Advanced Simulation Methods

Chapter 8 - Hidden Markov Models and Sequential Importance Sampling

The general hidden Markov models, which are described in Section 1.1, provide an extremely flexible framework for modeling time series. The great descriptive power of these models comes at the expense of intractability: it is impossible to obtain analytic solutions to the inference problems of interest, with the exception of a small number of particularly simple cases (finite state space, linear Gaussian recursions). Sequential Monte Carlo (SMC) methods, aka "particle methods", constitute a broad and popular class of Monte Carlo algorithms that have been developed over the past twenty years to provide approximate solutions to these intractable inference problems. For a more detailed treatment of SMC, see [1], [2] and [3]. In these notes, we introduce Sequential Importance Sampling, which is the precursor of SMC.

1 Inference in Hidden Markov Models

1.1 Hidden Markov Models

Consider an $\mathbb{X}-$valued discrete-time Markov process $(X_t)_{t\geq 1}$ such that

$$X_1 \sim \mu_\theta(\cdot) \text{ and for all } t \geq 2 \quad X_t | X_{t-1} = x_{t-1} \sim f_\theta(\cdot | x_{t-1}) \tag{1}$$

where μ_θ is a probability density function (called the "initial distribution") and $f_\theta(\cdot|x)$ denotes the probability density associated with the transition kernel of the Markov process (also called the "transition distribution"). The index θ corresponds to some parameter of the distributions (see examples below). We are interested in estimating $\{X_t\}_{t\geq 1}$ but we do not observe it. We only have access to the $\mathbb{Y}-$valued process $(Y_t)_{t\geq 1}$. It is assumed that, given $(X_t)_{t\geq 1}$ and a parameter value θ, the observations $(Y_t)_{t\geq 1}$ are statistically independent one to the other, and the marginal law of Y_t depends only on X_t, the hidden state at the curren time. In other words, the conditional laws of the observations are given by

$$\forall t \geq 1 \quad Y_t | (X_k = x_k)_{k\geq 1} \sim g_\theta(\cdot | x_t), \tag{2}$$

where θ denotes also the parameter of g_θ. The distribution g is sometimes called the measurement distribution, or the emission distribution, or observation distribution. Note that typically there would be some parameter θ_1 for the initial distribution, some parameter θ_2 for the transition distribution, and some parameter θ_3 for the measurement distribution. We write $\theta = (\theta_1, \theta_2, \theta_3)$, thus putting all the parameters in a single vector θ.

For the sake of simplicity, we have only considered case of homogeneous models here; that is, the transition and observation densities are independent of the time index t. The extension to the non-homogeneous case is straightforward. *It is assumed throughout these notes that the model parameter θ is known, thus we focus on the inference of the hidden process (X_t) given (Y_t).* We will come back to the inference on θ later; in the meantime we drop θ from the notation, and write μ, f, g, implicitly referring to a fixed value of θ.

Models specified as in Eq. (1)-(2) are known as hidden Markov models (HMM) or state space models (SSM). A representation of the dependence between the variables is shown on Figure (1). The following examples provide an illustration of several simple models within this framework.

Example 1 ***Finite State Space HMM.*** *In this case, we have $\mathbb{X} = \{1, ..., m\}$ so*

$$\forall kl, \in \mathbb{X} \quad \mathbb{P}(X_1 = k) = \mu(k), \; \mathbb{P}(X_t = k | X_{t-1} = l) = f(k | l).$$

If the observations are also in a finite state space $\mathbb{Y} = \{1, ..., n\}$, then the observation distribution corresponds to a collection of distributions $g(\cdot \mid x)$ on $\mathbb{Y}$ indexed by the hidden state x:

$$\forall j \in \mathbb{Y} \quad \forall k \in \mathbb{X} \quad \mathbb{P}(Y_t = j | X_t = k) = g(j | k).$$

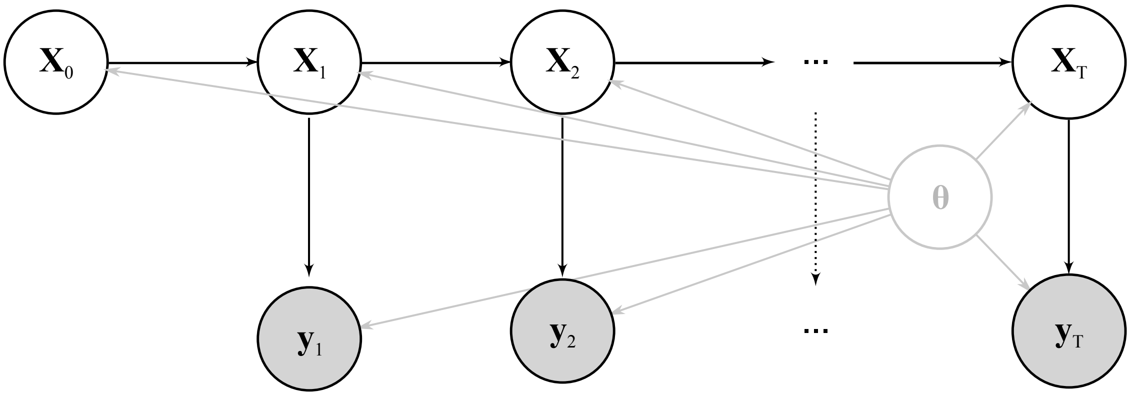

Figure 1: Graphical representation of the variables defining a state space model. Figure borrowed from L. Murray, *Bayesian State-Space Modelling on High-Performance Hardware Using LibBi.*

The parameter θ is then made of the probabilities $\mu(k)$ for all $k \in \mathbb{X}$, $f(k \mid l)$ for all $k, l \in \mathbb{X}$, and $g(j \mid k)$ for all $j \in \mathbb{Y}$, $k \in \mathbb{X}$. This type of model is extremely general and examples can be found in areas such as genetics in which they can describe genetic sequences observed with measurement errors, signal processing, and computer science in which they can describe arbitrary finite-state machines. Inference in finite state space HMMs can be performed exactly using specific algorithms, for instance the Viterbi algorithm, the forward-backward algorithm and Baum-Welch's algorithm.

Example 2 ***Linear Gaussian model***. *Here, $\mathbb{X} = \mathbb{R}^{n_x}$, $\mathbb{Y} = \mathbb{R}^{n_y}$, $X_1 \sim \mathcal{N}(0, \Sigma)$ and*

$$\begin{aligned} X_t &= AX_{t-1} + BV_t, \\ Y_t &= CX_t + DW_t \end{aligned}$$

where $V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{n_v})$, $W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{n_w})$ and A, B, C, D are matrices of appropriate dimensions; the parameter θ is made of A, B, C, D. In this case $\mu(x) = \mathcal{N}(x; 0, \Sigma)$, $f(x' | x) = \mathcal{N}\left(x'; Ax, BB^T\right)$ and $g(y | x) = \mathcal{N}\left(y; Cx, DD^T\right)$. Since inference is analytically tractable for this model using the Kalman filter, it has been extremely widely used for problems such as target tracking and signal processing.

