Filtering in continuous time by least action

Chris Rogers

First version: September 2008

August 25, 2010

0. The reason I volunteered for the working group on continuous-time filtering problems was because it seemed to me that in the other areas probably the best you could hope to achieve was some improvement of existing 'bread-and-butter' SMC methods, whereas in continuous time there was much more scope for novelty. In particular, in continuous time you have all the tools of stochastic calculus, so it should be possible to use these *before* discretizing, as a way of getting further. These notes present one way of carrying out this approach; perhaps all which follows is well known, but from my brief study of the references we have been looking at, it looks like it could do with being better known if so! **These notes are very preliminary, and will be expanded as time permits.**

1. The informal idea of least-action is that Wiener measure on $C([0,T],\mathbb{R}^d)$ has a 'density' with respect to 'Lebesgue' measure of the form

$$f(x) \propto \exp\Big\{ -\tfrac{1}{2} \int_0^T |\dot{x}_s|^2 \, ds \Big\}. \qquad (1)$$ density

While it is impossible to make rigorous sense of this, it is the intuition which drives a great deal of stochastic analysis, such as large deviations. If we were only to observe the process at multiples of some small time-step $h = T/N$, then the density of $(x_h, x_{2h}, \ldots, x_T)$ as a random vector in $\mathbb{R}^{Nd}$ would indeed be of the form (1), where for $\dot{x}_s$ we substitute the slope of the piecewise-linear interpolation of the x_{jh}. The expression $\frac{1}{2}\int_0^T |\dot{x}_s|^2 \, ds$ is often referred to as the *action integral*, and is in effect the negative of the log-likelihood.

2. Suppose that we have some (nice) SDE in $\mathbb{R}^d$

$$dZ_t = \sigma(t, Z_t)\, dW_t + \mu(t, Z_t) dt, \qquad (2)$$ dZ

where Z is partitioned[1] $Z = [X; Y]$ where X is d_1-dimensional, and Y is d_2-dimensional. We shall suppose that Y is observed, and the objective is to estimate X from these observations.

[1] ... using Scilab/Matlab notation for vectors and matrices ..

Formally rearranging (2) gives us

$$\frac{dW}{dt} = \sigma(t, Z_t)^{-1}\Big(\frac{dZ}{dt} - \mu(t, Z_t) \Big), \tag{3}$$ dWdt

so that the action integral here will be

$$\begin{aligned} A &= \tfrac{1}{2} \int_0^T \Big| \sigma(t, Z_t)^{-1}\Big\{ \begin{pmatrix} \dot{x}_t \\ \dot{Y}_t \end{pmatrix} - \mu(t, Z_t) \Big\} \Big|^2 \, dt \\ &\equiv \int_0^T \psi(t, x_t, \dot{x}_t) \, dt, \end{aligned} \tag{4}$$

say. Notice that we know Y, so it is only an issue to select the (least-action = maximum-likelihood) path x. If we suppose that x_0 has a prior density proportional to $\exp(-\varphi(x))$, then the log-likelihood to be minimized is

$$\varphi(x_0) + \int_0^T \psi(t, x_t, \dot{x}_t) \, dt \equiv \varphi(x_0) + \int_0^T \psi(t, x_t, p_t) \, dt, \tag{5}$$ LL

where we write $p \equiv \dot{x}$. This we attack by calculus of variations; if we have found the optimal x, then any perturbation to $x + \xi$ must to leading order make zero change to the objective. Integrating by parts gives

$$\begin{aligned} 0 &= \xi_0 D\varphi(x_0) + \int_0^T \big\{ \xi_t \cdot D_x\psi + \dot{\xi} \cdot D_p\psi \big\} \, dt \qquad (6) \\ &= \xi_0 D\varphi(x_0) + \big[\xi_t^j D_{p_j}\psi \big]_0^T + \int_0^T \xi_t^j \big\{ D_{x_j}\psi - D_t D_{p_j}\psi - \dot{x}_k D_{p_j} D_{x_k}\psi - \dot{p}_k D_{p_j} D_{p_k}\psi \big\} \, dt. \end{aligned}$$

Since ξ is arbitrary, we deduce the conditions for optimality:

$$\begin{aligned} 0 &= D_{p_j}\psi(0, x_0, p_0) - D_{x_j}\varphi(x_0) \qquad (7) \\ 0 &= D_{x_j}\psi - D_t D_{p_j}\psi - \dot{x}_k D_{p_j} D_{x_k}\psi - \dot{p}_k D_{p_j} D_{p_k}\psi \qquad (8) \\ 0 &= D_{p_j}\psi(T, x_T, p_T) \qquad (9) \end{aligned}$$

Thus we derive a (generally non-linear) first-order ODE (8) for (x, p), with initial condition (7) and terminal condition (9). This will generally have a unique solution, though the 'shooting' nature of the ODE is rather clumsy in practice[2].

