now (and for the coming two lectures) extend the principal goal of the course to the problem of estimating sequentially sequences $(\tau_n)_{n\geq 0}$ of expectations

$$\tau_n = \mathbb{E}_{f_n}(\phi(X_{0:n})) = \int_{\mathsf{X}_n} \phi(x_{0:n}) f_n(x_{0:n}) \, \mathrm{d}x_{0:n}$$

over spaces X_n of increasing dimension, where again the densities $(f_n)_{n\geq 0}$ are known up to normalizing constants only; i.e. for every $n \geq 0$,

$$f_n(x_{0:n}) = \frac{z_n(x_{0:n})}{c_n},$$

where c_n is an unknown constant and z_n is a known positive function on X_n.

As we will see, such sequences appear in many applications in statistics and numerical analysis.

We are here $\longrightarrow$ •

We are here $\longrightarrow$ •

Prelude: Markov chains

A Markov chain on $\mathsf{X} \subseteq \mathbb{R}^d$ is a family of random variables (= stochastic process) $(X_k)_{k \geq 0}$ taking values in X such that

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

for all (measurable) $\mathsf{B} \subseteq \mathsf{X}$. 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). Consequently,

$$\mathbb{P}(X_{k+1} \in \mathsf{B} | X_k = x_k) = \int_{\mathsf{B}} q(x_{k+1} | x_k) \, \mathrm{d}x_{k+1}.$$

Markov chains (cont.)

The following theorem provides the joint density $f_n(x_0, x_1, \ldots, x_n)$ of $X_0, X_1, \ldots, X_n$.

Theorem

Let (X_k) be Markov with initial distribution χ. Then for $n > 0$,

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

Corollary (Chapman-Kolmogorov equation)

Let (X_k) be Markov. Then for $n > 1$,

$$f_n(x_n|x_0) = \int \cdots \int \left(\prod_{k=0}^{n-1} q(x_{k+1}|x_k) \right) dx_1 \cdots dx_{n-1}.$$

Example: The AR(1) process

As a first example we consider a first order autoregressive process (AR(1)) in $\mathbb{R}$. Set

$$X_0 = 0, \quad X_{k+1} = \alpha X_k + \epsilon_{k+1},$$

where α is a constant and the variables $(\epsilon_k)_{k\geq 1}$ of the noise sequence are i.i.d. with density function f. In this case,

$$\begin{aligned}\mathbb{P}(X_{k+1} \leq x_{k+1}|X_k = x_k) &= \mathbb{P}(\alpha X_k + \epsilon_{k+1} \leq x_{k+1}|X_k = x_k) \\ &= \mathbb{P}(\epsilon_{k+1} \leq x_{k+1} - \alpha x_k|X_k = x_k) = \mathbb{P}(\epsilon_{k+1} \leq x_{k+1} - \alpha x_k),\end{aligned}$$

implying that

$$\begin{aligned}q(x_{k+1}|x_k) &= \frac{\partial}{\partial x_{k+1}}\mathbb{P}(X_{k+1} \leq x_{k+1}|X_k = x_k) \\ &= \frac{\partial}{\partial x_{k+1}}\mathbb{P}(\epsilon_{k+1} \leq x_{k+1} - \alpha x_k) = f(x_{k+1} - \alpha x_k).\end{aligned}$$

We are here $\longrightarrow$ •

Simulation of rare events for Markov chains

Let (X_k) be a Markov chain on $\mathsf{X} = \mathbb{R}$ and consider the rectangle $\mathsf{B} = \mathsf{B}_0 \times \mathsf{B}_1 \times \cdots \mathsf{B}_n \subseteq \mathbb{R}^n$, where every $\mathsf{B}_\ell = (a_\ell, b_\ell)$ is an interval. Here B can be a possibly extreme event.

