(Exact and efficient) Simulation of Conditioned Markov Processes

Omiros Papaspiliopoulos
Universitat Pompeu Fabra, Barcelona
http://www.econ.upf.edu/∼omiros

GrStoc α', $\Lambda EYKA\Delta A$ 2009

Acknowledgement: K.Kalogeropoulos for his steady driving!

Context: Markov processes

In this lecture we consider time-ordered processes. Why Markov?

- Statistical: first-order approximation on a sequence of random variables which cannot be treated independent
- Mathematical: the generator (and the initial conditions) identify the process. Questions of stationarity, convergence, ergodicity, etc, can be neatly formulated and investigated.
- Computational: Simulating exactly a Markov chain path $(X_1, X_2, \dots, X_T)$ can be done by $T-1$ applications of the same algorithm: Simulate X_2 given X_1. Break of the curse of dimensionality.

Spatial processes

The mathematics and modelling are rather different when the time-ordering is not available. The Markov property can be defined on arbitrary sets equipped with a neighbourhood structure (eg lattice processes, graphical models); Nial's lecture

Local smoothing; Gaussian processes

The methods of this lecture do not apply to arbitrary neighbourhood structures. The most straightforward extension is to trees

Forward-Backward simulation

We can name the problem of simulating a Markov process forward in time (some times we will say *unconditionally*, although we might condition on the first value) as forward simulation. This is at least in principle feasible, although doing it exactly when the time parameter is the positive real line has to be carefully thought.

Nevertheless, this lecture is particularly interested in the backward simulation problem: that of simulating X given information about the future, we will call this *conditional simulation*, hence simulation of conditioned Markov processes. In the passing we will address the **exact** aspect of the forward problem

Canonical problem (to have in mind):

$$\text{simulate}\, X\, \text{given}\, X_0 = u\, \text{and}\, X_T = v \tag{1}$$

Backward simulation

Here, we are interested in simulating from conditioned Markov processes, where the type of conditioning event is such that the conditioned process is still Markov. Example: the one given above.

Another point of view: simulating from certain changes of measures from a Markov probability measure (Feynman-Kac problems).

The aim of the lecture is to address the following questions: can we simulate exactly certain classes of conditioned Markov processes, and can we do it efficiently (i.e polynomial time)? Black-box methods?

We will approach this question by considering different types of state-spaces, time-parameters and conditioning events:

- discrete-time, discrete-space: Hidden Markov Models (HMMs)
- continuous-time, discrete space: discretely observed with error Markov Jump Processes (MJPs)
- continuous-time, continuous-space: diffusion bridges

This course will give a rather complete presentation of statistical inference and simulation for HMMs; it will characterize mathematically the dynamics of conditioned MJPs; it will introduce simulation of diffusion bridges. We will also demonstrate a few simple theoretical tools (dynamic programming, generalized Bayes formula, generalized importance sampling, Girsanov's theorem, likelihood ratios for conditioned processes, retrospective sampling)

My hope is that even a completely uninterested in the topic member in the audience will find at least 2-3 ideas/methods/intuitions which are of interest

Part I: Hidden Markov Models, estimation and simulation

- Introduce a class of parameter-driven time series models
- Re-formulate as local-property-preserving change-of-measure from a Markov chain measure. The resulting structure is fundamental in applied maths by appropriately specifying the state-space, the generator and the weights. Some diverse examples.
- Powerful machinery: forward-backward recursions (decomposition of Bayes formula) inspired by dynamic programming techniques.
- Efficient conditional simulation, marginal computations

HMMs as time-series models

Let $Y_i\,, i = 1, \ldots, T$ be a series of serially-dependent observations (taking values in a measurable space).

Let $X_i\,, i = 1, \ldots, T$ be a Markov Chain (MC) with state-space $\mathcal{X} = \{1, \ldots, d\}$, transition matrix γ, and initial distribution $X_1 \sim \delta$. We assume *time-homogeneity* only for notational simplicity, and identify $\mathcal{X}$ with $\{1, \ldots, d\}$ without loss of generality. Thus,

$$
\begin{aligned}
P[X_1 = j] &= \delta_j & (2)\\
P[X_k = i | X_{k-1} = j] &= \gamma_{ij}\ , 1 < k \leq T, i, j \in \mathcal{X} & (3)
\end{aligned}
$$

where the latter is (a version of) the regular conditional probability

A class of parameter-driven models

Let $p(y \mid x)$ be a family of kernel density functions (i.e probability densities in y and measurable in x), where y takes values in the same set as the Y_is and $x \in \mathcal{X}$. We are rather loose on the measure-theory, but it is ok.

We assume the following **parameter-driven model**:

$$\begin{tikzcd}
X_1 \arrow[r] \arrow[d] & X_2 \arrow[r] \arrow[d] & X_3 \arrow[r] \arrow[d] & \cdots \arrow[r] \arrow[d, phantom, "\cdots"] & X_T \arrow[d] \\
Y_1 & Y_2 & Y_3 & \cdots & Y_T
\end{tikzcd}$$

X is unobserved/latent, but the Markov dependence in X induces (higher-order) dependence in Y. $p(y|x)$ might be a stochastic matrix when the Y_i's are discetely-valued (e.g **binary or multinomial** data), or a conditional probability (kernel) function.