3. There's quite a rigour deficit here, so let's take a look at the simplest example to try to see what happens here. Suppose that the underlying signal X is a one-dimensional OU process

$$dX = \sigma_X dW - \beta_X X dt,$$

[2]We shall have more to say on this later.

and that $Y = X + y$, where y is an independent OU process

$$dy = \sigma_y dW' - \beta_y y dt$$

for independent Brownian motions W and W'. Assume that the prior for X_0 is a standard normal. The SDE satisfied by $Z = [X; Y]$ is thus

$$\begin{aligned} dZ &= \begin{pmatrix} \sigma_X & 0 \\ \sigma_X & \sigma_y \end{pmatrix} \begin{pmatrix} dW \\ dW' \end{pmatrix} + \begin{pmatrix} -\beta_X & 0 \\ -\beta_X + \beta_y & -\beta_y \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} dt \\ &\equiv \sigma \begin{pmatrix} dW \\ dW' \end{pmatrix} + A \begin{pmatrix} X \\ Y \end{pmatrix} dt \end{aligned}$$

Therefore we have

$$\psi(t, x, p) = \frac{1}{2} \left(\begin{pmatrix} p \\ \dot{Y} \end{pmatrix} - A \begin{pmatrix} x \\ Y \end{pmatrix} \right)^T M \left(\begin{pmatrix} p \\ \dot{Y} \end{pmatrix} - A \begin{pmatrix} x \\ Y \end{pmatrix} \right) \tag{10}$$

where $M = (\sigma\sigma^T)^{-1}$. This is a completely explicit action functional, whose derivatives are quite easy to work with.

For a given x_0, we can derive p_0 from (7), then numerically solve the ODE for $(x, p) \equiv (x, \dot{x})$ forward to time T, and finally check whether the terminal condition (9) holds. Classically, the approach to be used is to adjust x_0 until the terminal condition is satisfied, but the SMC approach suggests another way to do this; keep a cloud of starting points, and reweight according to the likelihood of the path which solves (7), (8). Figure 1 shows what happened when this procedure was carried out on this example with 200 time steps.

4. Let's see this in action on a rather different example, which could arise in modelling of defaults. Suppose that there is an unobserved CIR process x, satisfying

$$dx_t = \sigma \sqrt{x_t}\, dW_t + \beta(a - x_t) dt, \tag{11}$$ CIR

which serves as the stochastic intensity of a counting process N. The times $0 < \tau_1 < \tau_2 < \ldots < \tau_{N_T} < T$ of events are observed, and we have to filter x from these observations. For this example, the action functional is simply

$$\psi(t, x, p) = \psi(x, p) = \frac{(p - \beta(a - x))^2}{2\sigma^2 x}, \tag{12}$$ LAF2

and the action integral (= -log-likelihood) to be minimised is just

$$\varphi(x_0) + \int_0^T \psi(x_s, \dot{x}_s)\, ds + \int_0^T x_s\, ds - \sum_{i=0}^{n} \log x_{\tau_i}, \tag{13}$$ LL2

where we have abbreviated $n \equiv N_T$. Once again, we consider a small perturbation ξ away from the optimal x, and equate the first-order change to zero, which gives us