Say that we wish to to compute, sequentially as n increases, some expectation under the conditional distribution $f_{n|\mathsf{B}}$ of the states $X_{0:n} = (X_0, X_1, X_2, \ldots, X_n)$ given $X_{0:n} \in \mathsf{B}$, i.e.

$$\tau_n = \mathbb{E}_{f_n}(\phi(X_{0:n})|X_{0:n} \in \mathsf{B}) = \mathbb{E}_{f_{n|\mathsf{B}}}(\phi(X_{0:n}))$$

$$= \int_{\mathsf{B}} \phi(x_{0:n}) \underbrace{\frac{f(x_{0:n})}{\mathbb{P}(X_{0:n} \in \mathsf{B})}}_{=f_{n|\mathsf{B}}(x_{0:n}) = z_n(x_{0:n})/c_n} \mathrm{d}x_{0:n}.$$

Here the unknown probability $c_n = \mathbb{P}(X_{0:n} \in \mathsf{B})$ of the rare event B is often the quantity of interest.

Simulation of rare events for Markov chains (cont.)

As

$$c_n = \mathbb{P}(X_{0:n} \in \mathsf{B}) = \int \mathbb{1}_{\mathsf{B}}(x_{0:n}) f(x_{0:n}) \, \mathrm{d}x_{0:n}$$

a first—naive—approach could of course be to use standard MC and simply

1. simulate the Markov chain N times, yielding trajectories $(X_i^{0:n})_{i=1}^N$,
2. count the number N_{B} of trajectories that fall into B, and
3. estimate c_n using the MC estimator

$$c_n^N = \frac{N_{\mathsf{B}}}{N}.$$

Problem: if $c_n = 10^{-9}$ we may expect to produce a billion draws before obtaining a single draw belonging to B! As we will se, SMC methods solve the problem efficiently.

We are here $\longrightarrow$ •

Estimation in general hidden Markov models (HMMs)

A hidden Markov model (HMM) comprises two stochastic processes:

1. A Markov chain $(X_k)_{k\geq 0}$ with transition density q:

$$X_{k+1}|X_k = x_k \sim q(x_{k+1}|x_k).$$

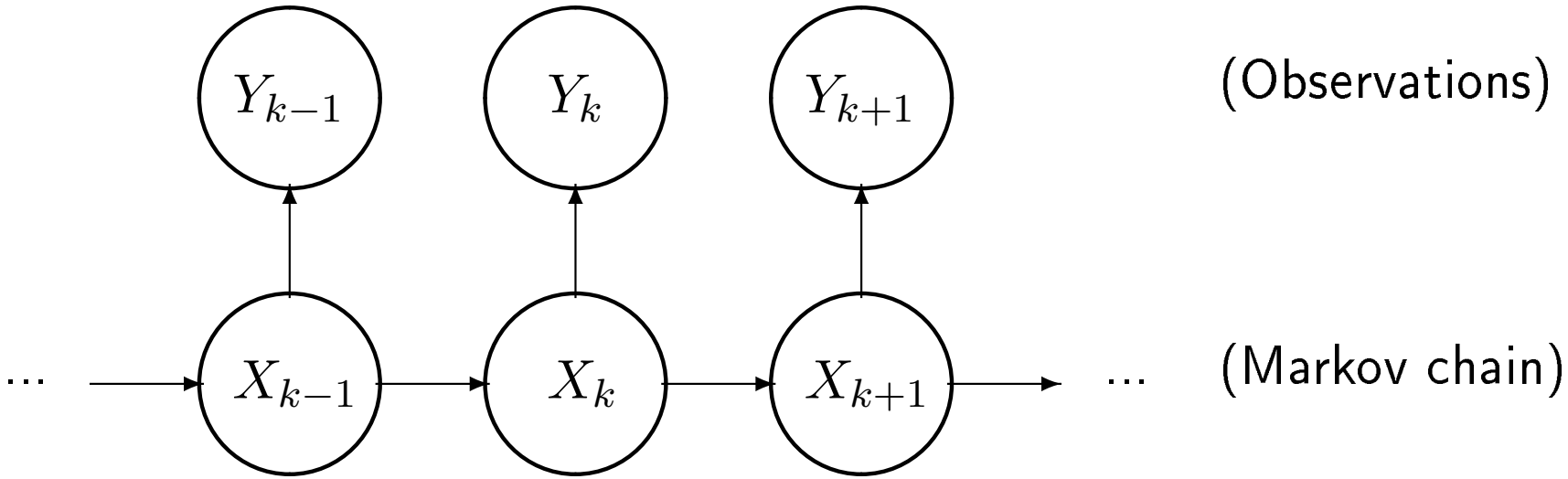
   The Markov chain is not seen by us (hidden) but partially observed through

2. an observation process $(Y_k)_{k\geq 0}$ such that conditionally on the chain $(X_k)_{k\geq 0}$,
   - (i) the Y_k's are independent with
   - (ii) conditional distribution of each Y_k depending on the corresponding X_k only.