Example 3 ***Stochastic Volatility model***. *We have $\mathbb{X} = \mathbb{Y} = \mathbb{R}$, $X_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right)$ and*

$$\begin{aligned} X_t &= \alpha X_{t-1} + \sigma V_t, \\ Y_t &= \beta \exp(X_t/2) W_t \end{aligned}$$

where $V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. In this case we have $\mu(x) = \mathcal{N}\left(x; 0, \frac{\sigma^2}{1-\alpha^2}\right)$, $f(x' | x) = \mathcal{N}\left(x'; \alpha x, \sigma^2\right)$ and $g(y | x) = \mathcal{N}\left(y; 0, \beta^2 \exp(x)\right)$. The parameter θ is then made of (α, σ, β). Note that this choice of initial distribution ensures that the marginal distribution of X_t is also $\mu(x)$ for all t. This type of model, and its generalizations, have been very widely used in various areas of economics and mathematical finance. It is not a linear Gaussian model, thus Kalman filters do not apply.

1.2 Inference in Hidden Markov Models as a Bayesian problem

Equations (1)-(2) can be seen a Bayesian model in which Eq. (1) defines the prior distribution of the process of interest $(X_t)_{t \geq 1}$ and Eq. (2) defines the likelihood function; that is:

$$p(x_{1:t}) = \mu(x_1) \prod_{k=2}^{t} f(x_k | x_{k-1}) \tag{3}$$

and

$$p(y_{1:t} | x_{1:t}) = \prod_{k=1}^{t} g(y_k | x_k), \tag{4}$$

where, for any sequence $(z_t)_{t\geq 1}$, and any $i \leq j$, we use the notation $z_{i:j} := (z_i, z_{i+1}, ..., z_j)$.

In such a Bayesian context, inference about $X_{1:t}$ given a realization of the observations $Y_{1:t} = y_{1:t}$ relies upon the posterior distribution

$$p\left(x_{1:t}\right| y_{1:t}) = \frac{p\left(x_{1:t}, y_{1:t}\right)}{p\left(y_{1:t}\right)}, \tag{5}$$

where

$$p\left(x_{1:t}, y_{1:t}\right) = p\left(x_{1:t}\right) p\left(y_{1:t}\right| x_{1:t}) , \tag{6}$$

$$\text{and } p\left(y_{1:t}\right) = \int_{\mathbb{X}^t} p\left(x_{1:t}, y_{1:t}\right) dx_{1:t}. \tag{7}$$

In the setting of HMMs, the distribution $p(x_t \mid y_{1:t})$ is called the filtering distribution, and the distribution $p\left(x_{1:t}\right| y_{1:t})$ is called the filtering distribution on the path space; the distribution of a past state x_s given $y_{1:t}$, where $s < t$, is called the smoothing distribution. The distribution of future states $p(x_{t+k} \mid y_{1:t})$, where $k \geq 1$, is called the prediction distribution.

Note that compared to the usual statistical setting, where the dimension of the unknown variable θ is fixed and independent of the numbers of observations n, here there is one new unknown variable X_t for each new observation y_t. This represents the difficulty of the problem: after t observations, there are t variables $X_1, \ldots, X_t$ to estimate.

1.3 Filtering recursions and marginal likelihood

We consider here the so-called problem of filtering: characterizing the distribution of the state X_t of the hidden Markov model at the current time t, given the information provided by all of the observations $y_1, \ldots, y_t$ received up to the current time. We first derive recursions, which express $p(x_t \mid y_{1:t})$ as an update of $p(x_{t-1} \mid y_{1:t-1})$, or similarly, $p(x_{1:t} \mid y_{1:t})$ as an update of $p(x_{1:t-1} \mid y_{1:t-1})$.

We recall that, following Eq. (1)-(2), the posterior distribution $p\left(x_{1:t}\right| y_{1:t})$ is defined by Eq. (5). The un-normalized posterior distribution $p\left(x_{1:t} \mid y_{1:t}\right)$ given in Eq. (5) satisfies

$$p\left(x_{1:t} \mid y_{1:t}\right) \propto p\left(x_{1:t-1}, y_{1:t-1}\right) f\left(x_t\right| x_{t-1}) \, g\left(y_t\right| x_t) \, . \tag{8}$$

Consequently, the posterior $p\left(x_{1:t}\right| y_{1:t})$ satisfies the following recursion

$$p\left(x_{1:t}\right| y_{1:t}) = p\left(x_{1:t-1}\right| y_{1:t-1}) \, \frac{f\left(x_t\right| x_{t-1}) \, g\left(y_t\right| x_t)}{p\left(y_t\right| y_{1:t-1})}, \tag{9}$$

where

$$p\left(y_t\right| y_{1:t-1}) = \int p\left(x_{t-1}\right| y_{1:t-1}) \, f\left(x_t\right| x_{t-1}) \, g\left(y_t\right| x_t) \, dx_{t-1:t} \tag{10}$$

In the literature, the recursion satisfied by the marginal distribution $p\left(x_t\right| y_{1:t})$ is often presented. It is straightforward to check (by integrating out $x_{1:t-1}$ in (9)) that we have

$$p\left(x_t\right| y_{1:t-1}) = \int f\left(x_t\right| x_{t-1}) \, p\left(x_{t-1}\right| y_{1:t-1}) \, dx_{t-1} \tag{11}$$

and

$$p\left(x_t\right| y_{1:t}) = \frac{g\left(y_t\right| x_t) \, p\left(x_t\right| y_{1:t-1})}{p\left(y_t\right| y_{1:t-1})}. \tag{12}$$

Equation (11) is known as the prediction step and (12) is known as the updating step.

If we can compute $p\left(x_{1:t}\right| y_{1:t})$ and thus $p\left(x_t\right| y_{1:t})$ sequentially, then the quantity $p\left(y_{1:t}\right)$, which is known as the marginal likelihood, can also clearly be evaluated recursively using

$$p\left(y_{1:t}\right) = p\left(y_1\right) \prod_{k=2}^{t} p\left(y_k\right| y_{1:k-1}) \tag{13}$$

where $p\left(y_k\right| y_{1:k-1})$ is of the form (10).

For a finite state-space model as in Example 1, the integrals correspond to finite sums and (11)-(12) read as follows

$$\begin{aligned}\mathbb{P}\left(X_t = j \middle| y_{1:t-1}\right) &= \sum_{i=1}^{m} \mathbb{P}\left(X_t = j \middle| X_{t-1} = i\right) \mathbb{P}\left(X_{t-1} = i \middle| y_{1:t-1}\right) \\ &= \sum_{i=1}^{m} f\left(j \middle| i\right) \mathbb{P}\left(X_{t-1} = i \middle| y_{1:t-1}\right)\end{aligned}$$

and

$$\mathbb{P}\left(X_t = j \middle| y_{1:t}\right) = \frac{g\left(y_t \middle| X_t = j\right) \mathbb{P}\left(X_t = j \middle| y_{1:t-1}\right)}{p\left(y_t \middle| y_{1:t-1}\right)}$$

where

$$p\left(y_t \middle| y_{1:t-1}\right) = \sum_{i=1}^{m} g\left(y_t \middle| X_t = i\right) \mathbb{P}\left(X_t = i \middle| y_{1:t-1}\right).$$

All these quantities can be computed exactly. In a linear Gaussian model as in Example 2, the recursion can also be worked out exactly, which is the basis of the Kalman filter (see the corresponding exercise of Problem Sheet 6).