Statistical estimation in HMMS

On the basis of a sequence of observations $y_1, \ldots, y_T$

- Filtering: Compute probabilities $\pi_t(j) = P[X_t = j | y_{1:t}]$
- Smoothing: Compute probabilities $P[X_t = j | y_{1:T}]$
- Estimation: estimate MC dynamics and other unknown parameters. Let θ be the parameter vector
- MAP estimate: find the most likely (**Viterbi**) path given the observations, i.e the mode of the posterior of the states
- Signal reconstruction: simulate paths according to the posterior distribution of the states

Note that computing the expected path is a trivial collorary of the solution to the smoothing problem

Joint posterior of the states

For concreteness, although we shall revisit such expressions, note that the joint probabilities of the unobserved states given observed data, are obtained via Bayes formula as:

$$\begin{aligned}
&P[X_1 = i_1, X_2 = i_2, \ldots, X_T = i_T | y_{1:T}] \\
=& \frac{P[X_1 = i_1, X_2 = i_2, \ldots, X_T = i_T] \prod_{t=1}^{T} p(y_t \mid X_t = i_t)}{p(y_1, \ldots, y_T)} \\
\propto& P[X_1 = i_1, X_2 = i_2, \ldots, X_T = i_T] \prod_{t=1}^{T} p(y_t \mid X_t = i_t)
\end{aligned} \tag{4}$$

For the role of HMMs in modelling discrete-valued time-series data and model comparisons against alternative models (eg MCs on the observed data, discrete ARMA and models based on thinning) see e.g [MacDonald and Zucchini, 1997].

Whereas in many (most) parameter driven models exact implementation of these tasks involve exponential in T computational cost (due to marginalizations) for HMMs efficient implementation is possible. The core of the efficient implementation is **dynamic programming**. Before proceeding in providing the solution to the aforementioned problems we give a different, more general formulation, and show that a variety of other problems are also a specific instance of this general formulation.

Change of measure

Let $\mathbb{P}_t$ be the joint probabilities of $X_{1:t}$ according to the MC (prior) law, i.e. $\mathbb{P}_t(i_1,\ldots,i_t) = \delta_{i_1}\gamma_{i_1 i_2}\cdots\gamma_{i_{t-1}i_t}$, for $t = 1,\ldots,T$, $i_j \in \mathcal{X}$

Let $w_t(j) \geq 0$, for $j \in \mathcal{X}$ and $t = 1,\ldots,T$, and let $\mathbb{Q}_t$ be a sequence of probability measures defined by the following change of measure w.r.t $\mathbb{P}_t$:

$$\frac{\mathrm{d}\mathbb{Q}_t}{\mathrm{d}\mathbb{P}_t}(X_{1:t}) \propto \prod_{k=1}^{t} w_k(X_k) \tag{5}$$

where in this discrete-valued framework it becomes

$$\frac{\mathbb{Q}_t(i_1,\ldots,i_t)}{\mathbb{P}_t(i_1,\ldots,i_t)} = \prod_{k=1}^{t} w_k(i_k) \tag{6}$$

We define also

$$L_t = \mathbb{E}_{\mathbb{P}_t}\left[\prod_{k=1}^{t} w_k(X_k)\right] \tag{7}$$

L_t is the normalizing constant of $\mathbb{Q}_t$.

Note that for bounded function f,

$$\mathbb{E}_{\mathbb{P}_t}\left[f(X_t)\prod_{k=1}^{t} w_k(X_k)\right]$$

is a Feynman-Kac formula, see eg [Del Moral and Miclo, 2000].

Generalized framework

Let us now consider the problem of computing marginal expectations and sampling from the sequence of measures $\mathbb{Q}_t$.

- Observation 1: statistical tasks for HMMs (16) are a special case of this framework: take the weights to be the likelihood $w_t(j) = p(y_t|X_t = j)$

For example, it is clear that the posterior probabilities of the states (4) are given by $\mathbb{Q}_T$ defined in (5)

In particular:

- Filtering: Compute t-th marginal of $\mathbb{Q}_t$
- Smoothing: Compute t-th marginal of $\mathbb{Q}_T$
- Estimation: Compute L_T
- MAP estimate: find the mode of $\mathbb{Q}_T$
- Signal reconstruction: simulate paths according to $\mathbb{Q}_T$

Conditional independence

- Observation 2 (loosely stated): If $X = (X_1, \dots, X_T)$ is distributed according to $\mathbb{Q}_T$ then X is a non-homegenous Markov chain, i.e this change of measure preserves the Markov property

In the special case of (9) this can be derived by a standard **graph theory** argument: the conditioned process is Markov. Let $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$.

To see that this is indeed so, note that:

1. **Restricted likelihood ratio** Let any two measures $\mathbb{P}$ and $\mathbb{Q}$ on a space $(\Omega, \mathcal{F})$ with density $\xi(\omega) = \mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$, and $\mathcal{G}$ be a sub-σ-algebra. Then the density of the two measures restricted on the information set $\mathcal{G}$ is $\mathbb{E}_{\mathbb{P}}[\xi \mid \mathcal{G}]$.

Therefore, in our case

$$\frac{\mathrm{d}\mathbb{Q}_T}{\mathrm{d}\mathbb{P}_T}|_{\mathcal{F}_t} = \frac{1}{L_T}\mathbb{E}_{\mathbb{P}_T}[\prod_{k=1}^{T} w_k(X_k) \mid \mathcal{F}_t] = \frac{1}{L_T}\prod_{k=1}^{t} w_k(X_k)\, g_t(X_t) \quad (8)$$

where

$$g_t(X_t) := \mathbb{E}_{\mathbb{P}_T}[\prod_{k=t+1}^{T} w_k(X_k) \mid \mathcal{F}_t] = \mathbb{E}_{\mathbb{P}_T}[\prod_{k=t+1}^{T} w_k(X_k) \mid X_t] \quad (9)$$

by the Markov property of $\mathbb{P}_T$

2 **Bayes theorem** The previous result directly yields the following important collorary for generic $\mathbb{Q}, \mathbb{P}, \xi, \mathcal{G}$ as before and any integrable random variable X on $(\Omega, \mathcal{F})$:

$$\mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}] = \frac{\mathbb{E}_{\mathbb{P}}[X\,\xi \mid \mathcal{G}]}{\mathbb{E}_{\mathbb{P}}[\xi \mid \mathcal{G}]} \tag{10}$$

