Ergodicity and Accuracy in Optimal Particle Filters

David Kelly and Andrew Stuart

Courant Institute
New York University
New York NY
www.dtbkelly.com

November 29, 2016

Mathsfest 2016, *Canberra, Australia*

What is data assimilation?

What is data assimilation?

Suppose u satisfies an evolution equation

$$\frac{du}{dt} = F(u)$$

with some **unknown** initial condition $u_0 \sim \mu_0$.

There is a true trajectory of u that is producing *partial, noisy* observations at times $t = h, 2h, \dots, nh$:

$$y_n = Hu_n + \xi_n$$

where H is a linear operator (think low rank projection), $u_n = u(nh)$, and $\xi_n \sim N(0, \Gamma)$ iid.

The aim of **data assimilation** is to characterize the conditional distribution of u_n given the observations $\{y_1, \dots, y_n\}$

The conditional distribution is updated via the **filtering cycle**.

Illustration (Initialization)

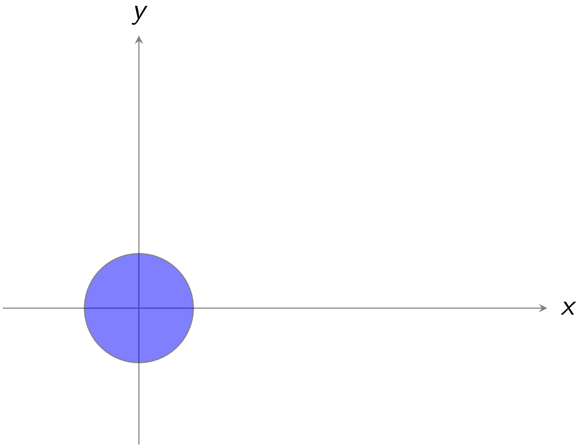


Figure: The blue circle represents our initial uncertainty $u_0 \sim \mu_0$.

Illustration (Forecast step)

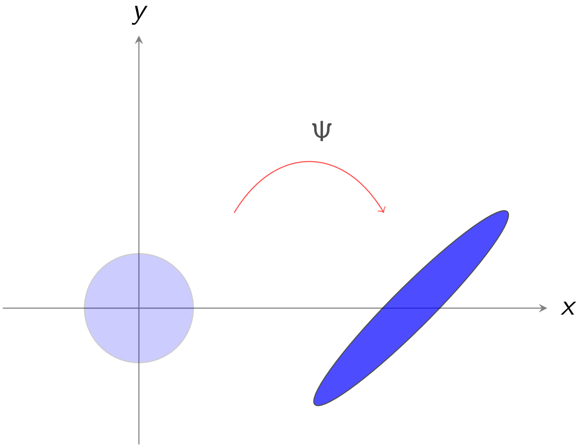


Figure: Apply the time h flow map Ψ. This produces a new probability measure which is our forecasted estimate of u_1. This is called the forecast step.

Illustration (Make an observation)

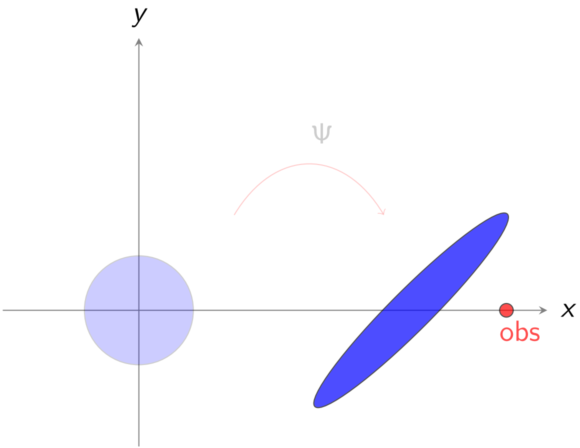


Figure: We make an observation $y_1 = Hu_1 + \xi_1$. In the picture, we only observe the x variable.