1.4 MCMC strategies

In the case of non-linear or/and non-Gaussian hidden Markov models, there are no analytic form for Eq. (11) and Eq. (12). In other words, even if we start from a simple parametric initial distribution μ for the first state X_1, there are no explicit formula giving the distribution of X_1 given y_1, the distribution of X_2 given y_1, the distribution of X_2 given y_1, y_2, etc. We thus need Monte Carlo methods to approximate the distribution $p(x_{1:t} \mid y_{1:t})$ (or the distributions $p(x_t \mid y_{1:t})$).

The most naive strategy would be to run a Metropolis-Hastings algorithm on the space $\mathbb{X}^t$. Starting from a path $x_{1:t}^{(0)}$, a candidate would be proposed from a global proposal distribution q on $\mathbb{X}^t$, or from a local proposal distribution $q(x' \mid x)$. The candidate would be accepted or not according to an acceptance ratio involving the target distribution

$$p(x_{1:t} \mid y_{1:t}) \propto \mu\left(x_1\right) \prod_{k=2}^{t} f\left(x_k \middle| x_{k-1}\right) \prod_{k=1}^{t} g\left(y_k \middle| x_k\right).$$

Given the dimension of the space, that is $t \times \dim\left(\mathbb{X}\right)$, it is most likely imposible to design a good proposal distribution q, and thus the resulting MCMC algorithm will converge slowly.

A Gibbs sampling strategy can also be implemented. It consists in sampling alternatively x_k given x_{-k} and $y_{1:t}$, for each $k \in \{1, \ldots, t\}$. The conditional independencies (that can be inferred from Figure (1)) imply

$$\begin{aligned}p(x_k \mid x_{-k}, y_{1:t}) &\propto p(x_k \mid x_{k-1}) p(x_k \mid y_k) p(x_{k+1} \mid x_k) \\ &= f(x_k \mid x_{k-1}) g(x_k \mid y_k) f(x_{k+1} \mid x_k)\end{aligned}$$

if $k \in \{2, \ldots, t-1\}$; the case $k = 1$ and $k = t$ can be worked out similarly. Thus, sampling from this conditional distribution can be envisioned, for instance using a Metropolis-Hastings step (which is now only on a space of dimension $\dim\left(\mathbb{X}\right)$ at each step). The Gibbs sampling approach has the benefit of breaking the high dimensional sampling problem into t smaller problems. However, Gibbs sampling typically converges slowly when the variables are highly correlated (remember the example of correlated bivariate normal distributions in Chapter 5). By the nature of hidden Markov models, we can expect each X_k to be strongly correlated with its neighbours X_{k-1} and X_{k+1}. Thus Gibbs sampling approaches typically perform poorly to sample from $p(x_{1:t} \mid y_{1:t})$, although historically they have been extensively used for this purpose.

Note also that MCMC strategies are not very convenient in the context of time series: every time a new observation y_t arrives, the whole MCMC algorithm has to be run conditional upon the whole dataset $y_{1:t}$. If the interest mainly lies in estimating the hidden state X_t given $y_{1:t}$ (aka the filtering problem), and not in the whole paths $X_{1:t}$, then it would seem more efficient to rely on the recursions of Eq. (11) and Eq. (12). This way we could update the past “knowledge” $p(x_{t-1} \mid y_{1:t-1})$, instead of starting from scratch every time.

2 Sequential Importance Sampling

SMC methods are a general class of Monte Carlo methods that allow us to sample approximately sequentially from the sequence of target posterior probability densities $\left\{p\left(x_{1:t}\middle| y_{1:t}\right)\right\}_{t\geq 1}$ and allows us to simultaneously approximate the sequence of marginal likelihoods $\left\{p\left(y_{1:t}\right)\right\}_{t\geq 1}$. At (algorithmic) time 1, we approximate $p\left(x_1\middle| y_1\right)$ and $p\left(y_1\right)$, then at time 2 we approximate $p\left(x_{1:2}\middle| y_{1:2}\right)$ and $p\left(y_{1:2}\right)$, etc.

The main building block of SMC methods is importance sampling, or more exactly, a sequential version of importance sampling described below. We first describe how importance sampling can be used to approximate the first filtering distribution, and then how the approximation can be sequentially updated upon the arrival of new observations.

2.1 Importance Sampling

Let us consider the problem of approximating the first filtering distribution $p(x_1 \mid y_1)$. We have

$$p(x_1 \mid y_1) = \frac{\mu(x_1)g(y_1 \mid x_1)}{\int_{\mathbb{X}} \mu(x_1)g(y_1 \mid x_1)dx_1} \propto \mu(x_1)g(y_1 \mid x_1).$$

Introduce a importance proposal distribution q_1 on $\mathbb{X}$, such that $\operatorname{supp} p(x_1 \mid y_1) \subset \operatorname{supp} q_1(x_1)$, i.e. $\forall x_1 \quad p\left(x_1\middle| y_1\right) > 0 \Rightarrow q_1\left(x_1\right) > 0$. By sampling $X_1^1, \dots, X_1^N \overset{\text{i.i.d}}{\sim} q_1$, we can compute

$$\forall i \in \{1, \dots, N\} \quad w_1^i = \frac{\mu(X_1^i)g(y_1 \mid X_1^i)}{q_1\left(X_1^i\right)}.$$

Then the normalized importance sampling (NIS) approximation of $p(x_1 \mid y_1)$ is given by the empirical distribution $\pi_1^N(x_1)$ defined as

$$\pi_1^N(x_1) = \frac{\sum_{i=1}^N w_1^i \delta_{X_1^i}(x_1)}{\sum_{j=1}^N w_1^j} = \sum_{i=1}^N W_1^i \delta_{X_1^i}(x_1),$$

where $W_1^i = w_1^i / \sum_{j=1}^N w_1^j$. It is an IS approximation of $p(x_1 \mid y_1)$ in the sense that for any test function φ_1 on $\mathbb{X}$,

$$I^N\left(\varphi_1\right) = \int \varphi_1(x)\pi_1^N(x_1)dx_1 = \sum_{i=1}^N W_1^i \varphi_1(X_1^i) \xrightarrow[N\to\infty]{a.s.} \int \varphi_1(x)p(x_1 \mid y_1)dx,$$

as described in the lecture notes on Importance Sampling. Note also that an estimator of the marginal likelihood $p(y_1) = \int_{\mathbb{X}} \mu(x_1)g(y_1 \mid x_1)dx_1$ is given by $p^N(y_1) = N^{-1}\sum_{i=1}^N w_1^i$, by a standard importance sampling argument:

$$p^N(y_1) = \frac{1}{N}\sum_{i=1}^N w_1^i = \frac{1}{N}\sum_{i=1}^N \frac{\mu(X_1^i)g(y_1 \mid X_1^i)}{q_1\left(X_1^i\right)} \xrightarrow[N\to\infty]{a.s.} \int \frac{\mu(x_1)g(y_1 \mid x_1)}{q_1\left(x_1\right)} q_1\left(x_1\right)dx_1 = p(y_1).$$

For a given test function φ_1, we have seen already that there is an expression of the proposal distribution minimizing the asymptotic variance of $I^N\left(\varphi_1\right)$. The expression is of little practical interest, because the optimal proposal involves intractable calculations (we come back to it later though). Another way to select a proposal distribution consists in assessing the effective sample size. It is defined as

$$\text{ESS} = \frac{\left(\sum_{i=1}^N w_1^i\right)^2}{\left(\sum_{i=1}^N w_1^i\right)} = \frac{1}{\sum_{i=1}^n \left(W_1^i\right)^2}.$$

One can check that $1 \leq \text{ESS} \leq N$. We have $\text{ESS} = N$ if $W_1^i = N^{-1}$ for all i; i.e. if $q_1\left(x_1\right) = p\left(x_1\middle| y_1\right)$. If we have a very poor proposal distribution, then there will exist i such that $W_1^i \approx 1$, and for all the remaining indexes $j \neq i$, $W_1^j \approx 0$. Then the ESS will be close to 1. As a rule of thumb, the higher the ESS the better our approximation, and this can be used to select among multiple proposal distributions.