The density of the conditional distribution $Y_k|(X_k)_{k\geq 0} \overset{\text{d.}}{=} Y_k|X_k$ will be denoted by $p(y_k|x_k)$.

Estimation in general HMMs (cont.)

Graphically:



$$Y_k|X_k = x_k \sim p(y_k|x_k) \qquad \text{(Observation density)}$$
$$X_{k+1}|X_k = x_k \sim q(x_{k+1}|x_k) \qquad \text{(Transition density)}$$
$$X_0 \sim \chi(x_0) \qquad \text{(Initial distribution)}$$

Example HMM: A stochastic volatility model

The following dynamical system is used in financial economy (see Taylor (1982)). Let

$$\begin{cases} X_{k+1} = \alpha X_k + \sigma \epsilon_{k+1}, \\ Y_k = \beta \exp\left(\frac{X_k}{2}\right) \varepsilon_k, \end{cases}$$

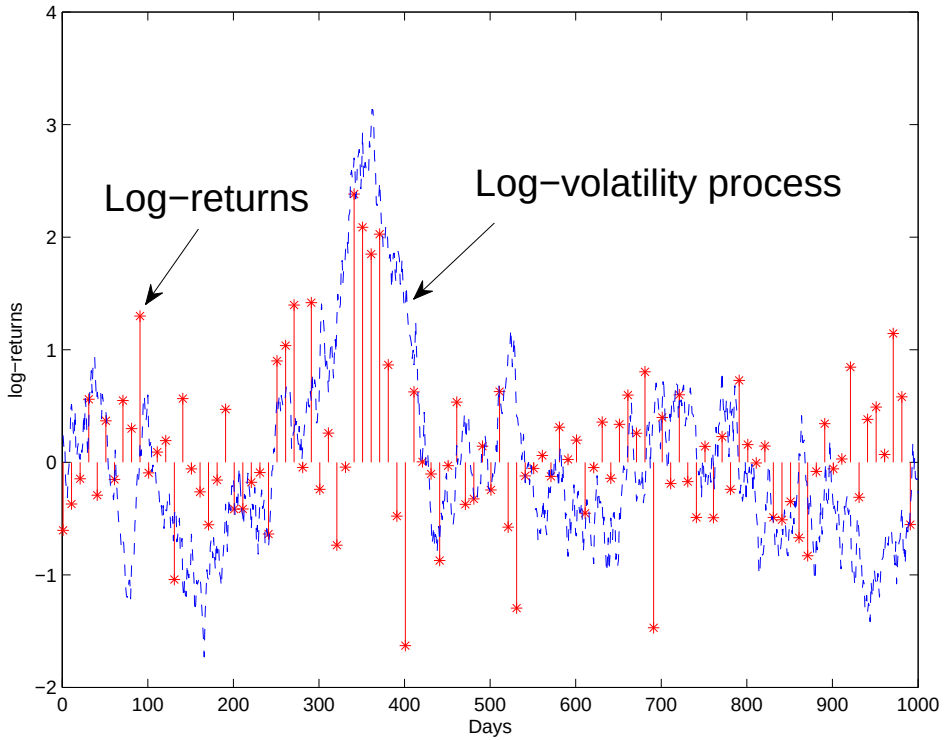
where $\alpha \in (0,1)$, $\sigma > 0$, and $\beta > 0$ are constants and $(\epsilon_k)_{k \geq 1}$ and $(\varepsilon_k)_{k \geq 0}$ are sequences of i.i.d. standard normal-distributed noise variables. In this model

- the values of the observation process (Y_k) are observed daily log-returns (from e.g. the Swedish OMXS30 index) and
- the hidden chain (X_k) is the unobserved log-volatility (modeled by a stationary AR(1) process).