$$\begin{aligned} 0 &= \varphi'(x_0)\xi_0 + \int_0^T \{\psi_x \xi + \psi_p \dot{\xi} + \xi\}\, dt - \sum_{i=1}^{n} \frac{\xi(\tau_i)}{x(\tau_i)} \\ &= \varphi'(x_0)\xi_0 + \sum_{j=0}^{n} \int_{\tau_j}^{\tau_{j+1}} \{\psi_x \xi + \psi_p \dot{\xi} + \xi\}\, dt - \sum_{i=0}^{n} \frac{\xi(\tau_i)}{x(\tau_i)} \\ &= \varphi'(x_0)\xi_0 + \sum_{j=0}^{n} \Bigg[\int_{\tau_j}^{\tau_{j+1}} \{\psi_x - \psi_{px}\dot{x} - \psi_{pp}\dot{p} + 1\}\xi\, dt + \Big[\xi_t \psi_p(x_t, \dot{x}_t)\Big]_{\tau_j}^{\tau_{j+1}} \Bigg] - \sum_{i=0}^{n} \frac{\xi(\tau_i)}{x(\tau_i)}. \end{aligned}$$

In order for this to be zero whatever perturbation ξ is used, we have to have a number of conditions:

$$\begin{aligned} 0 &= \psi_x - \psi_{px}\dot{x} - \psi_{pp}\dot{p} + 1 \quad \text{in each interval } (\tau_j, \tau_{j+1}) \quad ; && (14) \\ 0 &= \varphi'(x_0) - \psi_p(x_0, p_0)\ ; && (15) \\ 0 &= -\psi_p(x_{\tau_i}, p_{\tau_i +}) + \psi_p(x_{\tau_i}, p_{\tau_i -}) - x(\tau_i)^{-1}\ ; && (16) \\ 0 &= \psi_p(x_T, p_T). && (17) \end{aligned}$$

Thus the least-action path must be constructed piecewise in each of the intervals between observations. We are assisted in this by the fact that the ODE (14) to be solved can be solved quite explicitly:

$$x_t = \left(\frac{\beta^2 a^2}{A} - \frac{AB^2}{4}\right) e^{\beta t} - \frac{Ae^{-\beta t}}{4\beta^2} - \frac{AB}{2\beta} \qquad (18)$$

xsoln

for some constants A, B.

As in the first example, we have a shooting problem, where we have to pick a starting point x_0, find p_0 using (15), then solve the ODE out to τ_1 using the solution (18), use (16) to work out the value of p to the right of τ_1, and then continue to the end, finally checking whether (17) holds at T, adjusting x_0 until it does.

5. Some remarks. (1) Methodologically, what we are doing is to solve some ODE, and wiggle the initial condition around until we get a good fit. Notice that solving an ODE (even in quite large dimensions) is a well-studied problem in numerical analysis, and there are good accurate fast numerical schemes. Thus we should expect that this approach will cope better with high-dimensional problems than pure SMC methods.

(2) Another observation is that the methodology is similar to the Doss-Sussmann approach to solving an SDE; you calculate the stochastic flow of the associated ODE, and then make the initial condition diffuse. However, it is clear that the two approaches are actually radically different; the Doss-Sussmann approach gives us a *first-order* ODE, the least-action approach gives us a *second-order* ODE.

(3) If T gets large, the numerical feasibility of this least-action approach falls off; we will likely need unachievable precision in the location of x_0 to hit the final condition. This suggests that we should be thinking in a more recursive way about the approach, and this actually fits quite well with SMC, as was mentioned earlier. If we keep a cloud of particles, each the current position of a solution of the ODE (8) from a different starting point, then the SMC method gives us weights on these. Suppose we have solved out to time t, and have such a population of particles. Then to move forward to $t+1$, say, we could step forward from each of the existing particles (using the ODE (8)), and see how they do. Of course, likely none of them will hit the terminal condition (9) at time $t+1$, but by some interpolation between values at time t we may be able to get closer.

(4) Even if we want to keep close in spirit to classical SMC methodology, this least-action analysis can still be very helpful. Go back to the situation in Section 2, (2). Suppose that we have been able to compute the least-action path Z^* over $[0,T]$; then (3) would suggest that

$$\frac{dW}{dt} = b_t \equiv \sigma(t, Z_t^*)^{-1}\left(\frac{dZ^*}{dt} - \mu(t, Z_t^*) \right). \tag{19}$$ dWdt2

So we could use this to importance-sample in a good way; instead of simulating paths of W as ordinary Brownian motions, we could simulate them as Brownian motions plus the known drift $b_t dt$, and then use Cameron-Martin-Girsanov to importance-weight them. This would be quite easy to do, and would put paths where they could do a lot of work - always difficult in higher dimensions.