Therefore in our case for any Borel A

$$\begin{aligned}\mathbb{Q}_T[X_t \in A \mid \mathcal{F}_{t-1}] &= \frac{\mathbb{E}_{\mathbb{P}_T}[1[X_t \in A] \prod_{k=1}^{T} w_k(X_k) \mid \mathcal{F}_{t-1}]}{\mathbb{E}_{\mathbb{P}_T}[\prod_{k=1}^{T} w_k(X_k) \mid \mathcal{F}_{t-1}]} \\ = \frac{h_{t-1}(X_{t-1})}{g_{t-1}(X_{t-1})} &= \mathbb{Q}_T[X_t \in A \mid X_{t-1}]\end{aligned} \tag{11}$$

due to the conditional independence property of $\mathbb{P}_T$, where last equation follows from same argument as the first.

Hence, by a standard argument (see for example Prop.5.6 of [Kallenberg, 1997]) we have the conditional independence.

The MC dynamics are given in 2nd equation, where $h_t(X_t) := \mathbb{E}_{\mathbb{P}_T}[1[X_t \in A] \prod_{k=t+1}^{T} w_k(X_k) \mid X_t]$

All conditioned Markov processes we consider in this lecture are Markov.

Multitude of problems

- Observation 3: our generic framework (5) encapsulates a wide variety of problems of discrete probability other than inference for HMMs

For example:

1. our canonical problem (1) is a special case: take $w_t = 1$ for any $t < T$ and $w_T(y) = 1$ and $w_T(u) = 0$ for any $u \neq y$.

2 independent random variables conditioned on their sum: Let Z_i be independent positive discrete random variables, let $S_t = \sum_{i=1}^{t} Z_i$. We are interested in the probabilities (or simulation) of each Z_i conditionally on the value of $S_T = k$. However, $X_t = (Z_t, S_t)$ is a MC, and we set $w_t = 1$ for all $t < T$, and $w_T(j, l) = 1$ if $l = k$, and 0 otherwise

3 Computational game theory: efficient computation of the exponentially weighted average forecaster for the so-called online shortest path problem; easy to incorporate hard constraints (by state-space expansion and/or adjustments of the generator)

Methodology I: dynamic programming

The methodology we use to solve the various problems of interest is inspired by a dynamic programming approach for optimization of chain-structured objective functions with discrete-valued inputs. What we describe below is a generic scheme, and in our context solves directly the MAP problem. However, the simulation techniques are also elaborations of this idea. For a short description see S.2.4 of [Liu, 2008].

Minimization

Let $x_t \in \mathcal{X} = \{1, \ldots, d\}$ and the objective function have a chain structure

$$H_T(x) = \sum_{t=1}^{T-1} h_t(x_t, x_{t+1}) \tag{12}$$

hence it is a sum of terms which involve functions of pairs, and note we have no loops. In general a naive optimization of a function of T discrete-valued inputs by enumeration has computational cost exponentially large in T. The particular (Markovian in a sense) structure allows us to do much better.

Let

$$c_1(j) = \min_i h_1(i,j) \tag{13}$$

and for $t > 1$

$$c_t(j) = \min_i \{c_{t-1}(i) + h_t(i,j)\} \tag{14}$$

It should be clear that $\min_j c_{T-1}(j)$ yields the minimum of H_T (e.g consider $T = 2$). However, the significant advantage of the formulation is computational: $\mathcal{O}(Td^2)$

Hence, a forward computation yields the minimum of H_T. To find the minimizer we need to run a backward computation

Let

$$\hat{x}_T = \arg\min_j c_{T-1}(j)$$

no other value of the last argument can yield a smaller value for the objective function.

Similarly, for $t < T$, let

$$\hat{x}_t = \arg\min_j \{c_{t-1}(j) + h_t(j, \hat{x}_{t+1})\}$$

Derivation of the $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_T)$ can be done at $\mathcal{O}(Td)$ cost

It should be clear that $\hat{x} = \arg\min_x H_T(x)$ and it can be obtained at $\mathcal{O}(Td^2)$ cost

Viterbi path

Recall that it is the mode of $\mathbb{Q}_T$. Writing the density of $\mathbb{Q}_T$ w.r.t the counting measure,

$$q_T(x_1, \ldots, x_T) = \frac{1}{L_T}\delta_{x_1} \prod_{t=1}^{T-1} \gamma_{x_t, x_{t+1}} \prod_{t=1}^{T} w_t(x_t) \tag{15}$$

the Viterbi path is the **minimizer of the negative log-density** (energy), i.e

$$-\log\left\{\delta_{x_1} \prod_{t=1}^{T-1} \gamma_{x_t, x_{t+1}} \prod_{t=1}^{T} w_t(x_t)\right\} = \sum_{t=1}^{T-1} h_t(x_t, x_{t+1}) \tag{16}$$

therefore the dynamic programming yields the Viterbi path

Filtering

We will construct similar recursions to (14) for the evolution of the filtering probabilities, π_t, which recall that can be thought as the probabilities of X_t given $y_{1:t}$ in the HMM (9), or the t-th marginal of $\mathbb{Q}_t$ in (5).

