Particle Filters: Convergence Results and High Dimensions

Mark Coates
mark.coates@mcgill.ca

McGill University
Department of Electrical and Computer Engineering
Montreal, Quebec, Canada

Bellairs 2012

Outline

References

- Crisan, D. and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. IEEE Trans. Signal Processing, 50(3):736-746, Mar. 2002.
- Beskos, A., Crisan, D., & Jasra A. (2011). On the stability of sequential Monte Carlo methods in high dimensions. Technical Report, Imperial College London.
- Snyder, C., Bengtsson, T., Bickel, P., & Anderson, J. (2008). Obstacles to high-dimensional particle filtering. Month. Weather Rev., 136, 46294640.
- Bengtsson, T., Bickel, P., & Li, B. (2008). Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In Essays in Honor of David A. Freeman, D. Nolan & T. Speed, Eds, 316334, IMS.
- Quang, P.B., Musso, C. and Le Gland F. (2011). An Insight into the Issue of Dimensionality in Particle Filtering. Proc. ISIF Int. Conf. Information Fusion, Edinburgh, Scotland.

Discrete-time Filtering

- Fixed observations $y_1, \dots, y_n$ with $y_k \in \mathbb{R}^{d_y}$.
- Hidden Markov chain $X_0, \dots, X_n$ with $X_k \in E^d$.
- Initial distribution $X_0 \sim \mu(dx_0)$.
- Probability transition kernel $K(dx_t|x_{t-1})$ such that:

$$\Pr(X_t \in A|X_{t-1} = x_{t-1}) = \int_A K(dx_t|x_{t-1}) \tag{1}$$

- Observations conditionally independent of X and have marginal distribution:

$$\Pr(Y_t \in B|X_t = x_t) = \int_B g(dy_t|x_t) \tag{2}$$

Bayes' Recursion

- Paths of signal and observation processes from time k to l:

$$X_{k:l} = (X_k, X_{k+1}, \ldots, X_l); \qquad Y_{k:l} = (Y_k, Y_{k+1}, \ldots, Y_l).$$

- Define probability distribution:

$$\pi_{k:l|m}(dx_{k:l}) = P(X_{k:l} \in dx_{k:l} | Y_{1:m} = y_{1:m})$$

- Bayes theorem leads to the following relationship:

$$\pi_{0:t|t}(dx_{0:t}) \propto \mu(dx_0) \prod_{k=1}^{t} K(dx_k | x_{k-1}) g(y_k | x_k) \tag{3}$$

Bayes' Recursion

- Prediction:

$$\pi_{0:t|t-1}(dx_{0:t}) = \pi_{0:t-1|t-1}(dx_{0:t-1})K(dx_t|x_{t-1})$$

- Update:

$$\pi_{0:t|t}(dx_{0:t}) = \left[\int_{\mathbb{R}_y^d} \pi_{0:t|t-1}(dx_{0:t})\right]^{-1} g(y_t|x_t)\pi_{0:t|t-1}(dx_{0:t})$$

Particle Filtering

- Recursive algorithm.
- Produce particle cloud with empirical measure close to $\pi_{t|t}$.
- N particle paths $\{x_t^{(i)}\}_{i=1}^N$.
- Associated empirical measure:

$$\pi_{t|t}^N(dx_t) = \frac{1}{N}\sum_{i=1}^N \delta_{x_t^{(i)}}(dx_t) \tag{4}$$

Particle Filtering

- Initialization: Sample $x_0^{(i)} \sim \pi_{0|0}(dx_0)$.
- For $t \geq 1$
- Importance sampling: Sample $\tilde{x}_t^{(i)} \sim \pi_{t-1|t-1}^N K(dx_t)$.
- Weight evaluation:

$$w_t^{(i)} \propto g(y_t|\tilde{x}_t^{(i)}); \quad \sum_{i=1}^{N} w_t^{(i)} = 1 \tag{5}$$

- Resample: Sample $x_t^{(i)} \sim \tilde{\pi}_{t|t}^N(dx_t)$.

