$2^{\mathbf{nd}}$ November 2006
Cambridge

Sequential Monte Carlo methods: can we replace the resampling step? [1]

Dan Crisan (Imperial College London)

- Filtering in discrete time:
- Sequential Monte Carlo methods
- Filtering in continuous time.
- Kushner-Stratonovitch versus McKean-Vlasov
- An approximate representation

Dan Crisan, Jie Xiong, *Approximate McKean-Vlasov Representations for a class of SPDEs*, http://arxiv.org/find/grp-math/1/au:+Crisan
Dan Crisan, Jie Xiong, *Numerical solutions for a class of SPDEs over bounded domains*, in preparation.

[1] Recent Advances in Monte Carlo Based Inference, Cambridge 2006.

The filtering problem in discrete time

The signal process: $X = \{X_t; t \in \mathbb{N}\}$, Markov chain with state space $\mathbb{R}^d$, $X_0 \sim \pi_0(dx_0)$, $\Pr(X_t \in A | X_{t-1} = x) = k_t(x, A)$.
Example:

$$X_t = b(X_{t-1}) + \sigma(X_{t-1}) B_t, \qquad B_t \sim N(0,1) \text{ i.i.d.}$$

The observation process: Y associated stochastic process with state space $\mathbb{R}^m$ such that $\Pr\left(Y_t \in B | \mathcal{F}_t^-\right) = \Pr(Y_t \in B | X_t = x_t) = \int_B g_t(y, x_t)\, dy.$
Example:

$$Y_t = h(X_t) + V_t, \qquad V_t \sim N(0,1) \text{ i.i.d.}$$

The filtering problem consists in computing:

$$\pi_t \text{ - the conditional distribution of } X_t \text{ given } \left\{Y_{[0,t]} = y_{[0,t]}\right\}$$

where $Y_{[0,t]} \triangleq (Y_0, ..., Y_t)$, $y_{[0,t]} \triangleq (y_0, ..., y_t)$.

Bayes' recursion.[2]

$$\begin{array}{ll} \text{Prediction} & \rho_t = \pi_{t-1} k_t \\ \text{Updating} & \frac{d\pi_t}{d\rho_t} = C_t^{-1} g_t \end{array},$$

where $C_t \triangleq \int_{\mathbb{R}^d} g_t(y_t, x)\, \rho_t(dx)$.

[2] If μ is a measure and k is a kernel, then $\mu k(A) \triangleq \int \mu(dx)\, k(x, A)$.

Sequential Monte Carlo methods/Particle Filters

- Algorithms to approximate π_t using discrete measures of the form [3]

$$\sum_i \alpha_i \delta_{V_i},$$

  i.e., empirical distributions associated with a set of (random) particles with masses $\alpha_1, \alpha_2, ...$, and positions $V_1, V_2, ...$, respectively, in the state space of X, respectively of $X_{[0,t]}$.

- Recursive algorithms: The approximation for π_t and Y_{t+1} are the only information used in order to obtain the approximation for π_{t+1}. In other words, the information gained from $Y_1, ..., Y_t$ is embedded in the current approximation.

- Quite often: $\pi_t^n = \dfrac{1}{n}\sum_{k=1}^{n_t} \delta_{V_k^t}$ and $E[\pi_t^n] = \pi_t$. Consequently $E[n_t] = n$, with many of the existing algorithms keeping the number of particles constant: $n_t \equiv n$.

- The generic SMC method involves sampling from the prior distribution of the signal and then using a weighted bootstrap technique (or equivalent) with weights defined by the likelihood of the most recent observation data.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith, *Novel approach to nonlinear/non-gaussian bayesian state estimation*, IEE Proceedings on Radar and Signal Processing, vol. 140, pp. 107–113, 1993.

[3] δ_x is the Dirac delta distribution concentrated at x: $\delta_x(A) = \mathbf{1}_A(x)$.

Generic SMC method

At time $t = 0$. *Initialization*

For $i = 1, ..., N$, sample $x_0^{(i)} \sim \pi_0 (dx_0)$ and set $t = 1$.

At time $t \geq 1$,

Step 1: *Importance Sampling step*

For $i = 1, ..., N$, sample $\widetilde{x}_t^{(i)} \sim k_t \left(x_{t-1}^{(i)}, dx_t \right)$ (suboptimal choice $\widetilde{x}_t^{(i)} \sim \pi_{t-1}^N k_t$).

$$\widetilde{x}_t^{(i)} = b \left(x_{t-1}^{(i)} \right) + \sigma \left(x_{t-1}^{(i)} \right) b_t^i$$

For $i = 1, ..., N$, evaluate the normalized importance weights $\overline{w}_t^{(i)} \propto g_t \left(y_t, \widetilde{x}_t^{(i)} \right)$.