The same filtering recursion can be derived in various ways, e.g: i) sparce linear algebra: re-arrangment of the sum(integral) of (15) ii) Bayesian "elementary" calculation for HMMs (9) using regular conditional probability and conditional independence, iii) using the change of measure and the generalized Bayes formula (10)

All variants rely on a marginalization step which one way or another is a version of Fubini's theorem. ii) and iii) extend easily to other contexts, iii) being a more formal version of ii). We can identify iii) with the so-called reference probability approach, for a book-length description see [Elliott et al., 1995]

For pegagogical reasons (but also to give a perspective of different approaches in the literature) I will demonstrate the recursion with all three ways mentioned above (rather quickly):

filtering recursion

$$\pi_t(j) = \frac{L_{t-1}}{L_t} \sum_i \pi_{t-1}(i)\gamma_{i,j} w_t(j) \tag{17}$$

hence all filtering probabilities are obtained at a $\mathcal{O}(Td^2)$

Algebraic derivation

Approach common in engineering literature. We are interested in the t-th marginal of $\mathbb{Q}_t$, see [MacDonald and Zucchini, 1997]. Based on the Legesgue density (15):

$$\pi_t(j) = \mathbb{Q}_t[\cup\{(i_1, \ldots, i_{t-1}, j)\}] = \sum_{i_1,\ldots,i_{t-1}} q_t(i_1, \ldots, i_{t-1}, j)$$

$$= \frac{1}{L_t} \sum_{i_1,\ldots,i_{t-1}} \delta_{i_1} \prod_{k=1}^{t-2} \gamma_{i_k,i_{k+1}} \prod_{k=1}^{t-1} w_k(i_k) \gamma_{i_{t-1},j} w_t(j)$$

$$= \frac{L_{t-1}}{L_t} \sum_i \pi_{t-1}(i) \gamma_{i,j} w_t(j)$$

We can put this in matrix form for the flow of π_t

Bayesian calculation

Approach common in statistical non-linear filtering. We concentrate on the HMM case (9), e.g [Gordon et al., 1993]

$$
\begin{aligned}
\pi_t(j) &= P[X_t = j \mid y_{1:t}] = \sum_i P[X_t = j, X_{t-1} = i \mid y_{1:t}] \\
&= \sum_i \frac{p(y_t \mid X_t = i) P(X_t = j \mid X_{t-1} = j) P[X_{t-1} = i \mid y_{1:t-1}]}{p(y_t \mid y_{1:t-1})} \\
&= \frac{L_{t-1}}{L_t} \sum_i w_t(j) \gamma_{i,j} \pi_{t-1}(i)
\end{aligned}
$$

Change of measure

Approach inspired by stochastic analysis approaches to continuous-time filtering, see [Elliott et al., 1995] . Let $(X_1, \ldots, X_t) \sim \mathbb{Q}_t$ in (5).

By (10):

$$\pi_t(j) = \mathbb{E}_{\mathbb{Q}_t}[1[X_t = j]] = \frac{1}{L_t}\mathbb{E}_{\mathbb{P}_t}\left[1[X_t = j]\prod_{i=1}^{t} w_i(X_i)\right]$$

$$= \frac{1}{L_t}\mathbb{E}_{\mathbb{P}_t}\left[\prod_{i=1}^{t-1} w_i(X_i)\,\mathbb{E}_{\mathbb{P}_t}[1[X_t = j]w_t(X_t) \mid \mathcal{F}_{t-1}]\right]$$

$$= \frac{1}{L_t}\mathbb{E}_{\mathbb{P}_t}\left[\prod_{i=1}^{t-1} w_i(X_i)\,\mathbb{E}_{\mathbb{P}_t}[1[X_t = j]w_t(X_t) \mid X_{t-1}]\right]$$

$$= \frac{1}{L_t}\mathbb{E}_{\mathbb{P}_t}\left[\prod_{i=1}^{t-1} w_i(X_i)\, w_t(j)\mathbb{E}_{\mathbb{P}_t}[1[X_t = j] \mid X_{t-1}]\right]$$

$$= \frac{1}{L_t}\mathbb{E}_{\mathbb{P}_t}\left[\prod_{i=1}^{t-1} w_i(X_i)\, w_t(j)\gamma_{X_{t-1},j}\right] = \frac{1}{L_t}\mathbb{E}_{\mathbb{P}_{t-1}}\left[\prod_{i=1}^{t-1} w_i(X_i)\, w_t(j)\gamma_{X_{t-1},j}\right]$$

$$= \frac{L_{t-1}}{L_t}\mathbb{E}_{\mathbb{Q}_{t-1}}\left[w_t(j)\gamma_{X_{t-1},j}\right] = \frac{L_{t-1}}{L_t}\sum_i w_t(j)\gamma_{i,j}\pi_{t-1}(i)$$

Smoothing

We are interested in the t-th marginal of $\mathbb{Q}_T$: $\phi_t(j) = P[X_t = j|y_{1:T}]$. We will present the Bayesian calculation but the other approaches can be used

$$P[X_t = j|y_{1:T}] \quad \propto \quad \underbrace{P[y_{t+1:T} \mid X_t = j]}_{\text{backward } \phi_t(j)} \quad \times \quad \underbrace{P[X_t = j \mid y_{1:t}]}_{\text{forward } \pi_t(j)}$$

Note that

$$\begin{aligned} \phi_{T-1}(j) &= p(y_T \mid X_{T-1} = j) = \sum_i w_T(i)\gamma_{j,i} \qquad (18) \\ \phi_t(j) &= \sum_i p(y_{t:T}, X_t = i \mid X_{t-1} = j) = \sum_i \phi_{t+1}(i) w_t(i)\gamma_{j,i} \end{aligned}$$

hence at $\mathcal{O}(Td^2)$ we obtain the backward equations and hence the smoothing probabilities