Drawbacks

Variation of Importance Weights

- Distn. of particles $\{\tilde{x}_t^{(i)}\}_{i=1}^N$ is approx. $\pi_{t|t-1} = \pi_{t-1|t-1}K$.
- The algorithm can be inefficent if this is "far" from $\pi_{t|t}$.
- Then the ratio:

$$\frac{\pi_{t|t}(dx_t)}{\pi_{t|t-1}(dx_t)} \propto g(y_t|x_t)$$

can generate weights with high variance.

Drawbacks

Variation induced by resampling

- Proposed resampling generates $N_t^{(i)}$ copies of the i-th particle.
- These are drawn from a multinomial distribution, so:

$$\begin{aligned} E(N_t^{(i)}) &= N w_t^{(i)} \\ \mathrm{var}(N_t^{(i)}) &= N w_t^{(i)}(1 - w_t^{(i)}) \end{aligned}$$

Sequential Importance Sampling/Resampling

- Initialization: Sample $x_0^{(i)} \sim \pi_{0|0}(dx_0)$.
- For $t \geq 1$
- Importance sampling: Sample $\tilde{x}_t^{(i)} \sim \pi_{t-1|t-1}^N \tilde{K}(dx_t)$.
- Weight evaluation:

$$w_t^{(i)} \propto \frac{K(dx_t|x_{t-1}^{(i)})g(y_t|\tilde{x}_t^{(i)})}{\tilde{K}(dx_t|x_{t-1}^{(i)})}; \quad \sum_{i=1}^{N} w_t^{(i)} = 1 \tag{6}$$

- Resample: Sample $x_t^{(i)} \sim \tilde{\pi}_{t|t}^N(dx_t)$.

Sequential Importance Sampling/Resampling

- Algorithm is the same as the bootstrap with a new dynamic model.

$$
\begin{aligned}
Pr(X_t \in A | X_t = x_{t-1}, Y_t = y_t) &= \int_A \tilde{K}(dx_t | x_{t-1}, y_t) \\
Pr(Y_t \in B | X_{t-1} = x_{t-1}, X_t = x_t) &= \int_B w(x_{t-1}, x_t, dy_t)
\end{aligned}
$$

- Only true if we assume observations are fixed!
- With this model, $\rho_{0:t|t-1} \neq \pi_{0:t|t-1}$ but $\rho_{0:t|t} = \pi_{0:t|t}$.
- If $\tilde{K}$ has better mixing properties, or $w(x_{t-1}, x_t, y_t)$ is a flatter likelihood, then algorithm will perform better.

Almost Sure Convergence

Theorem

Assume that the transition kernel K is Feller and that the likelihood function g is bounded, continuous and strictly positive, then $\lim_{N \to \infty} \pi_{t|t}^N = \pi_{t|t}$ *almost surely.*

- Feller: for φ a continuous bounded function, $K\varphi$ is also a continous bounded function.
- Intuition: we want two realizations of the signal that start from "close" positions to remain "close" at subsequent times.
- Define $(\mu, \varphi) = \int \varphi \mu$.
- We write $\lim_{N \to \infty} \mu^N = \mu$ if $\lim_{N \to \infty} (\mu^N, \varphi) = (\mu, \varphi)$ for any continuous bounded function φ.

Proof discussion

- Let (E, d) be a metric space
- Let $(a_t)_{t=1}^{\infty}$ and $(b_t)_{t=1}^{\infty}$ be two sequences of continuous functions $a_t, b_t : E \to E$.
- Let k_t and $k_{1:t}$ be defined:

$$k_t = a_t \circ b_t \quad k_{1:t} = k_t \circ k_{t-1} \circ \cdots \circ k_1. \tag{7}$$

- Perturb k_t and $k_{1:t}$ using function c^N:

$$k_t^N = c^N \circ a_t \circ c^N \circ b_t \quad k_{1:t}^N = k_t^N \circ k_{t-1}^N \circ \cdots \circ k_1^N. \tag{8}$$

- Assume that as N becomes larger, perturbations become smaller; c^N converges to the identity function on E.
- Does this mean that k_t^N and $k_{1:t}^N$ converge?