2.2 Sequential Importance Sampling

Once the first filtering distribution $p(x_1 \mid y_1)$ has been approximated by an empirical distribution such as $\pi_1^N(x_1)$, the next question is: how to approximate $p(x_{1:2} \mid y_{1:2})$, and $p(x_2 \mid y_{1:2})$ and $p(y_{1:2})$ as by-products. At step t, assume that we have obtained an approximation π_{t-1}^N of $p(x_{1:t-1} \mid y_{1:t-1})$, made of N trajectories $X_{1:t-1}^i$ sampled from q_{t-1} and with associated weights $w_{t-1}^i \propto p(X_{1:t-1}^i \mid y_{1:t-1})/q_{t-1}(X_{1:t-1}^i)$. Let us follow Eq. (11) and Eq. (12) to obtain the approximation π_t^N of $p(x_{1:t} \mid y_{1:t-1})$.

Introduce a proposal distribution $q_{t|t-1}(x_t \mid x_{t-1})$, that is, a distribution indexed by x_{t-1}. If for each $i \in \{1, \ldots, N\}$ we draw $X_t^i \sim q_{t|t-1}(x_t \mid X_{t-1}^i)$, then the trajectory $X_{1:t}^i = (X_{1:t-1}^i, X_t^i)$ follows $q_t(x_{1:t})$ defined as $q_{t-1}(x_{1:t-1})q_{t|t-1}(x_t \mid x_{t-1})$ (this follows simply from the "sampling via composition" argument described in Chapter 2). The importance weight function is then defined as

$$\begin{aligned} w(x_{1:t}) &\propto \frac{p(x_{1:t} \mid y_{1:t})}{q_t(x_{1:t})} \propto \frac{p\left(x_{1:t-1}\middle| y_{1:t-1}\right) f\left(x_t\middle| x_{t-1}\right) g\left(y_t\middle| x_t\right)}{q_{t-1}(x_{1:t-1})q_{t|t-1}(x_t \mid x_{t-1})} \\ &\propto w(x_{1:t-1}) \frac{f\left(x_t\middle| x_{t-1}\right) g\left(y_t\middle| x_t\right)}{q_{t|t-1}(x_t \mid x_{t-1})}. \end{aligned}$$

Thus given the previous weights w_{t-1}^i for all i, we multiply them by

$$\omega_t^i := \omega_t\left(X_{t-1}^i, X_t^i\right) := \frac{f\left(X_t^i\middle| X_{t-1}^i\right) g\left(y_t\middle| X_t^i\right)}{q_{t|t-1}(X_t^i \mid X_{t-1}^i)}$$

to obtain the new weights. The terms ω_t^i are thus called the incremental weights:

$$w_t^i = w_{t-1}^i \times \omega_t^i.$$

Then the "particles" $\left(w_t^i, X_{1:t}^i\right)_{i=1}^N$ form an approximation π_t^N of $p(x_{1:t} \mid y_{1:t})$, in the sense that for any test function φ_t on $\mathbb{X}^t$,

$$I^N\left(\varphi_t\right) = \int \varphi_t(x_{1:t})\pi_t^N(x_{1:t})dx_{1:t} = \frac{\sum_{i=1}^N w_t^i \varphi_t(X_{1:t}^i)}{\sum_{i=1}^N w_t^i} \xrightarrow[N\to\infty]{a.s.} \int \varphi_t(x_{1:t})p(x_{1:t} \mid y_{1:t})dx_{1:t}. \tag{14}$$

Again this is simple importance sampling using q_t to target $p(x_{1:t} \mid y_{1:t})$. Additionnally, we have an estimate of

$$\begin{aligned} p(y_t \mid y_{1:t-1}) &= \int f\left(x_t\middle| x_{t-1}\right) g\left(y_t\middle| x_t\right) p(x_{1:t-1} \mid y_{1:t-1}) dx_{1:t-1}dx_t \\ &= \int \frac{f\left(x_t\middle| x_{t-1}\right) g\left(y_t\middle| x_t\right)}{q_{t|t-1}(x_t \mid x_{t-1})} \frac{p(x_{1:t-1} \mid y_{1:t-1})}{q_{t-1}(x_{1:t-1})} q_t\left(x_{1:t}\right) dx_{1:t-1}dx_t \end{aligned}$$

using importance sampling, e.g.

$$p^N(y_t \mid y_{1:t-1}) = \frac{\sum_{i=1}^N w_{t-1}^i \omega_t^i}{\sum_{i=1}^N w_{t-1}^i}. \tag{15}$$

Thus we can obtain an estimate of the marginal likelihood as

$$p^N\left(y_{1:t}\right) = p^N\left(y_1\right) \prod_{k=2}^T p^N\left(y_k \mid y_{1:k-1}\right). \tag{16}$$

Note that from the path approximation π_t^N, we can of course retain only the last components $\left(X_t^i\right)_{i=1}^N$ to approximate the filtering distribution $p(x_t \mid y_{1:t})$:

$$p^N(x_t \mid y_{1:t}) = \frac{\sum_{i=1}^N w_t^i \delta_{X_t^i}(x_t)}{\sum_{i=1}^N w_t^i} \approx_{N\to\infty} p(x_t \mid y_{1:t}).$$

The Sequential Importance Sampling (SIS) algorithm proceeds as in Algorithm 1, with each step carried out for each $i = 1, \ldots, N$. Note that the N particles can be propagated and weighted in parallel.

Algorithm 1 Sequential Importance Sampling

```
At time t = 1
• Sample $X_1^i \sim q_1(\cdot)$.
• Compute the weights
      $$w_1^i = \frac{\mu(X_1^i) g(y_1 \mid X_1^i)}{q_1\left(X_1^i\right)}.$$
At time t ≥ 2
• Sample $X_t^i \sim q_{t|t-1}(\cdot | X_{t-1}^i)$.
• Compute the weights
      $$\begin{aligned} w_t^i &= w_{t-1}^i \times \omega_t^i \\ &= w_{t-1}^i \times \frac{f\left(X_t^i \middle| X_{t-1}^i\right) g\left(y_t \middle| X_t^i\right)}{q_{t|t-1}(X_t^i \mid X_{t-1}^i)}. \end{aligned}$$
```

2.3 Choosing proposal distributions

2.3.1 Prior proposal

In many settings, the default choice for the proposal distributions q_1 and $q_{t|t-1}$ is to use μ and f, i.e. the model distributions. This simplifies the form of the weight functions as follows:

$$\begin{aligned} w(x_1) &= \frac{\mu(x_1) g(y_1 \mid x_1)}{\mu\left(x_1\right)} = g(y_1 \mid x_1) \\ \forall t \geq 2 \quad w(x_{1:t}) &= w(x_{1:t-1}) \frac{f(x_t \mid x_{t-1}) g\left(y_t \middle| x_t\right)}{f(x_t \mid x_{t-1})} = w(x_{1:t-1}) g(y_t \mid x_t). \end{aligned}$$

Thus, in this case the trajectories $X_{1:t}^i$ are drawn from $p(x_{1:t}) = \mu(x_1) \prod_{k=2}^t f(x_k \mid x_{k-1})$ and the weight function is simply $w(x_{1:t}) = p(y_{1:t} \mid x_{1:t}) = \prod_{k=1}^t g(y_k \mid x_k)$, as in Eq. 3 and Eq. 4. This proposal performs a "blind" exploration of the state space: the particle X_t^i is drawn from $f(x_t \mid X_{t-1}^i)$, irrespective of the observation y_t.