(5) In support of the basic notion that least-action corresponds in some sense to the maximum-likelihood path, suppose we take a simple SDE

$$dX_t = dW_t + \mu(X_t)dt$$

and approximate it by the Euler scheme:

$$dX_t^{(n)} = dW_t + \mu(X^{(n)}(t^{(n)}))\, dt, \tag{20}$$ Euler

where $t^{(n)} = 2^{-n}[2^n t]$, with conditionally Gaussian increments. Writing $h \equiv 2^{-n}$, and $x_j \equiv X^{(n)}(jh)$, the log likelihood is to within irrelevant constants

$$-\varphi(x_0) - \tfrac{1}{2}\sum_{j=0}^{N-1}(x_{j+1} - x_j - h\mu(x_j))^2.$$

Differentiating with respect to x_j gives us the conditions (writing $\mu_j \equiv \mu(x_j)$, $\mu_j' \equiv \mu'(x_j)$)

$$\begin{aligned}
0 &= (x_j - x_{j-1} - h\mu_{j-1}) - (1 + h\mu_j')(x_{j+1} - x_j - h\mu_j) \\
&= (2x_j - x_{j-1} - x_{j+1}) + h(\mu_j - \mu_{j-1}) - h\mu_j'(x_{j+1} - x_j - h\mu_j) \\
&= (2x_j - x_{j-1} - x_{j+1}) + h\mu_j'(x_j - x_{j-1}) - h\mu_j'(x_{j+1} - x_j - h\mu_j) + \varepsilon_j \\
&= (2x_j - x_{j-1} - x_{j+1})(1 + h\mu_j') + h^2\mu_j\mu_j' + \varepsilon_j,
\end{aligned}$$

where $\varepsilon_j = O(h^3)$ when we assume that $x_{j+1} - x_j = O(h)$. Dividing by h^2, we deduce

$$0 = -\frac{x_{j+1} - 2x_j + x_{j-1}}{h^2} + \mu_j \mu_j' + O(h); \tag{21}$$ FD1

this is a finite-difference scheme for the ODE (8), more simply expressed as equation (2) in JV's notes of November 10th.

6. Let's now look at some higher-order stuff. What we did was to maximize the log-likelihood by minimizing the action (5). Assuming that the log-likelihood is C^2 near its maximum, we have (as in classical ML theory) that the likelihood surface is approximately quadratic near the maximum, and the second derivative of the quadratic tells us the approximate Gaussian behaviour near to the MLE. So if we go back to the action (5) and consider expanding around the optimal x^* by perturbing to $x^* + \xi$ for small ξ, the second-order contribution we get is

$$\begin{aligned} Q(\xi) &\equiv \tfrac{1}{2} D_{x_i} D_{x_j} \varphi(x_0^*) \xi_0^i \xi_0^j + \int_0^T \left\{ \tfrac{1}{2} \xi^i \xi^j D_{x_i} D_{x_j} \psi + \xi^i \dot{\xi}^j D_{p_j} D_{x_i} \psi + \tfrac{1}{2} \dot{\xi}^i \dot{\xi}^j D_{p_i} D_{p_j} \psi \right\} dt \\ &\equiv \tfrac{1}{2} D_{x_i} D_{x_j} \varphi(x_0^*) \xi_0^i \xi_0^j + \int_0^T \left\{ \tfrac{1}{2} \xi_t^i A_t^{ij} \xi_t^j + \xi_t^i B_t^{ij} \dot{\xi}_t^j + \tfrac{1}{2} \dot{\xi}_t^i q_t^{ij} \dot{\xi}_t^j \right\} dt \end{aligned} \tag{22}$$

where the derivatives appearing in the integral are evaluated along the optimal path $(x_t^*, \dot{x}_t^*)$, and the matrix-valued functions of time A, B and q are defined by the obvious identifications. This quadratic functional of ξ characterizes the (approximate) Gaussian distribution of the perturbation.