The strength of this model is that it allows for volatility clustering, a phenomenon that is often observed in real financial time series.

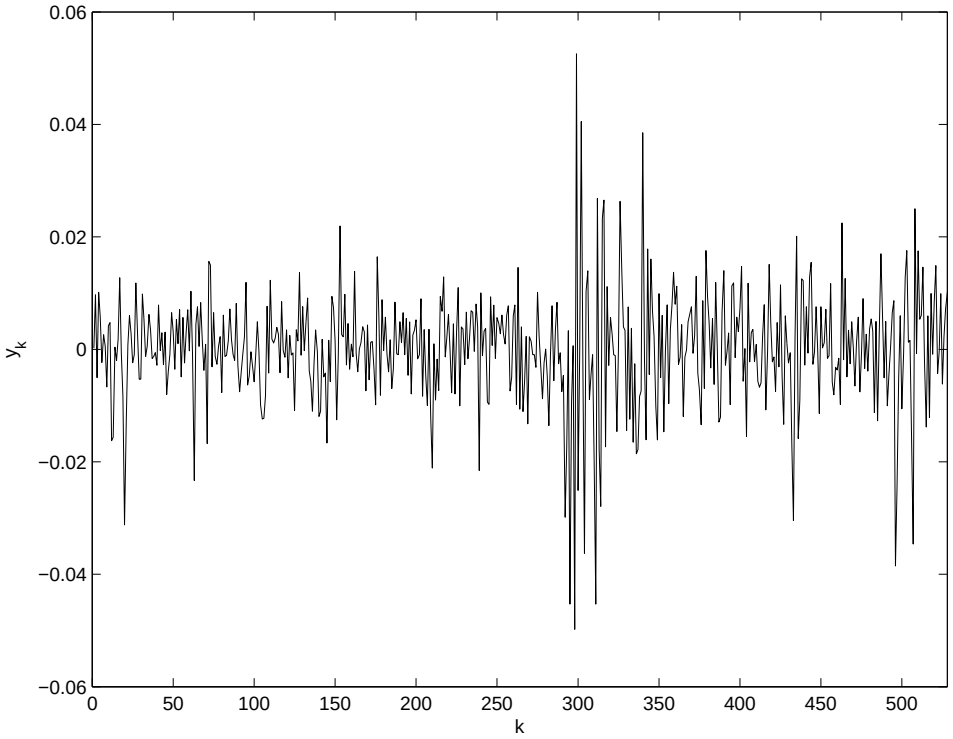
Example HMM: A stochastic volatility model

A realization of the the model looks like follows (here $\alpha = 0.975$, $\sigma = 0.16$, and $\beta = 0.63$).



Example HMM: A stochastic volatility model

Daily log-returns from the Swedish stock index OMXS30, from 2005-03-30 to 2007-03-30.



The smoothing distribution

When operating on HMMs, one is most often interested in the smoothing distribution $f_n(x_{0:n}|y_{0:n})$, i.e. the conditional distribution of a set $X_{0:n}$ of hidden states given $Y_{0:n} = y_{0:n}$.

Theorem (Smoothing distribution)

$$f_n(x_{0:n}|y_{0:n}) = \frac{\chi(x_0)p(y_0|x_0)\prod_{k=1}^{n} p(y_k|x_k)q(x_k|x_{k-1})}{L_n(y_{0:n})},$$

where $L_n(y_{0:n})$ is the likelihood function given by

$$L_n(y_{0:n}) = \text{density of the observations } y_{0:n}$$
$$= \int \cdots \int \chi(x_0)p(y_0|x_0) \prod_{k=1}^{n} p(y_k|x_k)q(x_k|x_{k-1})\, dx_0 \cdots dx_n.$$