Beyond its simplicity, an advantage of this simple choice is that $q_{t|t-1}$ cancels out f in the calculation of the weight function. Thus, in cases where $f(x_t \mid x_{t-1})$ can be sampled from but not evaluated point-wise, choosing $q_{t|t-1} = f$ is the only viable option.

2.3.2 Locally optimal proposal

A sensible approach consists in selecting a proposal $q_{t|t-1}\left(x_t \middle| x_{t-1}\right)$ that minimizes the variance of the incremental weights $(\omega_t^i)_{i=1}^N$.

Proposition. The proposal $q_{t|t-1}\left(x_t \middle| x_{t-1}\right)$ minimizing $\mathbb{V}_{q_t(x_{1:t})}\left(w\left(X_{1:t}\right)\right)$ is given by

$$q_{t|t-1}^{\text{opt}}\left(x_t \middle| x_{t-1}\right) = \frac{f\left(x_t \middle| x_{t-1}\right) g\left(y_t \middle| x_t\right)}{p\left(y_t \middle| x_{t-1}\right)} \tag{17}$$

and the associated incremental weight is given by

$$\omega_t^{\text{opt}}\left(x_{t-1}, x_t\right) = p\left(y_t \middle| x_{t-1}\right).$$

Proof. We have by the variance decomposition formula

$$\begin{aligned} \mathbb{V}_{q_t(x_{1:t})}\left(w\left(X_{1:t}\right)\right) = & \mathbb{V}_{q_{t-1}(x_{1:t-1})}\left[\mathbb{E}_{q_{t|t-1}(x_t|x_{t-1})}\left[w\left(X_{1:t}\right)\middle| X_{1:t-1}\right]\right] \\ & + \mathbb{E}_{q_{t-1}(x_{1:t-1})}\left[\mathbb{V}_{q_{t|t-1}(x_t|x_{t-1})}\left[w\left(X_{1:t}\right)\middle| X_{1:t-1}\right]\right] \end{aligned}$$

but

$$\mathbb{E}_{q_{t|t-1}(x_t|x_{t-1})}\left[w\left(X_{1:t}\right)\middle| X_{1:t-1}\right] = w_{t-1}\left(X_{1:t-1}\right) p\left(y_t \middle| X_{t-1}\right)$$

is independent of $q_{t|t-1}\left(x_t|\, x_{t-1}\right)$. So minimizing $\mathbb{V}_{q_t(x_{1:t})}\left(w\left(X_{1:t}\right)\right)$ w.r.t $q_{t|t-1}\left(x_t|\, x_{t-1}\right)$ is equivalent to minimizing $\mathbb{V}_{q_t(\left.x_t\right|x_{t-1})}\left\{\left.w\left(X_{1:t}\right)\right|X_{1:t-1}\right\}$ w.r.t $q_{t|t-1}\left(x_t|\, x_{t-1}\right)$. This is achieved for $q^{\text{opt}}_{t|t-1}\left(x_t|\, x_{t-1}\right)$ as given in Eq. (17), because

$$\begin{aligned}\frac{\mathbb{V}_{q^{\text{opt}}_{t|t-1}(\left.x_t\right|x_{t-1})}\left\{\left.w\left(X_{1:t}\right)\right|X_{1:t-1}\right\}}{w^2\left(X_{1:t-1}\right)} &= \int\left\{\frac{f\left(x_t|\, x_{t-1}\right)g\left(y_t|\, x_t\right)}{q^{\text{opt}}_{t|t-1}\left(x_t|\, x_{t-1}\right)}\right\}^2 q^{\text{opt}}_{t|t-1}\left(x_t|\, x_{t-1}\right)dx_t \\ &-\left\{\int\frac{f\left(x_t|\, x_{t-1}\right)g\left(y_t|\, x_t\right)}{q^{\text{opt}}_{t|t-1}\left(x_t|\, x_{t-1}\right)}q^{\text{opt}}_{t|t-1}\left(x_t|\, x_{t-1}\right)dx_t\right\}^2 \\ &= p^2\left(y_t|\, x_{t-1}\right)-p^2\left(y_t|\, x_{t-1}\right)=0.\end{aligned}$$

Hence the result follows. ■

Example 4 *Consider the following nonlinear model*

$$f\left(x'|\, x\right)=\mathcal{N}\left(x';\varphi\left(x\right),\sigma_V^2\right),\ g\left(y|\, x\right)=\mathcal{N}\left(y;x,\sigma_W^2\right)$$

then we have

$$q^{opt}_{t|t-1}\left(x_t|\, x_{t-1}\right)=\mathcal{N}\left(x_t;\mu\left(x_{t-1}\right),\sigma^2\left(x_{t-1}\right)\right)$$

where

$$\begin{aligned}\sigma^2\left(x_{t-1}\right) &= \frac{\sigma_V^2\sigma_W^2}{\sigma_V^2+\sigma_W^2}, \\ \mu\left(x_{t-1}\right) &= \sigma^2\left(x_{t-1}\right)\left[\frac{\varphi\left(x_{t-1}\right)}{\sigma_V^2}+\frac{y_t}{\sigma_W^2}\right], \\ \omega_t^{opt}\left(x_{t-1},x_t\right) &= \frac{1}{\sqrt{2\pi\left(\sigma_V^2+\sigma_W^2\right)}}\exp\left(-\frac{\left(y_t-\varphi\left(x_{t-1}\right)\right)^2}{2\left(\sigma_V^2+\sigma_W^2\right)}\right).\end{aligned}$$

The benefits of using the locally optimal proposal compared to the prior proposal will be significant if the observation y_t is very informative; i.e. if $g\left(y_t|\, x_t\right)$ is "peaky". Indeed, we see that the optimal proposal uses y_t to propagate each X^i_{t-1} to X^i_t. Note that this optimization is only performed for one time step, which is why it is called "locally" optimal. Thus, it does *not* correspond to finding the optimal proposal $q_t(x_{1:t})$ on the path space, i.e. it does *not* correspond to minimizing the variance of weights $w(x_{1:t})$ with respect to $q_t(x_{1:t})$.

2.3.3 Approximation to the locally optimal proposal

Practically it might be impossible to sample from $q^{\text{opt}}_{t|t-1}\left(x_t|\, x_{t-1}\right)$ and/or to compute $\omega_t^{\text{opt}}\left(x_{t-1},x_t\right)$. However, we can perhaps use this distribution as a guideline to construct approximations.