Step 2: *Selection step*

Multiply/Discard particles $\left\{ \widetilde{x}_t^{(i)} \right\}_{i=1}^N$ with respect to high/low importance weights $\overline{w}_t^{(i)}$ to obtain N particles $\left\{ x_t^{(i)} \right\}_{i=1}^N$.

Set $t = t + 1$.

Step 2 can be done by means of sampling with replacement (SIR algorithm), stratified sampling, Bernoulli sampling, Carpenter-Clifford-Fearnhead-Whitley genetic algorithm, Crisan-Lyons TBBA algorithm (see Kunsch 2005).

Let π_t^N denote the associated empirical measure $\pi_t^N \triangleq \frac{1}{N}\sum_{i=1}^{N} \delta_{x_t^{(i)}}$.

Theorem. [D.C., P. Del Moral, A. Doucet, T. Lyons, L. Miclo].

Let

$$\mathbf{Error}(t,N) = \sup\left\{ E\left[\left|\int_{\mathbb{R}^d} \varphi\pi_t^N(dx) - \int_{\mathbb{R}^d} \varphi\pi_t(dx)\right|\right]; ||\varphi||_\infty \leq 1 \right\}.$$

Then

$$\mathbf{Error}(t,N) \leq \frac{c_t}{\sqrt{N}},$$

Remarks:

- If the number of updating stages becomes large, the repeated application of the weighted bootstrap may lead to what the literature describes as "impoverished sample" or "sample attrition".
- The sample being carried forward will have fewer and fewer distict values.
- Can we eliminate the weighted bootstrap/resampling step ?

Proposed method:

At time $t = 0$

Initialization.

For $i = 1, ..., N$, sample $x_0^{(i)} \sim \pi_0 (dx_0)$ and set $t = 1$.

At time $t \geq 1$

Interacting importance sampling.

For $i = 1, ..., N$, sample $x_t^{(i)} \sim k_t' \left(y_t, \pi_{t-1}^N, x_{t-1}^{(i)}, dx_t \right)$.

$$\widetilde{x}_t^{(i)} = b \left(x_{t-1}^{(i)} \right) + \lambda \left(x_{t-1}^{(i)}, \pi_{t-1}^N, y_t, \varepsilon \right) + \sigma \left(x_{t-1}^{(i)} \right) b_t^i$$

where

$$\begin{aligned}
\lambda \left(x, \pi_{t-1}^N, y_t, \varepsilon \right) &= \frac{1}{\omega_d \pi_{t-1}^{N,\varepsilon} (x)} \int_{\mathbb{R}^d} \frac{z - x}{||z - x||^d} \bar{d} \left(z, \pi_{t-1}^N, y_t, \varepsilon \right) \pi_{t-1}^N (dx) \\
\pi_{t-1}^{N,\varepsilon} (x) &= \frac{1}{(2\pi\varepsilon)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \exp \left(-\frac{||z - x||^2}{2\varepsilon} \right) \pi_{t-1}^N (dx) \\
d (x, y_t) &= \sum_{k=1}^m h_k (x) y_t^k - \frac{1}{2} h_k^2 (x) \\
\bar{d} \left(x, \pi_{t-1}^N, y_t, \varepsilon \right) &= d (x, y_t) - \int_{\mathbb{R}^d} d (z, y_t) \pi_{t-1}^N (dx)
\end{aligned}$$

Set $t = t + 1$.

Remarks:

- The method does not contain the resampling step.
- The method replaces the sampling from the prior step with sampling from a distribution that depends on the entire (existing) sample and the most recent observation data: The particles are no longer "blind".
- It is inspired by the analogous one for the continuous time filtering problem.
- The continuous time method is motivated by a certain McKean-Vlasov representation for the solution of the the filtering problem.
- The resampling step can be added, if required.

Filtering in continuous time

$(\Omega, \mathcal{F}, P)$ complete probability space. $B = \{\left(B_t^i\right)_{i=1}^p, t \geq 0\}$, $V = \{\left(V_t^i\right)_{i=1}^m, t \geq 0\}$ independent Wiener processes.