Estimation of smoothed expectations

Being a high-dimensional (say $n \approx 1000$ or $10,000$) integral over complicated integrands, $L_n(y_{0:n})$ is in general unknown. However by writing

$$\tau_n = \mathbb{E}(\phi(X_{0:n})|Y_{0:n} = y_{0:n}) = \int \cdots \int \phi(x_{0:n}) f_n(x_{0:n}|y_{0:n}) \, \mathrm{d}x_0 \cdots \mathrm{d}x_n$$
$$= \int \cdots \int \phi(x_{0:n}) \frac{z_n(x_{0:n})}{c_n} \, \mathrm{d}x_0 \cdots \mathrm{d}x_n,$$

with

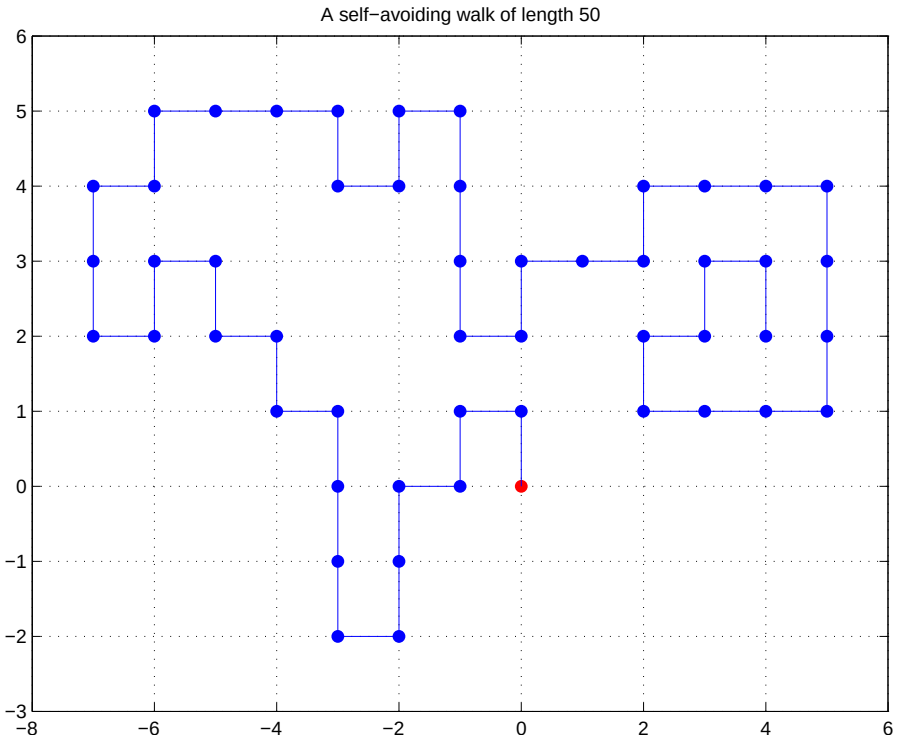
$$\begin{cases} z_n(x_{0:n}) = \chi(x_0) p(y_0|x_0) \prod_{k=1}^n p(y_k|x_k) q(x_k|x_{k-1}), \\ c_n = L_n(y_{0:n}), \end{cases}$$

we may cast the problem of computing τ_n into the framework of self-normalized IS. In particular we would like to update sequentially, in n, the approximation as new data (Y_k) appears.