Assume for example that we have

$$f\left(x'|\, x\right)=\mathcal{N}\left(x';\varphi\left(x\right),\sigma_V^2\right),\ g\left(y|\, x\right)=\mathcal{N}\left(y;\zeta\left(x\right),\sigma_W^2\right)$$

for $\varphi:\mathbb{R}\rightarrow\mathbb{R}$ and $\zeta:\mathbb{R}\rightarrow\mathbb{R}$ some nonlinear functions. It would be possible to sample from $q^{\text{opt}}_{t|t-1}\left(x_t|\, x_{t-1}\right)=\frac{f\left(x_t|x_{t-1}\right)g\left(y_t|x_t\right)}{p\left(y_t|x_{t-1}\right)}$ using rejection sampling, by proposing from $X\sim f\left(x_t|\, x_{t-1}\right)$ and accepting the proposal with probability

$$\frac{\mathcal{N}\left(y_t;\zeta\left(X\right),\sigma_W^2\right)}{1/\sqrt{2\pi\sigma_W^2}}$$

as $\sup_x g\left(y_t|\, x\right)\leq 1/\sqrt{2\pi\sigma_W^2}$. However, this does not help us as we do not necessarily know how to compute analytically the associated incremental importance weight

$$\omega_t^{\text{opt}}\left(x_{t-1},x_t\right)=\int\mathcal{N}\left(x;\varphi\left(x_{t-1}\right),\sigma_V^2\right)\mathcal{N}\left(y_t;\zeta\left(x\right),\sigma_W^2\right)dx.$$

A sensible alternative consists in approximating $g\left(y_t|\, x\right)$ using a local linearization approach:

$$\zeta\left(x\right)\approx\zeta\left(\varphi\left(x_{t-1}\right)\right)+\left.\zeta'\right|_{\varphi\left(x_{t-1}\right)}\left(x-\varphi\left(x_{t-1}\right)\right),$$

which suggests the following Gaussian approximation $\widehat{g}\left(y|\,x\right)$ to $g\left(y|\,x\right)$:

$$\widehat{g}\left(y|\,x\right)=\mathcal{N}\left(y;\underbrace{\zeta\left(\varphi\left(x_{t-1}\right)\right)-\left.\frac{d\zeta}{dx}\right|_{\varphi\left(x_{t-1}\right)}\varphi\left(x_{t-1}\right)}_{m\left(x_{t-1}\right)}+\underbrace{\left.\frac{d\zeta}{dx}\right|_{\varphi\left(x_{t-1}\right)}}_{\beta\left(x_{t-1}\right)}x,\sigma_W^2\right).$$

It is a "local" approximation in the sense that it depends on x_{t-1}: for each sample X_{t-1}^i, it will build a different Gaussian approximation.

This Gaussian approximation suggests an IS proposal

$$q_{t|t-1}\left(x_t|\,x_{t-1}\right)=\frac{f\left(x_t|\,x_{t-1}\right)\widehat{g}\left(y_t|\,x_t\right)}{\int f\left(x_t'|\,x_{t-1}\right)\widehat{g}\left(y_t|\,x_t'\right)dx_t'}=\mathcal{N}\left(x_t;\mu\left(x_{t-1}\right),\sigma^2\left(x_{t-1}\right)\right)$$

where

$$\begin{aligned}\frac{1}{\sigma^2\left(x_{t-1}\right)}&=\frac{1}{\sigma_V^2}+\frac{\beta^2\left(x_{t-1}\right)}{\sigma_W^2},\\ \mu\left(x_{t-1}\right)&=\sigma^2\left(x_{t-1}\right)\left[\frac{\varphi\left(x_{t-1}\right)}{\sigma_V^2}+\frac{\beta\left(x_{t-1}\right)\left(y_t-m\left(x_{t-1}\right)\right)}{\sigma_W^2}\right]\end{aligned}$$

which is a normal distribution from which one can easily sample. The associated incremental importance weight can also be computed analytically:

$$\begin{aligned}\omega_t\left(x_{t-1},x_t\right)&=\frac{f\left(x_t|\,x_{t-1}\right)g\left(y_t|\,x_t\right)}{q_{t|t-1}\left(x_t|\,x_{t-1}\right)}\\&=\frac{\mathcal{N}\left(x_t;\varphi\left(x_{t-1}\right),\sigma_V^2\right)\mathcal{N}\left(y_t;\zeta\left(x_t\right),\sigma_W^2\right)}{\mathcal{N}\left(x_t;\mu\left(x_{t-1}\right),\sigma^2\left(x_{t-1}\right)\right)}.\end{aligned}$$

3 Linear Gaussian example

3.1 Empirical performance

We illustrate the performance of SIS on a simple linear Gaussian model; i.e.

$$\forall t\geq 1\quad X_t=\phi X_{t-1}+\sigma_V V_t, \tag{18}$$

$$\forall t\geq 1\quad Y_t=X_t+\sigma_V W_t, \tag{19}$$

with $X_0\sim\mathcal{N}\left(0,1\right)$, $V_t,W_t\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(0,1\right)$, $\phi=0.95$, $\sigma_V=1$, $\sigma_W=1$. In this example, we can compute all the quantities of interest exactly, using the Kalman filter. Thus it is possible to assess the performance of Monte Carlo methods. We simulate $T=100$ observations from this model. The generated data is plotted on Figure 2, along with the means of the filtering distributions $p(x_t\mid y_{1:t})$, calculated using the Kalman filter.

We propose to estimate the filtering means using sequential importance sampling, with either the prior proposal or the optimal proposal. We observe the evolution of the ESS over time when using the prior proposal and the locally optimal proposal within the SIS procedure based on $N=1000$ particles, see Figure 3. We see that the ESS quickly goes down to 1, thus we do not expect the sequential importance sampling method to perform well in this scenario. The optimal proposal seems to result in a slower decay of the ESS, but it still reaches nearly 1 after 25 steps.

Figure 4 shows the estimation results: SIS is used to estimate the filtering means $\mathbb{E}\left(x_t\mid y_{1:t}\right)$ and the filtering variances $\mathbb{V}\left(x_t\mid y_{1:t}\right)$, for all $t\geq 1$. The results are compared with the exact means and variances computed using the Kalman filter. We see that the optimal proposal allows to keep track of the filtering means, even though the ESS is very low. However, the variances is poorly estimated, for both proposals.

Finally, we estimate the log likelihood $\log p(y_{1:t})$ for all t. The results are shown in Figure 5. We see that the optimal proposal manages to estimate the log-likelihood fairly well, whereas the prior proposal completely fails after about 35 time steps.

This bad performance of SIS should not be a surprise. We have seen previously on a toy example that IS typically scales exponentially with the dimension of the target distribution. SIS is nothing but a special