Now suppose that we have some C^1 symmetric-matrix-valued function of time, θ, such that $\theta_T = 0$. Then we may write

$$\begin{aligned} Q(\xi) &= Q(\xi) + \tfrac{1}{2} \xi_0 \cdot \theta_0 \xi_0 + \left[\tfrac{1}{2} \xi_t \cdot \theta_t \xi_t \right]_0^T \\ &= \tfrac{1}{2} \xi_0 \cdot (D^2 \varphi(x_0^*) + \theta_0) \xi \\ &\qquad + \int_0^T \left\{ \tfrac{1}{2} \xi_t \cdot A_t \xi_t + \xi_t \cdot B_t \dot{\xi}_t + \tfrac{1}{2} \dot{\xi}_t \cdot \, q_t \dot{\xi}_t + \tfrac{1}{2} \xi_t \cdot \dot{\theta}_t \xi_t + \xi_t \cdot \theta_t \dot{\xi}_t \right\} dt. \end{aligned} \tag{23}$$

The quadratic for inside the integral is

$$\tfrac{1}{2} \dot{\xi}_t \cdot q_t \dot{\xi}_t + \xi_t \cdot (B_t + \theta_t) \dot{\xi}_t + \tfrac{1}{2} \xi_t \cdot (A_t + \dot{\theta}_t) \xi_t = \tfrac{1}{2} (\dot{\xi}_t + K_t \xi_t) \cdot q_t (\dot{\xi}_t + K_t \xi_t) \tag{24}$$ CTS

where $K_t \equiv q_t^{-1}(B_t^T + \theta_t)$ *provided*

$$A_t + \dot{\theta}_t = K_t^T q_t K_t = (B_t + \theta_t) q_t^{-1} (B_t^T + \theta_t) \tag{25}$$ eq6

This gives an ODE for θ to be solved with the boundary condition $\theta_T = 0$, which presents no real problems. Once we have solved this, we conclude that the perturbation ξ solves

$$d\xi_t = -K_t \xi_t \, dt + q_t^{-1/2} \, dW_t \tag{26}$$ dxi

which represents the perturbation as a zero-mean Gaussian process. The covariance of ξ can be easily obtained from an Itô expansion of $\xi\xi^T$:

$$d\xi\xi^T \doteq (-K_t\xi_t\xi_t^T - \xi_t\xi_t^T K_t^T + q_t^{-1})dt, \tag{27}$$ dxixi

where $\doteq$ signifies that the two sides differ by a (local) martingale. Hence

$$\dot{V}_t = -K_t V_t - V_t K_t^T + q_t^{-1} \tag{28}$$ dV

and this allows us to calculate the covariance at time t.

7. In any given example, the action functional takes the form

$$\psi(t,x,p) = \tfrac{1}{2}\,(v - b(t,x)) \cdot q(t,x)(v - b(t,x)), \tag{29}$$ psi2

where

$$v \equiv \begin{pmatrix} p \\ 0 \end{pmatrix}, \qquad q(t,x) \equiv (\sigma(t,z)\sigma(t,z)^T)^{-1}, \qquad b(t,x) \equiv \mu(t,z) - \begin{pmatrix} 0 \\ \dot{Y}_t \end{pmatrix} \tag{30}$$ wq

where we assume that

$$z = \begin{pmatrix} x \\ Y_t \end{pmatrix}$$

when we evaluate σ and μ. This will typically introduce time dependence into b and q even though there may have been no time dependence in μ and σ.

The primitives of an example will be the functions μ and σ, and it is convenient to re-express ψ and its relevant derivatives in terms of μ, σ and their derivatives. Writing $w = v - b(t,x)$ as a convenient abbreviation, we shall then have

$$\begin{aligned}
D_{x_j}\psi &= \tfrac{1}{2}\, w \cdot (D_{x_j} q)\, w - (D_{x_j} b) \cdot q\, w \\
D_{x_j x_k}\psi &= \tfrac{1}{2}\, w \cdot (D_{x_j x_k} q)\, w - (D_{x_j} b) \cdot (D_{x_k} q)\, w - (D_{x_k} b) \cdot (D_{x_j} q)\, w + \\
&\qquad + (D_{x_j} b) \cdot q\, (D_{x_k} b) - (D_{x_k x_j} b) \cdot q\, w \\
D_{p_k}\psi &= (q\, w)_k \\
D_{p_j p_k}\psi &= q_{jk} \\
D_{x_j p_k}\psi &= (\,(D_{x_j} q)\, w)_k - (q\,(D_{x_j} b))_k \\
D_t D_{p_k}\psi &= (D_t q\, w)_k - (q D_t b)_k
\end{aligned}$$