We are here $\longrightarrow$ •

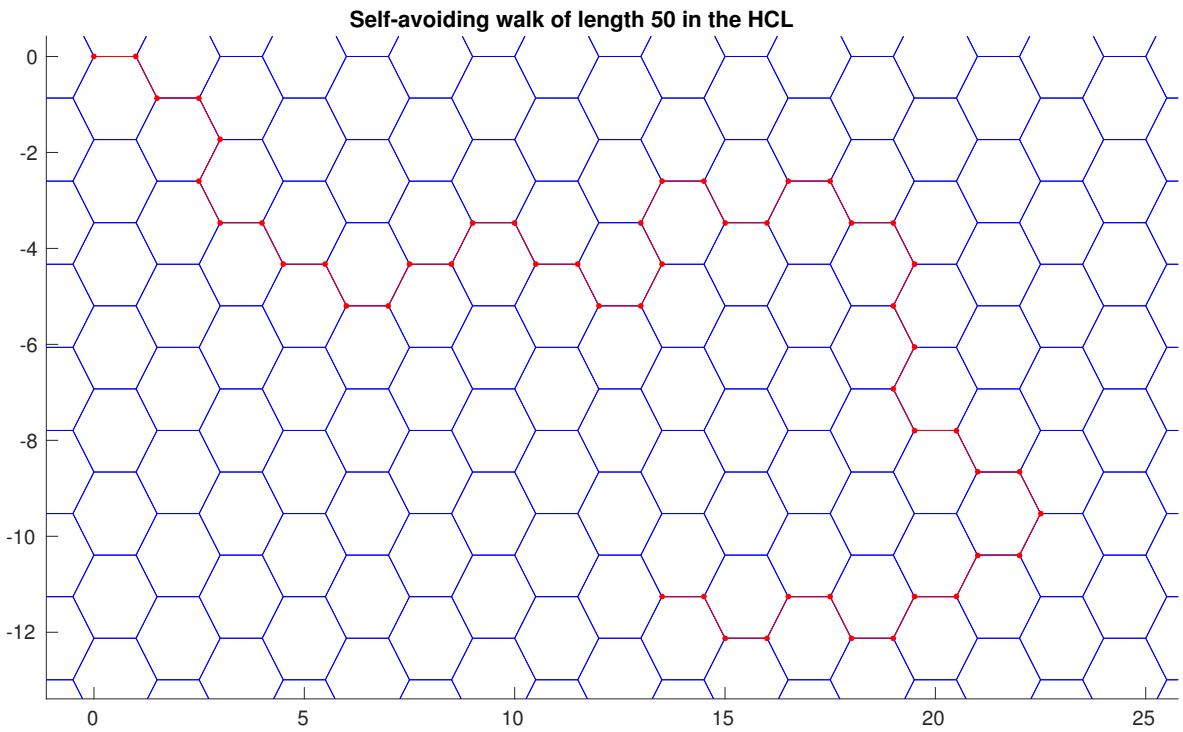
Self-avoiding walks (SAWs) in the 2-dim integer ($\mathbb{Z}^2$) lattice

Let $\mathsf{S}_n \stackrel{\text{def}}{=} \{x_{0:n} \in \mathbb{Z}^{2n} : x_0 = \mathbf{0}, |x_{k+1} - x_k| = 1, x_k \neq x_\ell, \forall 0 \leq \ell < k \leq n\}$ be the set of n-step self-avoiding walks in $\mathbb{Z}^2$.



Self-avoiding walks (SAWs) in the honeycomb lattice (HCL)

Let
$\mathsf{S}_n \overset{\text{def}}{=} \{x_{0:n} \in \text{HCL} : x_0 = \mathbf{0}, |x_{k+1} - x_k| = 1, x_k \neq x_\ell, \forall 0 \leq \ell < k \leq n\}$ be the set of n-step self-avoiding walks in HCL.



Application of SAWs

In addition, let

$$c_n = |\mathsf{S}_n| = \text{The number of possible SAWs of length } n.$$

SAWs are used in

- polymer science for describing long chain polymers, with the self-avoidance condition modeling the excluded volume effect.
- statistical mechanics and the theory of critical phenomena in equilibrium.

However, computing c_n (and in analyzing how c_n depends on n) is known to be a very challenging combinatorial problem!

An MC approach to SAWs

Trick: let $f_n(x_{0:n})$ be the uniform distribution on S_n:

$$f_n(x_{0:n}) = \frac{1}{c_n} \underbrace{\mathbb{1}_{\mathsf{S}_n}(x_{0:n})}_{=z(x_{0:n})}, \quad x_{0:n} \in \mathbb{Z}^{2n},$$

We may thus cast the problem of computing the number c_n (= the normalizing constant of f_n) into the framework of self-normalized IS based on some convenient instrumental distribution g_n on $\mathbb{Z}^{2n}$.

In addition, solving this problem for $n = 1, 2, 3, \ldots, 508, 509, \ldots$ calls for sequential implementation of IS.

This will be the topic of HA2!