The signal process X d-dimensional stochastic process

$$X_t = X_0 + \int_0^t b\left(X_s\right) ds + \int_0^t \sigma\left(X_s\right) dB_s$$

The observation process Y m-dimensional stochastic process

$$Y_t = \int_0^t h\left(X_s\right) ds + V_t$$

$b : \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times p}$, $h = (h_i)_{i=1}^m : \mathbb{R}^d \to \mathbb{R}^m$ Lipschitz functions.

$$\mathcal{Y}_t = \sigma(Y_s, s \in [0,t]), \quad t \geq 0.$$

π_t - the conditional distribution of X_t given $\mathcal{Y}_t$

$$\pi_t\left(\varphi\right) = E[\varphi(X_t)|\mathcal{Y}_t], \quad t \geq 0.$$

π - $\mathcal{Y}_t$-adapted, $\mathcal{P}\left(\mathbb{R}^d\right)$-valued process.

Kushner-Stratonovitch versus McKean-Vlasov

Fujisaki-Kallianpur-Kunita/Kushner-Stratonovitch equation

$$d\pi_t(\varphi) = \pi_t\left(\frac{1}{2}\sum_{i,j=1}^{d} a_{ij}\partial_i\partial_j\varphi + \sum_{i=1}^{d} b_i\partial_i\varphi - d_{\pi_t}\varphi\right)dt + \sum_{k=1}^{m}\pi_t(\beta^k_{\pi_t}\varphi)dY^k_t.$$

where $a = \sigma\sigma^\top$ and

$$d_{\pi_t} = \sum_{k=1}^{m}\left(h_k - \pi_t(h_k)\right)\pi_t(h_k), \qquad \beta^k_{\pi_t} = h_k - \pi_t(h_k).$$

McKean-Vlasov equation

$$dv_t(\varphi) = v_t\left(\frac{1}{2}\triangle\varphi - \sum_{i=1}^{d} b^i_{v_t}\varphi\right)dt,$$

where

$$\begin{aligned} b^i_{v_t}(x) &= \int_{\mathbb{R}^d} b^i(x,y)\,v_t(dy) = v_t\left(b^i(x,\cdot)\right) \\ b^i &: \mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R},\ i=1,...,d \quad \text{interaction kernels.} \end{aligned}$$

Question:

How do you approximate π/v using averages of interactive dynamical systems ?

Statistical mechanics

Macroscopic phenomena are the reflection of the mean behaviour of a large number microscopic particles.

Contributions

McKean (1967), Hitsuda, Mitoma (1986), Ben Arous, Brunaud (1990), Chiang, Kallianpur, Sundar (1991), Florchinger, LeGland (1992), Graham (1992), Dawson (1993) Bernard, Talay, Tubaro (1994), Kallianpur, Xiong (1994), Del Moral (1994, 1999), Kotelenez (1995), Meleard (1996), Dawson, Vaillancourt (1995), Morien (1996), Kurtz, Protter (1996), Bossy, Talay (1996), Graham, Meleard (1997), Ben Arous, Zeitouni (1999), Kurtz, Xiong (1999, 2000), Del Moral, Miclo (2000), Crisan, Lyons (1997,1999), Crisan, Gaines, Lyons (1998), Crisan (2002).

McKean-Vlasov equation

$$dv_t(\varphi) = v_t\left(\frac{1}{2}\triangle\varphi - \sum_{i=1}^{d} b^i_{v_t}\varphi\right)dt,$$

where

$$\begin{aligned} b^i_{v_t}(x) &= \int_{\mathbb{R}^d} b^i(x,y)\, v_t(dy) = v_t\left(b^i(x,\cdot)\right) \\ b^i &: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R},\ i = 1,...,d \quad \text{interaction kernelsboundedLipschitz} \end{aligned}$$

Theorem. Consider the nonlinear diffusion

$$\begin{aligned} dX_t &= b_{\tilde{v}_t}(X_t)\, dt + dW_t \\ &\quad \tilde{v}_t \ \text{isthelawof X}_\text{t} \end{aligned}$$

Then $\tilde{v}_t = v_t$ for all $t \geq 0$.

Interacting particle approximation for M-V

Theorem. Let $\{X^{i,N},\, i = 1, ..., N\}$ be the solution of the following system of SDEs

$$dX_t^{i,N} = \frac{1}{N}\sum_{j=1}^{N} b\left(X_t^{i,N}, X_t^{j,N}\right) dt + dW_t^i, \quad i = 1, ..., N$$

$$\Updownarrow$$

$$\begin{aligned} dX_t^{i,N} &= b_{\tilde{v}_t^N}\left(X_t^{i,N}\right) dt + dW_t^i, \quad i = 1, ..., N \\ \tilde{v}_t^N &= \frac{1}{N}\sum_{j=1}^{N} \delta_{X_t^{j,N}} \end{aligned}$$

Then $\lim_{N\to\infty} \tilde{v}_t^N = v_t$ for all $t \geq 0$.