Illustration (Analysis step)

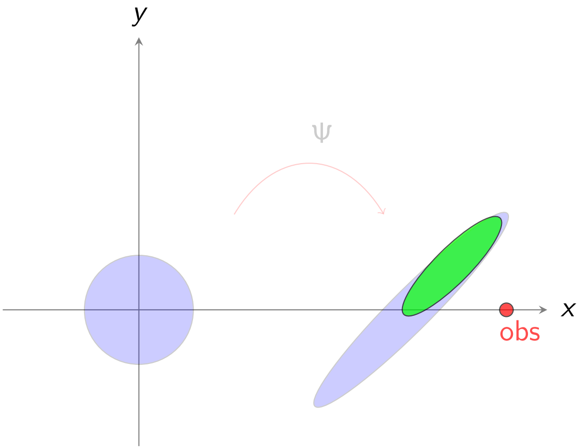


Figure: Using Bayes formula we compute the conditional distribution of $u_1|y_1$. This new measure (called the posterior) is the new estimate of u_1. The uncertainty of the estimate is reduced by incorporating the observation. The forecast distribution steers the update from the observation.

Bayes' formula filtering update

Let $Y_n = \{y_1, \ldots, y_n\}$. We want to compute the conditional density $\mathbf{P}(u_{n+1}|Y_{n+1})$, using $\mathbf{P}(u_n|Y_n)$ and y_{n+1}.

By Bayes' formula, we have

$$\mathbf{P}(u_{n+1}|Y_{n+1}) = \mathbf{P}(u_{n+1}|Y_n, y_{n+1}) \propto \mathbf{P}(y_{n+1}|u_{n+1})\mathbf{P}(u_{n+1}|Y_n)$$

But we need to compute the integral

$$\mathbf{P}(u_{n+1}|Y_n) = \int \mathbf{P}(u_{n+1}|Y_n, u_n)\mathbf{P}(u_n|Y_n)du_n \, .$$

In general there are **no closed formulas** for the Bayesian densities. One typically approximates the densities with a sampling procedure.

Applications can be very high dimensional (eg. robotics, numerical weather prediction) which makes the sampling problem highly non-trivial.

Particle filters approximate the posterior empirically

$$\mathbf{P}(u_k | Y_k) \approx \sum_{n=1}^{N} \frac{1}{N} \delta(u_k - u_k^{(n)})$$

the particles $\{u_k^{(n)}\}_{n=1}^{N}$ can be updated in different ways, giving rise to different particle filters.

Model Assumption

We always assume a **conditionally Gaussian** model

$$u_{k+1} = \psi(u_k) + \eta_k \quad , \quad \eta_k \sim N(0, \Sigma) \text{ i.i.d.}$$

where ψ is deterministic, with observations

$$y_{k+1} = Hu_{k+1} + \xi_{k+1} \quad , \quad \xi_k \sim N(0, \Gamma) \text{ i.i.d.}$$

This facilitates the implementation and theory for particles filters and is a realistic assumption for many practical problems.

We denote the posterior, with density $\mathbf{P}(u_k | Y_K)$, by μ_k and denote the particle filter approximations by μ_k^N.

The standard particle filter

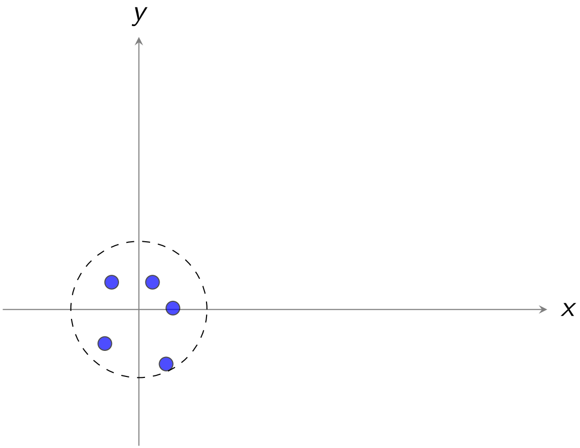


Figure: Start with N particles $\{u_k^{(n)}\}_{n=1}^N$ giving an empirical approx of μ_k.