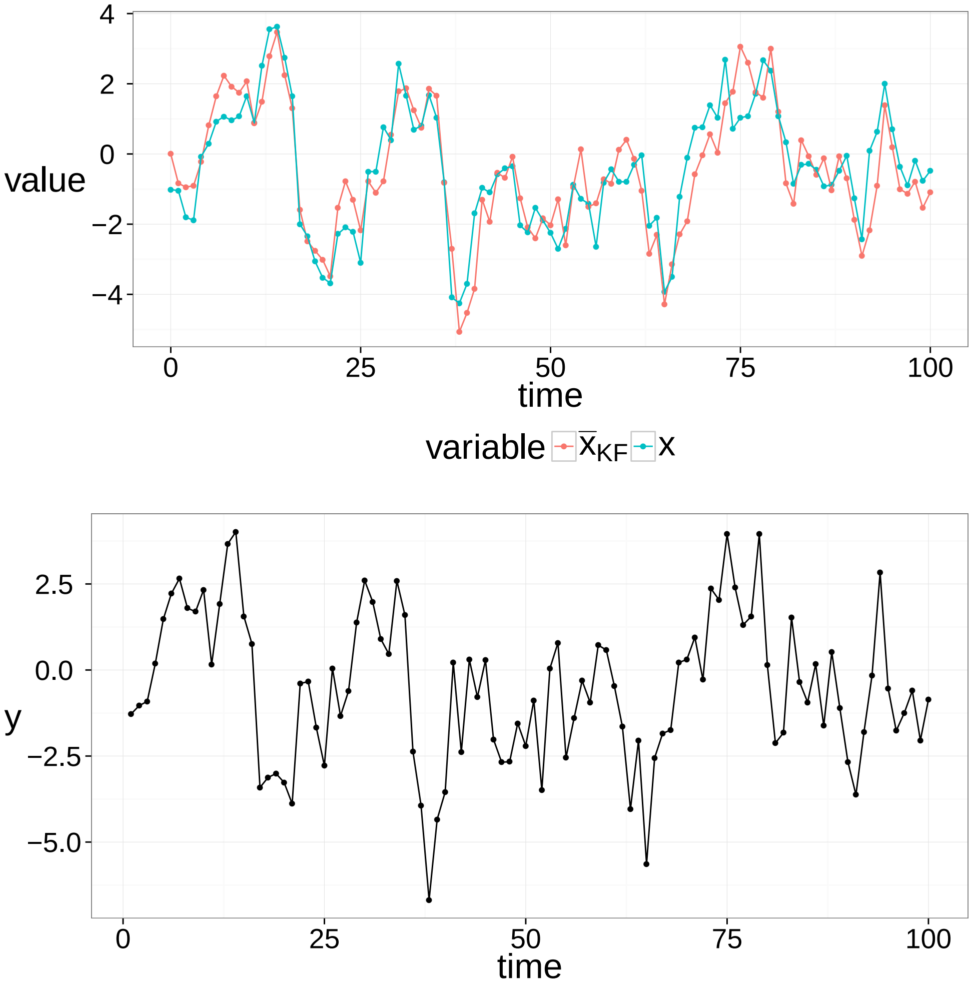

Figure 2: Synthetic dataset from the linear Gaussian model. Top: generated hidden process $(X_t)_{t\geq 1}$, along with the filtering mean calculated using the Kalman filter, $\left(\bar{X}_{KF}\right)$. Bottom: generated observations $(Y_t)_{t\geq 1}$.

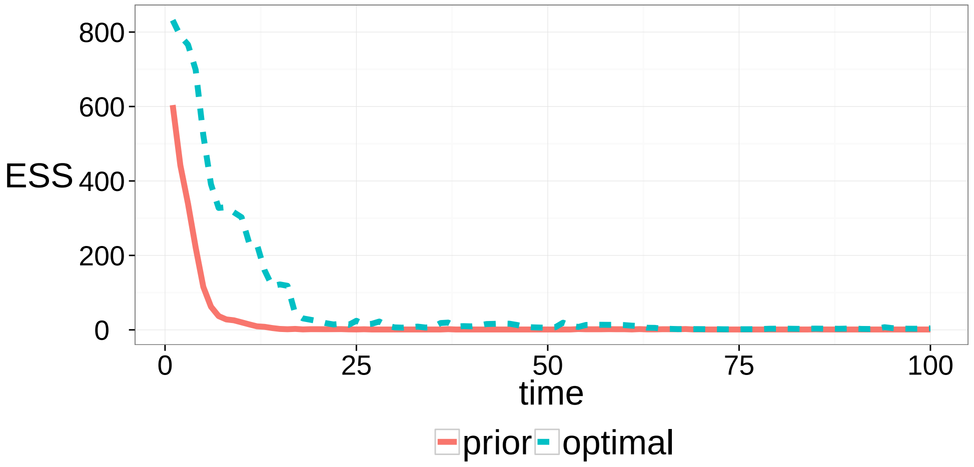

Figure 3: Evolution over time of the Effective Sample Size (ESS) using Sequential Importance Sampling, with the prior proposal and the optimal proposal. Here $N = 1000$, so the ESS is between 1 and 1000.

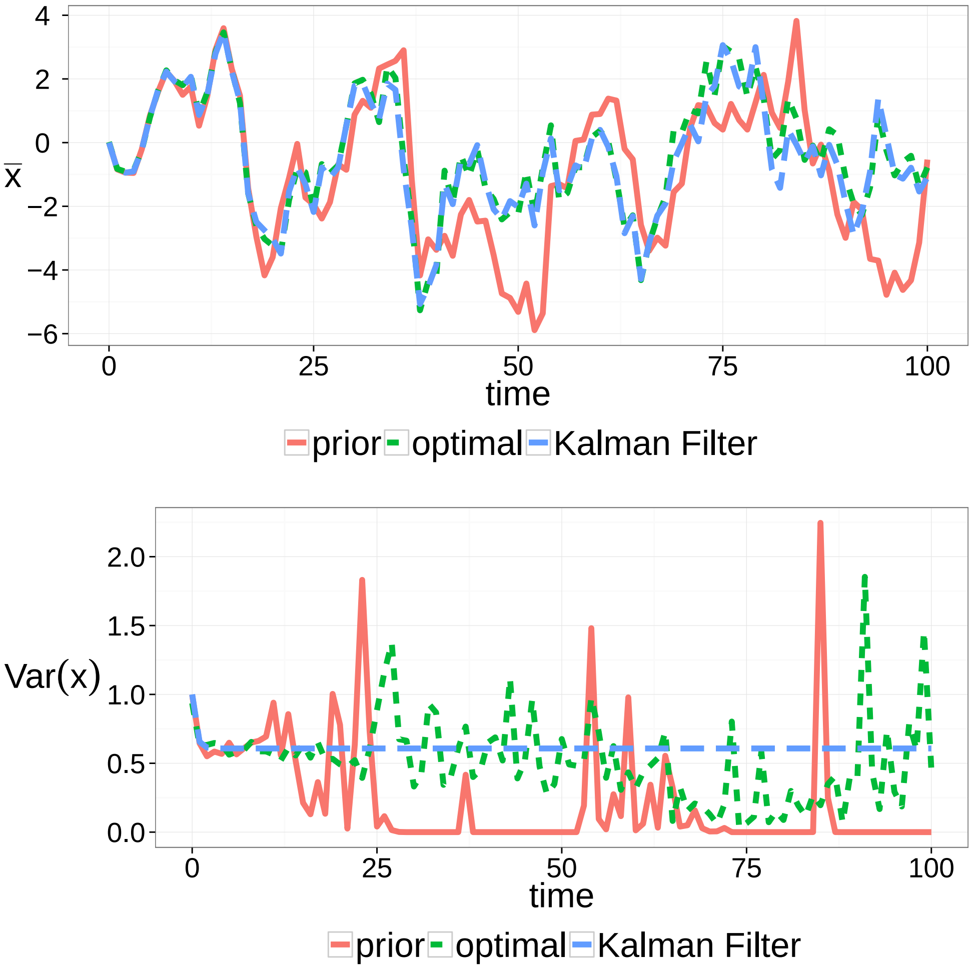

Figure 4: Estimation of the filtering means $\mathbb{E}\left(x_t \mid y_{1:t}\right)$ and the filtering variances $\mathbb{V}\left(x_t \mid y_{1:t}\right)$, using Sequential Importance Sampling, compared to the exact values calculated with the Kalman filter. The optimal proposal SIS manages to keep track of the filtering means, while the prior proposal fails completely. Both proposals fail to keep track of the filtering variances.