Counterexample

- Let $E = [0,1]$ and $d(\alpha, \beta) = |\alpha - \beta|$.
- Let a_t and b_t be equal to identity i on E; so k_t is also identity.

$$c^N(\alpha) = \begin{cases} \alpha + \dfrac{\alpha}{N}, & \text{if } \alpha \in [0, 1/2] \\ 1 - (N-1)|\dfrac{1}{2} + \dfrac{1}{2N} - \alpha|, & \text{if } \alpha \in (\dfrac{1}{2}, \dfrac{1}{2} + \dfrac{1}{N}) \\ \alpha + \dfrac{\alpha - 1}{N-2}, & \text{if } \alpha \in (\dfrac{1}{2} + \dfrac{1}{N}, 1) \end{cases}$$

- Now $\lim_{N \to \infty} c^N(\alpha) = \alpha$ for all $\alpha \in [0,1]$.
- But $\lim_{N \to \infty} k^N(\frac{1}{2}) = \lim_{N \to \infty} c^N \left(\frac{1}{2} + \frac{1}{2N} \right) = 1$

Proof Discussion

- So successive small perturbations may not always lead to a small perturbation overall.
- We need a stronger type of convergence for c^N: a uniform manner.
- For all $\epsilon > 0$ there exists $N(\epsilon)$ such that $d(c^N(e), i(e)) < \epsilon$ for all $N \geq N(\epsilon)$.
- This implies that $\lim_{N\to\infty} e^N = e \Rightarrow \lim_{N\to\infty} c^N(e_N) = e$.
- Then $\lim_{N\to\infty} k_t^N = k_t$ and $\lim_{N\to\infty} k_{1:t}^N = k_{1:t}$

- $E = P(\mathbb{R}^d)$: set of probability measures over $\mathbb{R}^d$ endowed with topology of weak convergence.
- μ_N converges weakly if $lim_{N\to\infty}(\mu_N, \varphi) = (\mu, \varphi)$ for all continuous bounded functions φ.
- Here $(\mu, \varphi) = \int \varphi\mu$.
- Define $b_t(\nu)(dx_t) = \int_{\mathbb{R}^d} K(dx_t|x_{t-1})\nu(dx_{t-1})$.
- So $\pi_{t|t-1} = b_t(\pi_{t-1|t-1})$.
- Let $a_t(\nu)$ be a probability measure: $(a_t, \nu) = (\nu, g)^{-1}(\nu, \varphi g)$ for any continuous bounded function φ.
- Then $\pi_{t|t} = a_t(\pi_{t|t-1}) = a_t \circ b_t(\pi_{t-1|t-1})$.

Filtering Application

- Assume a_t is continuous; slight variation in conditional distribution of X_t will not result in big variation in conditional distribution after y_t taken into account.
- One way: assume $g(y_t|\cdot)$ is continuous, bounded strictly positive function.
- Positivity ensures the normalizing denominator is never 0.
- Particle filtering: perturbation c^N is random, but with probability 1 we have the properties outlined above.

Convergence of the Mean Square Error

- Different convergence: $\lim_{N\to\infty} E[((\mu_N, \varphi) - (\mu, \varphi))^2] = 0.$
- Expectation over all realizations of the random particle method.
- Assumption: $g(y_t|\cdot)$ is a bounded function in argument x_t.

Theorem

There exists $c_{t|t}$ independent of N such that for any continous bounded function φ:

$$E\left[((\pi^N_{t|t}, \varphi) - (\pi_{t|t}, \varphi))^2\right] \leq c_{t|t}\frac{||\varphi||^2}{N} \tag{9}$$

Convergence of the Mean Square Error

- If one uses a kernel $\tilde{K}$ instead of K, we need that $||w|| < \infty$.
- "In other words, particle filtering methods beat the *curse of dimensionality* as the rate of convergence is independent of the state dimension d."
- "However to ensure a given precision on the mean square error...the number of particles N also depends on $c_{t|t}$, which can depend on d." [Crisan and Doucet, 2002]

Uniform Convergence

- We have shown that $(\pi^N_{t|t}, \varphi)$ converges to $(\pi_{t|t}, \varphi)$ in the mean-square sense.
- Rate of convergence is in $1/N$.
- But how does $c_{t|t}$ behave over time?
- If the true optimal filter doesn't forget its initial conditions, then errors accumulate over time.
- Need mixing assumptions on dynamic model (and thus on the true optimal filter).
- Uniform convergence results can be obtained [Del Moral 2004].