$$\mathbf{P}(u_k | Y_k) \approx \frac{1}{N} \sum_{n=1}^{N} \delta(u_k - u_k^{(n)})$$

Apply dynamics to particles

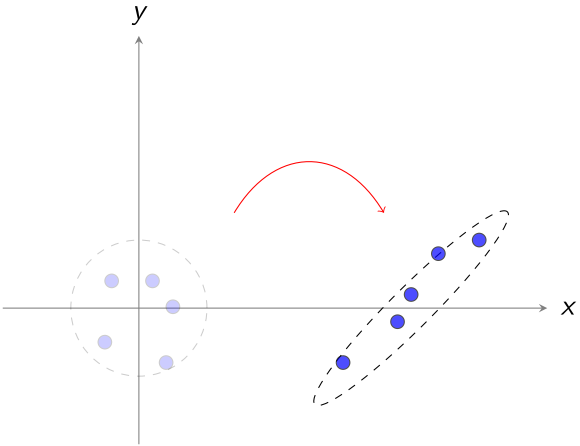
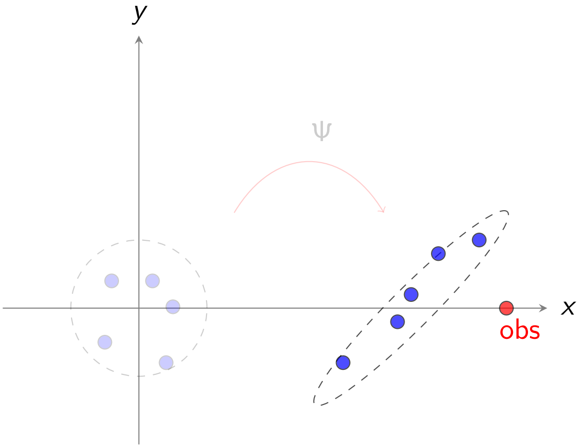


Figure: Apply the dynamics to each particle $\widehat{u}_{k+1}^{(n)} = \psi(u_k^{(n)}) + \eta_k^{(n)}$

$$\mathbf{P}(u_{k+1}|Y_k) \approx \sum_{n=1}^{N} \frac{1}{N}\delta(u_{k+1} - \widehat{u}_{k+1}^{(n)})$$

Make an observation



Weight the forecast particles

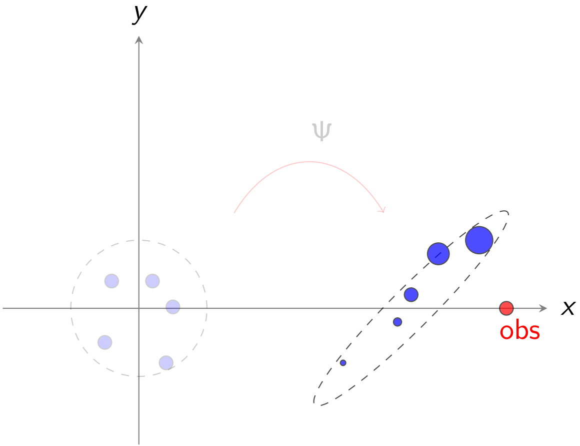


Figure: Assign weights $w_{k+1}^{(n)}$ to the particles, closer agreement with obs = larger weight. Weights are normalized $\sum_{n=1}^{N} w_{k+1}^{(n)} = 1$.