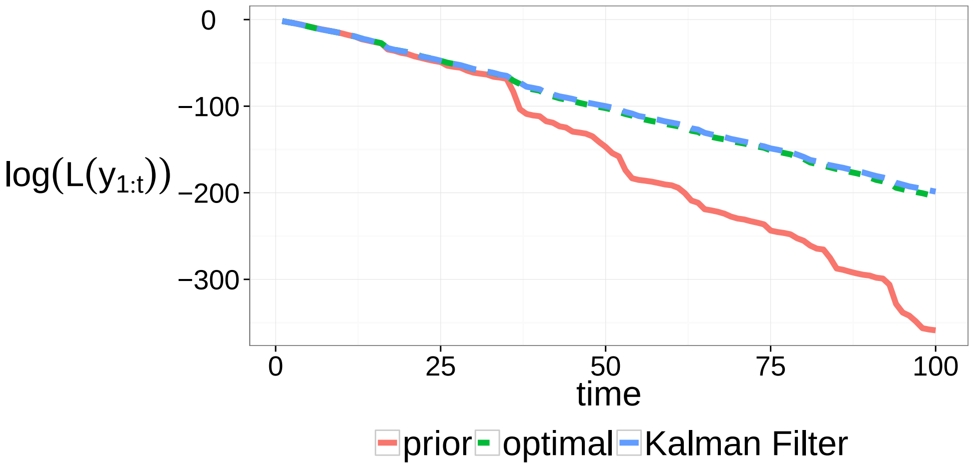

Figure 5: Estimation of the log likelihood $\log p(y_{1:t})$ using Sequential Importance Sampling, compared to the exact values computed with the Kalman filter.

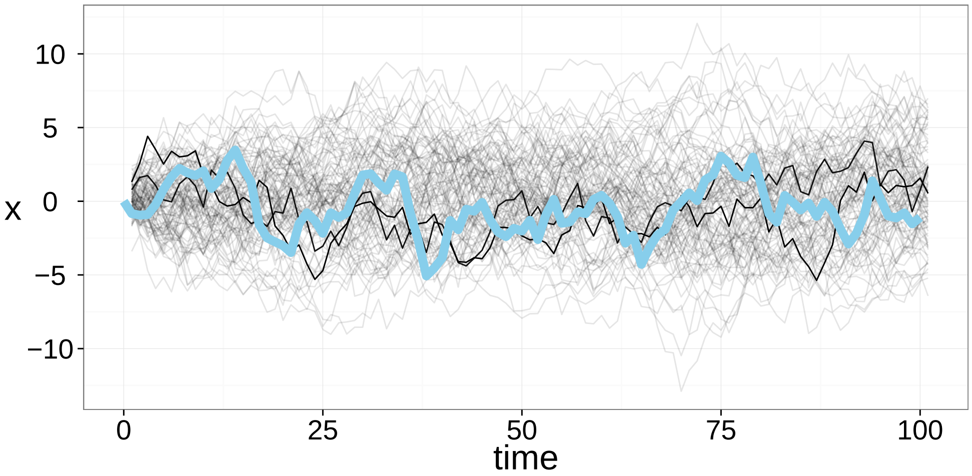

Figure 6: A hundred paths drawn from the prior proposal (black lines), and the filtering means calculated by Kalman filter (thick blue line). The intensity of black corresponds to the weight of each path: we see that only two paths carry most of the weights.

case of IS, so we cannot expect it to do much better than IS. To illustrate how the proposal distribution deteriorates with the number of time steps, 100 paths from the prior proposal are shown on Figure 6. The colours indicate the associated weights, and we can see that only two paths have significant weights. As the number of time steps grows, the weights degenerate until only one path has a significant weight.

3.2 A simple convergence result

The SIS estimates $p^N(y_{1:t})$ and $I^N(\varphi_t)$, of the marginal likelihood $p(y_{1:t})$ and of the filtering quantity $I(\varphi_t)$ respectively, satisfy central limit theorems, with convergence rate $1/\sqrt{N}$ as in the standard Monte Carlo setting. Their respective asymptotic variances are given by

$$\int \frac{p^2(x_{1:t}|\, y_{1:t})}{q_t(x_{1:t})} dx_{1:t} \tag{20}$$

and

$$\int \frac{p^2(x_{1:t}|\, y_{1:t})}{q_t(x_{1:t})} \left(\varphi_t(x_{1:t}) - I(\varphi_t)\right)^2 dx_{1:t}. \tag{21}$$

We provide here a very simple example showing analytically that SIS estimators have a variance increasing exponentially fast with t. We revisit the previous toy example defined by Eq. (18)-(19), except that we make it even simpler by selecting now $\phi = 0$ and that the observed sequence of observations is $y_1 = y_2 = \ldots = y_T = 0$. Using $\phi = 0$ ensures that $X_t \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \sigma_V^2\right)$ so

$$p(x_{1:t}, y_{1:t}) = \frac{1}{(2\pi)^t (\sigma_V \sigma_W)^t} \exp\left(-\frac{\sum_{k=1}^t x_k^2}{2}\left(\frac{1}{\sigma_W^2} + \frac{1}{\sigma_V^2}\right)\right)$$

and by integrating out $x_{1:t}$

$$p(y_{1:t}) = \frac{1}{\left(2\pi\left(\sigma_V^2 + \sigma_W^2\right)\right)^{t/2}}.$$

If we use the prior as a proposal then the variance of the log-weights is

$$\mathbb{V}_{p(x_{1:t})}\left[\log w_t(X_{1:T})\right] = \frac{\sigma_V^2}{4\sigma_W^4} t$$

and the relative variance of the weights is finite whenever $2\sigma_V^2 > \sigma_W^2$ and increases exponentially fact with

$$\mathbb{V}_{p(x_{1:t})}\left[\frac{w_t(X_{1:T})}{p(y_{1:t})}\right] = \left(\frac{\left(1 + \frac{\sigma_V^2}{\sigma_W^2}\right)^2}{2\frac{\sigma_V^2}{\sigma_W^2} - 1}\right)^{t/2} - 1$$

where $\left(1+\frac{\sigma_V^2}{\sigma_W^2}\right)^2 / \left(2\frac{\sigma_V^2}{\sigma_W^2}-1\right) > 1$ when $2\sigma_V^2 > \sigma_W^2$. Thus, in order to control the variance of the time index, we would need to choose N as exponential of t.

4 Summary

We have seen that SIS provides estimates whose variance increases typically exponentially with t, the number of observations. Resampling techniques are a key ingredient of SMC methods which (partially) solve this problem in some important scenarios. The next lecture notes will introduce this resampling component, leading to Sequential Monte Carlo methods aka particle filters. We will see that for those methods, the variance of the filtering estimates is stable over time: the number of particles N can be chosen independently of t.

References

[1] Del Moral, P. (2004) *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Series: Probability and Applications, Springer-Verlag, New York.

[2] Doucet, A., de Freitas, N. and Gordon, N.J. (eds.) (2001) *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, New York.

[3] Doucet, A. and Johansen, A.M. (2011), A tutorial on particle filtering and smoothing: fifteen years later. In *Handbook of Nonlinear Filtering*, Cambridge University Press.