Reconstruction

Simulation of $X = (X_1 \ldots, X_T) \sim \mathbb{Q}_T$. This is precisely our original problem: simulation of (class of) conditioned Markov processes

We do not need both forward and backward filters for this. Either is enough. Suppose we have computed the forward filter. Then

$$
\begin{aligned}
& X_T \sim \pi_T \quad (T\text{-th marginal of}\, \mathbb{Q}_T) \\
& P[X_t = j \mid y_{1:T}, X_{t+1} = i] \propto P[y_{t+1:T}, X_{t+1} = i \mid X_t = j, y_{1:t}] \\
& \times P[X_t = j \mid y_{1:t}] \\
& \propto \gamma_{j,i} \pi_t(j)
\end{aligned}
$$

Likelihood computation

For generic parameter-driven problems is an exponentially hard problem:

$$L_T = \sum_{i_1,\ldots,i_T} p_T(i_1,\ldots,i_T) \prod_j w(i_j)$$

- Same techniques developed above compute it at $\mathcal{O}(Td^2)$
- Direct optimization (e.g Nelder-Mead) or EM algorithm for parameter estimation. Stories...

Part II: Conditioned Markov Jump Processes (MJPs)

We now consider continuous-time discrete state-space MCs, sometimes called MJPs. Unlike the discrete case I will not present the most general setup where the mathematics are explicit, instead I will consider a natural extension of the previous framework where the observations are noisy versions of the signal, but the latter evolves continuously in time.

Additionally, unlike HMMs I will only try to obtain a mathematical description of the conditioned process and mention how simulation can be done.

Effectively, this second part (as the final third part) focuses on the canonical problem (1)

MJP

Again for simplicity we consider a time-homegeneous signal (although again conditioning will yield a time-inhomeneous process), $X = (X_t, t \geq 0)$, with state-space $\mathcal{X} = \{1, \ldots, d\}$. The process is characterized by its generator Q, which in this case is a $d \times d$ matrix with elements $Q_{ij} \geq 0$ for $j \neq i$ and $Q_{ii} = -\sum_{j \neq i} Q_{i,j}$.

Let $p_{i,j}(t) = P[X_t = j \mid X_0 = i]$. Then for small Δ

$$p_{i,j}(\Delta) \approx 1[i = j] + \Delta Q_{i,j}$$

The constraint of Q implies a sum-to-1 constraint on p. Let also δ be the initial distribution

Master Equation/Kolmogorov forward equation

Let $p_i(t) = P[X_t = i]$. Then, (for example) by a conditioning argument we get

$$p_j(t+\Delta) = \sum_i p_i(t) p_{i,j}(\Delta) \approx p_j(t) \left(1 - \Delta \sum_{j \neq i} Q_{j,i} \right) + \Delta \sum_{j \neq i} Q_{i,j} p_i(t)$$

where by taking limits we get the so-called Master Equation or KFE for the flow of marginal probabilities, or effectively for the transition density,

$$\partial p_j(t)/\partial t = \sum_{j \neq i} (-Q_{j,i} p_j(t) + Q_{i,j} p_i(t))$$

or in matrix form the system of ODEs

$$\partial p(t)/\partial t = Qp$$

Note that $s \rightarrow p(s)$ is a continuous map

Partial observations

Consider a sequence of times (chosen indpendently of X) $0 \leq t_1, \ldots, t_n \leq T$, and conditionally independent observations $y_i, i = 1, \ldots, n$ with $p(y_i \mid X_{t_i} = j)$

- Skeleton dynamics: our HMM solution directly gives us the dynamics of the conditioned skeleton $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$.
- From a simulation perspective we expect the complicated part to be to generate the MJP bridge
- The conditioned process is also an MJP

In the next couple of slides we will give the mathematical description of X given $y_{1:n}$ (i.e filter and smoother). We will see that they follow same lines as the discrete-time case with ODEs replacing recursions. I will follow a Bayesian calculation approach to derive expressions. Then I will comment on the simulation (reconstruction).

Filter

Let $y_{[0,t]}$ denote all available observations in $[0, t]$, and let $\pi_i(t) = P[X_t = i \mid y_{[0,t]}]$. Clearly, $\pi_i(s) = p_i(s)$ for all $s < t_1$. On the other hand

$$\pi_i(t_1) \propto p_i(t_1) p(y_1 \mid X_{t_1} = i)$$

hence the flow of filtering probabilities has a right discontinuity at t_1. Hence, $\pi(s)$ solves the master equation as $p(s)$ for all $s < t_1$ (and the same initial condition) and has a jump at t_1. The same argument shows that on $[t_1, t_2)$ $\pi(s)$ solves the master equation with an updated initial condition.

Smoother

One approach is to work as before by treating separately each term of Bayes formula. We show this below.

$$\begin{aligned} P[X_s = i \mid y_{1:n}] &= P[X_s = i \mid y_{[0,s]}, y_{(s,T]}] \\ &\propto P[y_{(s,T]} \mid X_s = i]\pi_i(s) \end{aligned}$$

Let $\phi_i(t) = P[y_{(t,T]} \mid X_t = i]$. By conditioning argument we get for $t_{n-1} < t \leq t_n$

$$
\begin{aligned}
\phi_i(t) &= \sum_j p(y_n, X_{t+\Delta} = j \mid X_t = i] \\
&= \sum_j \phi_j(t+\Delta) p_{i,j}(\Delta) \\
&\approx \phi_i(t+\Delta)(1 - \Delta \sum_{i \neq j} Q_{i,j}) + \sum_{j \neq i} \phi_j(t+\Delta) \Delta\, Q_{i,j}
\end{aligned}
$$

hence

$$\partial \phi_i(t) / \partial t = \sum_{j \neq i} Q_{i,j} (\phi_i(t) - \phi_j(t))$$

with initial condition $\phi_{t_n}(i) = p(y_n \mid X_{t_n} = i)$. Then $\phi_i(t)$ is right-continuous and
$\lim_{t \downarrow t_{n-1}} \phi_i(t_{n-1}) = \lim_{t \downarrow t_{n-1}} \phi_i(t) p(y_{n-1} \mid X_{t_{n-1}} = i)$