Remarks:

- The law of $X^{i,N}$ converges to the law of X.
- Propagation of chaos occurs.

Aims:

∘ Derive a nonlinear diffusion whose one-dimensional time distribution coincides with the solution of the K-S equation.

∘ Approximate the solution of the K-S equation with the empirical distribution of a system of (interacting) particles whose trajectories are solutions of a system of SDEs.

Hurdles:

∘ The K-S equation is an SPDE - the M-V is a PDE.

∘ The K-S equation has zero order terms - the M-V does not.

Step1

Approximate the SPDE with a PDE (by means of Clark's representation formula). Choose a partition $T = (t_i)_i$, $t_i = i\delta$. Define $Y^\delta = Y^\delta_{t_i} + \dfrac{Y^\delta_{t_{i+1}} - Y^\delta_{t_i}}{\delta}(t - t_i), \qquad t \in [t_i, t_{i+1})$. Consider the PDE

$$d\pi^\delta_t(\varphi) = \pi^\delta_t \left(\frac{1}{2} \sum_{i,j=1}^{d} a_{ij} \partial_i \partial_j \varphi + \sum_{i=1}^{d} b_i \partial_i \varphi - d_{\pi^\delta_t} \varphi \right) dt,$$

where $d_{\pi^\delta_t} = d^\delta_t - \pi^\delta_t(d^\delta_t)$ and $d^\delta_t = \displaystyle\sum_{k=1}^{m} h_k \frac{Y^\delta_{t_{i+1}} - Y^\delta_{t_i}}{\delta} - \frac{1}{2} h_k^2, \quad t \in [t_i, t_{i+1})$. Then $\lim_{\delta \to 0} \pi^\delta_t = \pi_t$ for all $t \geq 0$. Moreover, we can choose $\delta = \delta(\omega, n)$ and $k = k(\omega)$ such that

$$\left|\left| \pi^\delta_t - \pi_t \right|\right| \leq \frac{k}{n}.$$

Step2

Replace the zero order term with a first order term. Assume that π_t^δ is absolutely continuous w.r.t. Lebesgue measure and has a strictly positive, square integrable density $x \to \pi_t^\delta(x)$, then

$$\pi_t^\delta\left(d_{\pi_t^\delta}\varphi\right) = \pi_t^\delta\left(\sum_{i=1}^d \lambda^i_{\pi_t^\delta}\partial_i\varphi\right),$$

where

$$\lambda_{\pi_t^\delta}(x) = \frac{1}{\omega_d \pi_t^\delta(x)} \int_{\mathbb{R}^d} \frac{y-x}{||y-x||^d} d_{\pi_t^\delta}(y)\, \pi_t^\delta(dy).$$

Hence π^δ is the solution of a Mckean-Vlasov type equation with non-Lipschitz coefficients

$$d\pi_t^\delta(\varphi) = \pi_t^\delta\left(\frac{1}{2}\sum_{i,j=1}^d a_{ij}\partial_i\partial_j\varphi + \sum_{i=1}^d\left(b_i + \lambda^i_{\pi_t^\delta}\right)\partial_i\varphi\right)dt$$

Classical results do not apply !

Results

Theorem. Let $\left(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}\right)$ be a probability space on which there exists a *good* process $X^{\delta} = \left(X_t^{\delta}\right)_{t \geq 0}$ which satisfies equation

$$\begin{aligned} dX_t^{\delta} &= \left(b + \lambda_{\tilde{\pi}_t^{\delta}}\right)\left(X_t^{\delta}\right) dt + \sigma\left(X_t^{\delta}\right) dB_t \\ & \tilde{\pi}_t^{\delta} \ \text{isthelawof} \ \mathrm{X_t^{\delta}} \end{aligned} \tag{1}$$

Then $\tilde{\pi}_t^{\delta} = \pi_t^{\delta}$ for all $t \geq 0$.

Theorem [Uniqueness] There exists at most one solution of (1) which is a good process.

Theorem [Existence] If $x \rightarrow \lambda_{\pi_t^{\delta}}(x)$ is Lipschitz, then (1) has a solution which is a good process.

Example. If X is the solution of an SDE with reflection at $\partial B(0,R)$ along the normal

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t - dK_t,$$

where K is a bounded variation process satisfying

$$\begin{aligned} |K_t| &= \int_0^t \mathbf{1}_{\{X_s \in \partial B(0,R)\}} d|K_s| \\ K_t &= \int_0^t \frac{X_s}{R} d|K_s|. \end{aligned}$$

See Lions and Sznitman 1984 for details.