Re-sample the weighted particles

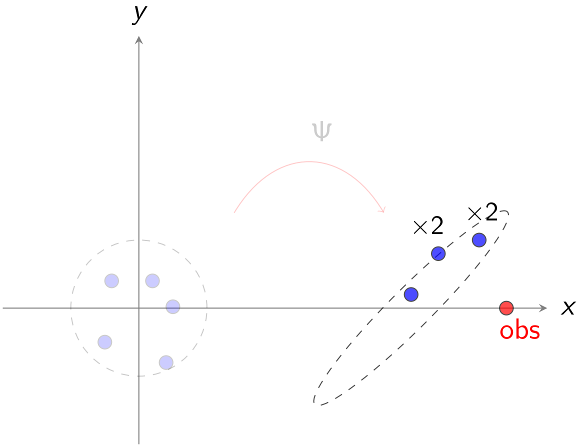


Figure: Sample $\{u_{k+1}^{(n)}\}_{n=1}^{N}$ from $\{\widehat{u}_{k+1}^{(n)}\}_{n=1}^{N}$ with weights $\{w_{k+1}^{(n)}\}_{n=1}^{N}$.

$$\mathbf{P}(u_{k+1}|Y_{k+1}) \approx \sum_{n=1}^{N} \frac{1}{N}\delta(u_{k+1} - u_{k+1}^{(n)})$$

The standard particle filter

We represent the standard particle filter as a random dynamical system

$$\widehat{u}_{k+1}^{(n)} = \psi(u_k^{(n)}) + \eta_k^{(n)}$$

$$u_{k+1}^{(n)} = \sum_{m=1}^{N} \mathbf{1}_{[x_{k+1}^{(m)}, x_{k+1}^{(m+1)})}(r_k^{(m)}) \widehat{u}_k^{(m)}$$

where $r_{k+1}^{(n)}$ is uniformly distributed on $[0,1]$ and

$$x_{k+1}^{(m+1)} = x_{k+1}^{(m)} + w_{k+1}^{(m)} \quad , \quad x_{k+1}^{(1)} = 0$$

ie. pick $\widehat{u}_{k+1}^{(m)}$ with probability $w_{k+1}^{(m)}$.

The motivation: importance sampling

If $p(x)$ is a probability density, the empirical approximation of p is given by

$$p(x) \approx \frac{1}{N}\sum_{n=1}^{N}\delta(x - x^{(n)})$$

where $x^{(n)}$ are samples from p.

When p is difficult to sample from, we can instead use the importance sampling approximation

$$p(x) \approx \frac{1}{N}\sum_{n=1}^{N}\frac{p(\widehat{x}^{(n)})}{q(\widehat{x}^{(n)})}\delta(x - \widehat{x}^{(n)})$$

where $\widehat{x}^{(n)}$ are samples from a different probability density q.

The motivation: standard particle filter

We have **samples** $\{u_k^{(n)}\}_{n=1}^N$ from $\mathbf{P}(u_k|Y_k)$ that we wish to update into **samples** from $\mathbf{P}(u_{k+1}|Y_{k+1})$.

Note that $u_k|Y_k$ is a Markov chain with kernel

$$p_k(u_k, du_{k+1}) = Z^{-1}\mathbf{P}(y_{k+1}|u_{k+1})\mathbf{P}(u_{k+1}|u_k)$$

If we could draw $u_{k+1}^{(n)}$ from $p_k(u_k^{(n)}, du_{k+1})$ then we would have $u_{k+1}^{(n)} \sim \mathbf{P}(u_{k+1}|Y_{k+1})$.

The motivation: standard particle filter

It is too difficult to sample directly, so we instead draw $\widehat{u}_{k+1}^{(n)}$ from $q(u_{k+1}) = \mathbf{P}(u_{k+1} | u_k^{(n)})$ and get the importance sampling approximation

$$\mathbf{P}(u_{k+1} | Y_{k+1}) \approx \frac{1}{N} \sum_{n=1}^{N} Z^{-1} \mathbf{P}(y_{k+1} | \widehat{u}_{k+1}^{(n)}) \delta(u_{k+1} - \widehat{u}_{k+1}^{(n)})$$

Since we cannot compute Z, approximate the weights by

$$w_{k+1}^{(n),*} = \mathbf{P}(y_{k+1} | \widehat{u}_{k+1}^{(n)}) \propto \exp\left(-\frac{1}{2} |y_{k+1} - H\widehat{u}_{k+1}^{(n)}|_\Gamma^2 \right)$$

$$w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{n=1}^{N} w_{k+1}^{(n),*}}$$