Smoothed rates

The conditioned MJP we consider are also MJPs with modified generator. Again a Bayesian calculation (and right-continuity of ϕ) gives

$$\begin{aligned}
P[X_{t+\Delta} = j \mid X_t = i, y_{1:n}] &= P[X_{t+\Delta} = j \mid X_t = i, y_{[0,t]}, y_{(t,T]}] \\
&= \frac{P[X_{t+\Delta} = j, X_t = i, y_{[0,t]}, y_{(t,T]}]}{P[X_t = i, y_{[0,t]}, y_{(t,T]}]} \\
&= \frac{p_{i,j}(\Delta)\pi_i(t)\phi_j(t+\Delta)}{\pi_i(t)\phi_i(t)} \\
&\approx \Delta Q_{i,j}\frac{\phi_j(t)}{\phi_i(t)} \quad \text{for } j \neq i
\end{aligned}$$

which shows that the conditioned process is a non-homogenous MJP with rate function

$$Q_{i,j}(t) = Q_{i,j}\frac{\phi_j(t)}{\phi_i(t)} \tag{19}$$

Note that the transition density of this MJP are the smoothing probabilities, hence we can get ODEs for them from the corresponding master equation.

Variational derivations

For discrete-time HMM we saw a variety of approaches to derive recursions and to characterize the conditioned process. For MJPs (and also for diffusion processes considered later) there is a really neat argument due to Manfred Opper, see for example [Opper and Sanguinetti, 2008].

Idea is to compute the KL divergence between a measure on the path space and the posterior law of X on the path space, and minimize this distance over the class of MJPs using constrained calculus of variation. Since the true posterior process is an MJP we get the exact dynamics. Motivation

Simulation of conditioned MJPs

My impression is that the literature is rather incomplete on the exact solution of (1). I am aware of the very interesting [Fearnhead and Sherlock, 2006], based on thinning of Poisson processes. Suggested references most welcome

Part III: conditioned diffusion processes

We move to certain class of continuous time and space Markov processes. We model d-dimensional stochastic process $X \in R^d$ as the solution of an SDE of the type:

$$\mathrm{d}X_s = b(s, X_s; \theta)\,\mathrm{d}s + \sigma(s, X_s; \theta)\,\mathrm{d}B_s, \quad s \in [0, T] \ ; \tag{20}$$

B is an m-dimensional standard Brownian motion, $b(\cdot, \cdot) : R_+ \times R^d \to R^d$ is the *drift*, $\sigma(\cdot, \cdot) : R_+ \times R^d \to R^{d \times m}$ is the *diffusion coefficient*. The initial point V_0 can be taken as fixed or elicited with a distribution, depending on the context. Also let

$$\Gamma = \sigma\sigma^*$$

We assume that coefficients are sufficiently regular so that (20) has a unique weak non-explosive solution

Simulation of diffusions

Compared to the previous problems this is much harder. Any path X is now a really infinite-dimensional object with no obvious sparse representation (as for example an MJP). Hence, it is both exact and efficient that are under serious scrutiny. Recent advances, e.g [Beskos et al., 2006], have shown the at first remarkable possibility of doing both exact and efficient simulation of unconditioned diffusions. Exact (but inefficient) simulation of conditioned diffusions is also possible.

Diffusion bridges

We are interested in (1). The theory of *h-transforms*, see for example Chapter IV.39 [Rogers and Williams, 2000], allows us to derive its SDE:

$$\begin{aligned} \mathrm{d}X_s &= \tilde{b}(s, X_s)\,\mathrm{d}s + \sigma(s, X_s)\,\mathrm{d}B_s, \quad s \in [0, T]\,, X_0 = u; \\ \tilde{b}(s, u) &= b(s, x) + [\sigma\sigma^*](s, x)\,\nabla_x \log p_{s,T}(x, v) \end{aligned} \tag{21}$$

- the local characteristics of the unconditioned and conditioned processes are the same
- the drift of the conditioned process includes an extra term which forces the process to hit v at time T. Note the similarity to (19)
- (21) is typically intractable since the drift is expressed in terms of the transition density (which we need as a function of the starting point)

Outline

The target conditioned process has intractable dynamics. Our strategy (which to a large extent is the state-of-the-art in this problem) will be to use importance sampling (IS) whereby we propose paths from another tractable process and weight accordingly. Hence, The development is again based on a change of measure. However, the simulation will be also based on a change of measure by means of importance/rejection sampling.

- Identify valid and tractable proposals
- Work out the likelihood ratio and turn on the IS machine

Importance sampling

Importance sampling (IS) is a classic Monte Carlo technique for obtaining samples from a probability measure $\mathbb{P}$ using samples from another probability measure $\mathbb{Q}$, see for example Chapter 2.5 of [Liu, 2008] for an introduction. Mathematically it is based on the concept of *change of measure*.