Curse of dimensionality

- Let's consider the batch setting.
- Observe $Y_{1:n}$; try to estimate the hidden state X_n.
- Let $g(y_{1:n}|x)$ be the likelihood and $f(x)$ the prior density.
- Suppose $f(x)$ is chosen as the importance density.
- RMSE convergence can be bounded [Leglande, Oudjane 2002] as:

$$E\left[((\pi_{t|t}^N,\varphi)-(\pi_{t|t},\varphi))^2\right]^{1/2}\leq\frac{c_0}{\sqrt{N}}I(f,g)||\varphi|| \quad (10)$$

where

$$I(f,g)=\frac{\sup_x g(y_{1:n}|x)}{\int_{\mathbb{R}^d} g(y_{1:n}|x)f(x)dx} \quad (11)$$

Curse of dimensionality

- We can consider that the term $I(f, g)$ characterizes the Monte Carlo (MC) error.
- As $\int_{\mathbb{R}^d} g(y_{1:n}|x)f(x)dx$ tends towards zero, the MC error increases.
- The integral represents the discrepancy between the prior and the likelihood.
- Weight variance:

$$\mathrm{var}(w^{(i)}) \approx \frac{1}{N^2} \left(\frac{\int_{\mathbb{R}^d} g(y_{1:n}|x)^2 f(x)\, dx}{\left(\int_{\mathbb{R}^d} g(y_{1:n}|x) f(x)\, dx\right)^2} - 1 \right) \tag{12}$$

- (Quang et al. 2011) provide a case-study showing that the MC error grows exponentially with the dimension.

More advanced algorithms?

- Insert an annealing SMC sampler between consecutive time steps, updating entire trajectory $x_{1:n}$.
- Algorithm is stable as $d \to \infty$ with cost $\mathcal{O}(n^2 d^2 N)$.
- Not an online algorithm.
- Assumes MCMC kernels have uniform mixing with respect to time; probably not true unless one increases the computational effort with time.
- Can we just sample x_n (freezing the other coordinates)?

SMC sampler

- Consider example where $g(y_k|x_k) = exp\left(\sum_{j=1}^{d} h(y_k, x_{k,j})\right)$ and transition density $F(x_k|x_{k-1}) = \prod_{j=1}^{d} f(x_{k,j}|x_{k-1,j})$.
- In idealized case, we sample exactly from the final target density of the SMC sampler.
- This is the conditionally optimal proposal and the incremental weight is

$$\int_{E^d} g(y_n|x_n)F(x_n|x_{n-1}) = \prod_{j=1}^{d} \int_E e^{h(y_n,x_{n,j}} f(x_{n,j}|x_{n-1,j})dx_n, j.$$

- Then weights generally have exponentially increasing variance in d.

Daum-Huang filter

- Use log-homotopy to smoothly migrate the particles from the prior to the posterior.
- Flow of particles is similar to the flow in time induced by the Fokker-Planck equation.
- Since Bayes' rule operates at discrete points in time, it is difficult to create a flow in time.
- Insert a scalar valued parameter λ acting as time, which varies from 0 to 1 at each discrete time.

Daum-Huang filter

- Unnormalized Bayes' rule can be written as $p(x) = f(x)g(x)$
- Here $g(x) = p(x_k|y_{1:k-1})$ is the predicted prior density and $h(x) = p(y_k|x_k)$ is the likelihood.
- Take the logarithm of both sides: $\log(p(x)) = \log(f(x)) + \log(g(x))$.
- Then define a homotopy function: $\log(p(x,\lambda)) = \log(f(x)) + \lambda \log(g(x))$.

Summary

- Particle filter convergence depends heavily on the properties of the likelihood function and the Markov kernel.
- Best case: relatively flat likelihood and strongly mixing kernel.
- MSE converges at rate $\mathcal{O}(1/N)$.
- But: be careful of dimensionality!
- Number of particles required for given accuracy grows exponentially in the state dimension.
- No particle filtering algorithm has been proven stable as the dimension grows.
- Techniques like Daum-Huang offer a promising approach to mitigating effects of high-dimension.