Notation: $|\cdot|_A = \langle A^{-1} \cdot, \cdot \rangle$

A different approach

Another approach

$$
\begin{aligned}
p_k(u_k^{(n)}, du_{k+1}) \\
&\propto \mathbf{P}(y_{k+1} | u_{k+1}) \mathbf{P}(u_{k+1} | u_k^{(n)}) \\
&= Z_\Gamma^{-1} \exp\left(-\frac{1}{2}|y_{k+1} - Hu_{k+1}|_\Gamma^2\right) Z_\Sigma^{-1} \exp\left(-\frac{1}{2}|u_{k+1} - \psi(u_k^{(n)})|_\Sigma^2\right) \\
&= Z_S^{-1} \exp\left(-\frac{1}{2}|y_{k+1} - H\psi(u_k^{(n)})|_S^2\right) Z_C^{-1} \exp\left(-\frac{1}{2}|u_{k+1} - m_{k+1}^{(n)}|_C^2\right)
\end{aligned}
$$

by product of Gaussian densities formulae , and

$$
\begin{aligned}
C^{-1} &= \Sigma^{-1} + H^T \Gamma^{-1} H \\
S &= H\Sigma H^T + \Gamma \\
m_{k+1}^{(n)} &= C(\Sigma^{-1}\psi(u_k^{(n)}) + H^T\Gamma^{-1} y_{k+1}) = (I - KH)\psi(u_k) + K y_{k+1}
\end{aligned}
$$

If $q(u_{k+1}) = Z_C^{-1} \exp\left(-\frac{1}{2}|u_{k+1} - m_{k+1}^{(n)}|_C^2\right)$ then the importance sampling approximation is given by

$$\mathbf{P}(u_{k+1}|Y_{k+1}) \approx \frac{1}{N}\sum_{n=1}^{N} Z^{-1} \exp\left(-\frac{1}{2}|y_{k+1} - H\psi(u_k^{(n)})|_S^2\right)\delta(u_{k+1} - \widehat{u}_{k+1}^{(n)})$$

where $\widehat{u}_{k+1}^{(n)}$ are sampled from q, ie

$$\widehat{u}_{k+1}^{(n)} = m_{k+1}^{(n)} + \zeta_{k+1}^{(n)} = (I - KH)\psi(u_k^{(n)}) + Ky_{k+1} + \zeta_{k+1}^{(n)}$$

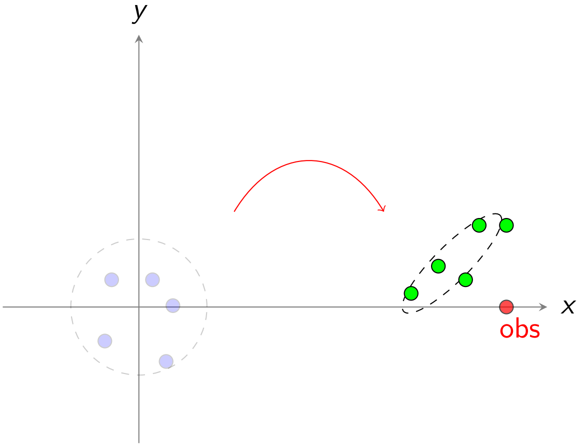
where $\zeta_{k+1}^{(n)} \sim N(0, C)$.

Since we cannot compute Z, approximate the weights by

$$w_{k+1}^{(n),*} = \exp\left(-\frac{1}{2}|y_{k+1} - H\psi(u_k^{(n)})|_S^2\right) \quad , \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{n=1}^{N} w_{k+1}^{(n),*}}$$