Suppose that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ with Radon-Nikodym density $f(x) = \mathbb{P}(\mathrm{d}x)/\mathbb{Q}(\mathrm{d}x)$. Then, in its simplest form IS consists of constructing a set of weighted particles (x_i, w_i), $i = 1, \ldots, N$, where $x_i \sim \mathbb{Q}$, and $w_i = f(x_i)$. This set gives a Monte Carlo approximation of $\mathbb{P}$, in the sense that for suitably integrable functions g, we have that

$$\frac{\sum_{i=1}^{N} g(x_i) w_i}{N} \,. \tag{22}$$

is an unbiased and consistent estimator of

$$\mathbb{E}_{\mathbb{P}}[g] := \int g(x)\mathbb{P}(\mathrm{d}x)\,.$$

IS can be cast in much more general terms, an extension particularly attractive in the context of stochastic processes. First, note that in most applications f is known only up to a normalising constant, $f(x) = cf_u(x)$, where only f_u can be evaluated and

$$c = \mathbb{E}_{\mathbb{Q}}[f_u]\,. \tag{23}$$

The notion of a properly weighted sample refers to a set of weighted particles (x_i, w_i), where $x_i \sim \mathbb{Q}$ and w_i is an unbiased estimator of $f_u(x_i)$, that is

$$\mathbb{E}_{\mathbb{Q}}[w_i \mid x_i] = f_u(x_i)\,.$$

Then for any integrable g

$$\mathbb{E}_{\mathbb{Q}}[gw] = \mathbb{E}_{\mathbb{P}}[g]\,\mathbb{E}_{\mathbb{Q}}[w]\,. \tag{24}$$

Rearranging the expression we find that a consistent estimator of $\mathbb{E}_{\mathbb{P}}[g]$ is given by

$$\frac{\sum_{i=1}^{N} g(x_i) w_i}{\sum_{i=1}^{N} w_i} . \tag{25}$$

When w_i is an unbiased estimator of $f(x_i)$ we have the option of using (22), thus yielding an unbiased estimator. However, (25) is a feasible estimator when c is unknown.

Although the first moment of w (under $\mathbb{Q}$) exists by construction, the same is not true for its second moment. Hence it is a minimal requirement of a "good" proposal distribution $\mathbb{Q}$ that $\mathbb{E}_{\mathbb{Q}}[w^2] < \infty$. In this case, and using the Delta method for ratio of averages it can be shown that (25) is often preferable to (22) in a mean square error sense because the denominator acts effectively as a control variable.

IS includes exact simulation as a special case when $\mathbb{Q} = \mathbb{P}$. Another special case is rejection sampling (RS), which assumes further that $f_u(x)$ is bounded in x by some calculable $K < \infty$. Then, if we accept each draw x_i with probability $f_u(x_i)/K$, the resulting sample (of random size) consists of independent draws from $\mathbb{P}$. This is a special case of the generalised IS where w_i is a binary 0-1 random variable taking the value 1 with probability $f_u(x_i)/K$.

IS for diffusions

For the sake of simplicity we consider the simplest case where $d = 1$, the process is time-homogenous and $\sigma = 1$:

$$\mathrm{d}X_s = b(X_s)\mathrm{d}s + \mathrm{d}B_s \quad s \in [0, T], X_0 = u;$$

Recall that out ultimate target is to simulate from X given also $X_T = v$. For the general case see for example [Papaspiliopoulos and Roberts, 2009]

Our main tool will be an expression for the likelihood ratio of unconditional Itô processes

Girsanov's theorem

Let $\mathbb{P}_b$ and $\mathbb{P}_0$ be the probability laws implied by the (20) with drift b and 0 (i.e the Brownian motion) respectively. Then, under certain conditions $\mathbb{P}_b$ and $\mathbb{P}_0$ are equivalent with density (*continuous time likelihood*) on $\mathcal{F}_t = \sigma(X_s, s \le t)$, $t \le T$, given by

$$\xi = \left.\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_0}\right|_t = \exp\left\{ \int_0^t b(X_s)\mathrm{d}X_s - \frac{1}{2}\int_0^t b^2(X_s)\mathrm{d}s \right\} . \qquad (26)$$

Think of this as the ratio of probabilities that a given paths has been generated by the diffusion relative to have been generated by the BM. Note that a naive way to simulate X unconditionally would be to simulate X under $\mathbb{P}_0$ and weight by (30). This still raises the issue of doing it **exactly**. But what about the process conditioned on the end point $X_T = v$?

Change of measure for conditioned processes

Let $\mathbb{P}_b^*$ and $\mathbb{P}_0^*$ denote the laws of the corresponding diffusion bridges conditioned on $X_T = v$. $\mathbb{P}_0^*$ generates a nice, tractable, Gaussian process, the Brownian bridge

Also let $p_{0,T}(u,v)$ and $\mathcal{G}_{0,T}(u,v)$ denote the transition densities of the two processes, the latter simply being a Gaussian density, the former being intractable.

What can we say about the density between $\mathbb{P}_b^*$ and $\mathbb{P}_0^*$? If we had that then we could switch the IS on...

However, we already know how to do get it: see (30). Let $\mathcal{G} = \sigma(X_0, X_T)$ then, from one side we have that

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_0}|_{\mathcal{G}} = \frac{p_{0,T}(u,v)}{\mathcal{G}_{0,T}(u,v)} \tag{27}$$

but from the other

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_0}|_{\mathcal{G}} = \mathbb{E}_{\mathbb{P}_0^*}[\xi] \tag{28}$$

and from (10) we get that

$$\frac{\mathrm{d}\mathbb{P}_b^*}{\mathrm{d}\mathbb{P}_0^*} = \frac{\xi}{\mathbb{E}_{\mathbb{P}_0^*}[\xi]} = \frac{\mathcal{G}_{0,T}(u,v)}{p_{0,T}(u,v)}\xi \tag{29}$$

Therefore, at least approximately we can do IS...Can we do better?