8. Here is another interesting example to study. Suppose we have a hidden process

$$dX_t = \sigma_X dW_t - \beta \sin(aX_t)dt \tag{31}$$ dX4

for positive constants σ_X, β and a, which we observe with additive OU noise z:

$$Y_t = X_t + z_t,$$

where

$$dz_t = \sigma_z dW'_t - \lambda z_t dt \tag{32}$$

for other positive constants σ_z and λ. This is an interesting dynamic, because X is mean-reverting to any point of the form $2n\pi/a$ for integer n, and mean-fleeing from any point of the form $(2n+1)\pi/a$. Thus it will tend to stick around integer multiples of $2\pi/a$, occasionally crossing over to a neighbouring point. The SDE for Z is thus

$$dZ_t \equiv \begin{pmatrix} dX_t \\ dY_t \end{pmatrix} = \begin{pmatrix} \sigma_X & 0 \\ \sigma_X & \sigma_z \end{pmatrix} \begin{pmatrix} dW_t \\ dW'_t \end{pmatrix} + \begin{pmatrix} -\beta \sin(aX_t) \\ -\beta \sin(aX_t) - \lambda(Y_t - X_t) \end{pmatrix} dt. \tag{33}$$

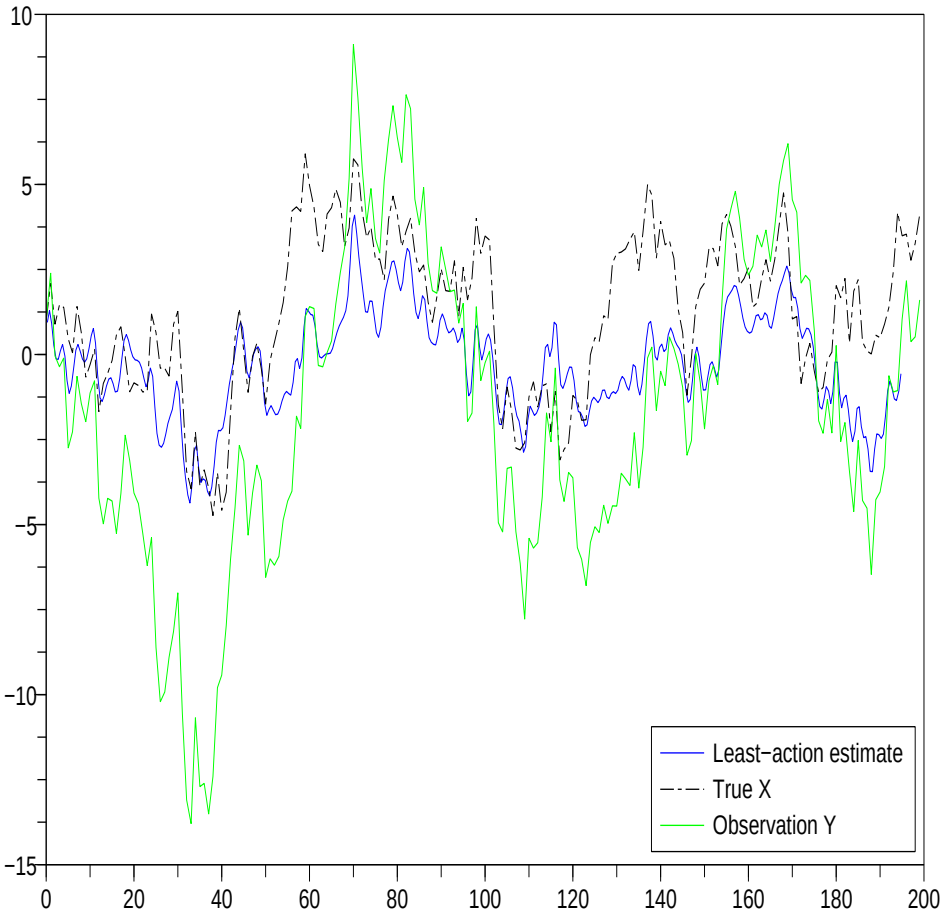

Figure 1: Result of least-action filtering.

pic1