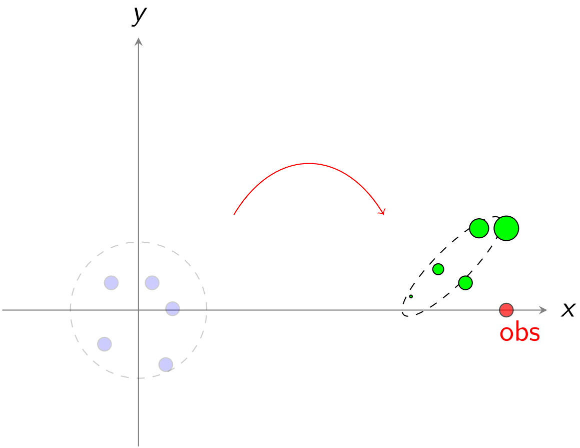
The optimal particle filter

Propagate the particles



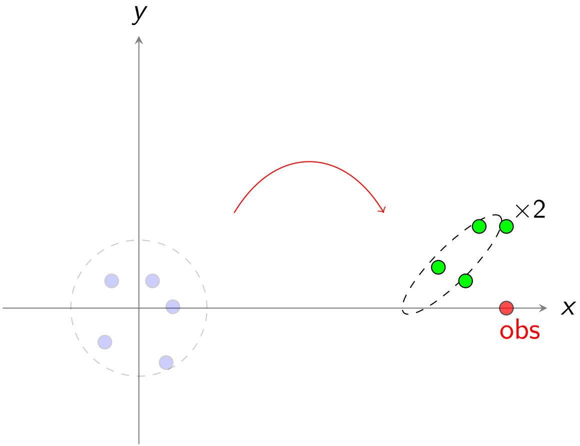
$$\widehat{u}_{k+1}^{(n)} = (I - KH)\psi(u_k^{(n)}) + K y_{k+1} + \zeta_{k+1}^{(n)} \quad , \quad \zeta_{k+1}^{(n)} \sim N(0, C) \text{ i.i.d.}$$

Weight the particles using the observation



$$w_{k+1}^{(n),*} = \exp\left(-\frac{1}{2}|y_{k+1} - H\psi(u_k^{(n)})|_S^2\right) \quad , \quad w_{k+1}^{(n)} = \frac{w_{k+1}^{(n),*}}{\sum_{n=1}^{N} w_{k+1}^{(n),*}}$$

Resample the weighted particles



$$\mathbf{P}(u_{k+1} | Y_{k+1}) \approx \sum_{n=1}^{N} \frac{1}{N} \delta(u_{k+1} - u_{k+1}^{(n)})$$

The optimal particle filter

We represent the optimal particle filter as a random dynamical system

$$\widehat{u}_{k+1}^{(n)} = (I - KH)\psi(u_k^{(n)}) + K y_{k+1} + \zeta_k^{(n)} \quad , \quad \zeta_k^{(n)} \sim N(0, C) \text{ i.i.d.}$$

$$u_{k+1}^{(n)} = \sum_{m=1}^{N} \mathbf{1}_{[x_{k+1}^{(m)}, x_{k+1}^{(m+1)}]}(r_{k+1}^{(n)}) \widehat{u}_{k+1}^{(m)} .$$

where $r_{k+1}^{(n)}$ is uniformly distributed on $[0, 1]$ and

$$x_{k+1}^{(m+1)} = x_{k+1}^{(m)} + w_{k+1}^{(m)}$$

ie. pick $\widehat{u}_{k+1}^{(m)}$ with probability $w_{k+1}^{(m)}$.

Note that $U_k = (u_k^{(1)}, \dots, u_k^{(n)})$ is a Markov chain.

What do we know about particle filters?

Theory for filtering distributions

The true posterior (filtering distribution) μ_k is known to be **accurate**:

$$\limsup_{k\to\infty} \mathbf{E}\|m_k - u_k\|^2 = O(\text{obs noise})$$

where u_k is the trajectory producing Y_k, $m_k = \mathbf{E}(u_k|Y_k)$ and we take $\mathbf{E}$ over all randomness.

And **conditionally ergodic**: If μ'_k, μ''_k are two copies of the filtering distribution with $\mu'_0 = \delta_{u'_0}$ and $\mu''_0 = \delta_{u''_0}$ then

$$d_{TV}(\mu'_k, \mu''_k) = O(\delta^k)$$

as $k \to \infty$, where $\delta \in (0,1)$.