Exact Simulation of Diffusion Bridges

Apart from Girsanov's theorem, another valueable tool is at hand, Itô's Lemma. This allows us to do integration by parts and simplify ξ in (30). Let $B(x)' = b(x)$, and assume that b is integrable then

$$\begin{aligned}\log \xi &= \int_0^t b(X_s)\mathrm{d}X_s - \frac{1}{2}\int_0^t b^2(X_s)\mathrm{d}s\,. \\ &= B(X_T) - B(X_0) - \int_0^t \frac{1}{2}(b^2 + b')(X_s)\mathrm{d}s\end{aligned} \tag{30}$$

Assume now that $-\infty < \ell < (b^2 + b')/2 < \ell + r < \infty$. Then define

$$\phi(u) = ((b^2 + b')/2 - \ell)/r \tag{31}$$

Simulating an event of equal probability

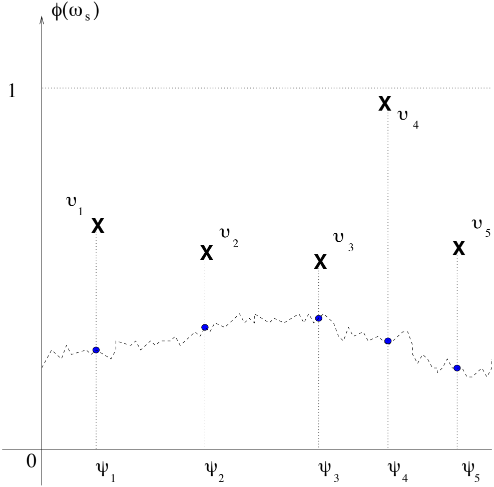
Hence putting everything together we get that

$$\frac{\mathrm{d}\mathbb{P}_b^*}{\mathrm{d}\mathbb{P}_0^*} \propto \xi \propto \exp\left\{-r\int_0^T \phi(X_s)\mathrm{d}s\right\} \leq 1 \tag{32}$$

Theorem

Let Φ *be a homogeneous Poisson process of intensity* r *on* $[0, T] \times [0, 1]$ *and* N *is the number of points of* Φ *below the graph* $s \mapsto \phi(X_s)$, $s \in [0, T]$, *then:*

$$\mathrm{P}\left[\, N = 0 \,|\, X \,\right] = \exp\left\{-r\int_0^T \phi(X_s)\mathrm{d}s\right\}$$


φ(ω_s)
1
0
υ_1
υ_2
υ_3
υ_4
υ_5
X
X
X
X
X
ψ_1
ψ_2
ψ_3
ψ_4
ψ_5

Idealized rejection sampler

1. Simulate $X \sim \mathbb{P}_0^*$
2. Simulate a $Po(r)$ process $\Phi = \{z_1, z_2, \ldots, z_\kappa\}$, $z_j = (z_{j,1}, z_{j,2})$, $z_{j,1} \in [0, t], z_{j,2} \in [0, 1]$, $1 \leq j \leq \kappa$
3. Compute the acceptance indicator I:

$$I := \prod_{j=1}^{\kappa} \mathbb{I}\left[\phi(X_{z_{j,1}}) < z_{j,2}\right]$$

4. If $I = 1$ accept X, otherwise return to 1 and retry.

Retrospective Exact Simulation

1. Simulate $\Phi = \{z_1, z_2, \dots, z_\kappa\}$
2. Simulate the values of $X \sim \mathbb{P}_0^*$, at the time instances $z_{j,1}$, $1 \leq j \leq \kappa$, therefore:

$$S(X) = \{(0, x), (z_{1,1}, X_{z_{1,1}}), \dots, (z_{\kappa,1}, X_{z_{\kappa,1}}), (t, y)\}$$

3. Compute the acceptance indicator I.
4. If $I = 1$ then accept and return the proposed skeleton $S(X)$; otherwise return to 1 and retry.

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2006).
Retrospective exact simulation of diffusion sample paths with applications.
Bernoulli, 12:1077–1098.

Del Moral, P. and Miclo, L. (2000).
Branching and interacting particle systems. Approximations of Feymann-Kac formulae with applicationc to non-linear filtering, volume 1729.
Springer, Berlin.

Elliott, R. J., Aggoun, L., and Moore, J. B. (1995).
Hidden Markov models, volume 29 of *Applications of Mathematics (New York)*.
Springer-Verlag, New York.
Estimation and control.

Fearnhead, P. and Sherlock, C. (2006).
An exact Gibbs sampler for the Markov-modulated Poisson process.
J. R. Stat. Soc. Ser. B Stat. Methodol., 68(5):767–784.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993).
Novel approach to nonlinear/non-Gaussian Bayesian state estimation.
IEE Proceedings Part F: Communications, Radar and Signal Processing, 140:107–113.

Kallenberg, O. (1997).
Foundations of modern probability.
Probability and its Applications (New York). Springer-Verlag, New York.

Liu, J. S. (2008).
Monte Carlo strategies in scientific computing.
Springer Series in Statistics. Springer, New York.

MacDonald, I. L. and Zucchini, W. (1997).
Hidden Markov and other models for discrete-valued time series, volume 70 of *Monographs on Statistics and Applied Probability*.
Chapman & Hall, London.

Opper, M. and Sanguinetti, G. (2008).
Variational inference for markov jump processes.
In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, page 11051112. MIT Press, Cambridge, MA.

Papaspiliopoulos, O. and Roberts, G. (2009).
Importance sampling techniques for estimation of diffusion models.
In *SEMSTAT*. Chapman and Hall.

Rogers, L. C. G. and Williams, D. (2000).
Diffusions, Markov processes, and martingales. Vol. 1.
Cambridge Mathematical Library. Cambridge University Press, Cambridge.
Foundations, Reprint of the second (1994) edition.