Consistency of particle filters

Most particle filters (including the standard and optimal PFs) are **consistent**:

The empirical measure converges to the true filtering distribution and moreover

$$d(\mu_k^N, \mu_k) \leq C_{d,k} N^{-1/2}$$

But the constant $C_{d,k}$ scales badly with dimension.

eg. (Bickel et al) For a class of linear models, if $d \to \infty$ then we must have$N \geq C \exp(d)$ for consistency.

Works better than consistency theory suggests

Figure: Lorenz equations, only observing x variable. Particle filter with $N = 5$ exhibits **accuracy** and **forgets its initialization**.

Theory for optimal particle filter with fixed N.

Accuracy

Assumption The map $(I - KH)\psi(\cdot)$ is a contraction wrt some norm $\|\cdot\|$. (generalization of observability to nonlinear systems)

Theorem (K, Stuart 16)

$$\limsup_{k\to\infty} \mathbf{E} \max_n \| u_k^{(n)} - u_k \|^2 = O(obs\ noise)$$

for each $n = 1, \ldots, N$.

Accuracy Proof

Let $e_k^{(n)} = u_k^{(n)} - u_k$ and $\widehat{e}_k^{(n)} = \widehat{u}_k^{(n)} - u_k$ then

$$\begin{aligned}
\widehat{e}_{k+1}^{(n)} &= (I - KH)\psi(u_k^{(n)}) + K y_{k+1} + \zeta_{k+1}^{(n)} - (\psi(u_k) + \eta_k) \\
&= (I - KH)(\psi(u_k^{(n)}) - \psi(u_k)) + K(y_{k+1} - H\psi(u_k)) \\
&+ (\zeta_{k+1} - \eta_k) \\
&= (I - KH)(\psi(u_k^{(n)}) - \psi(u_k)) + (K\xi_{k+1} + \zeta_{k+1}^{(n)} - \eta_k)
\end{aligned}$$

Note that all the noises are independent, and by the contraction assumption $\|(I - KH)(\psi(u_k^{(n)}) - \psi(u_k))\| \leq \alpha \|e_k^{(n)}\|$ for some $\alpha \in (0, 1)$.

Accuracy Proof

And moreover

$$e_{k+1}^{(n)} = \sum_{m=1}^{N} 1_{[x_{k+1}^{(m)}, x_{k+1}^{(m+1)})}(r_{k+1}^{(n)}) \widehat{e}_{k+1}^{(m)}$$

and note that exactly one of the terms in the sum is non-zero.

It follows easily that $\max_n \|e_{k+1}^{(n)}\| \leq \max_n \|\widehat{e}_{k+1}^{(n)}\|$.

From the contraction assumption and independence it follows that

$$\mathbf{E} \max_n \|e_{k+1}^{(n)}\|^2 \leq \alpha \mathbf{E} \max_n \|e_k^{(n)}\|^2 + \beta$$

for $\alpha \in (0,1)$ and $\beta > 0$. Result follows by Gronwall.

Conditional Ergodicity: Preliminaries

Let $U'_k = (u_k^{(1)'}, \ldots, u_k^{(N)'})$ and $U''_k = (u_k^{(1)''}, \ldots, u_k^{(N)''})$ be two optimal PFs driven by the same observations Y_k, but with different initializations u'_0 and u''_0 respectively.

Recall that these are Markov chains, and denote their transition kernels by $p_{k+1}(U_k, \cdot)$. Denote the law of U_k by $p^k(U_0, \cdot)$

Conditional Ergodicity: The result

Theorem (K, Stuart 16)

The optimal PF is conditionally ergodic in the sense that

$$d_{TV}\left(p^k(U_0', \cdot), p^k(U_0'', \cdot)\right) = O(\delta^k)$$

as $k \to \infty$, for $\delta \in (0, 1)$.

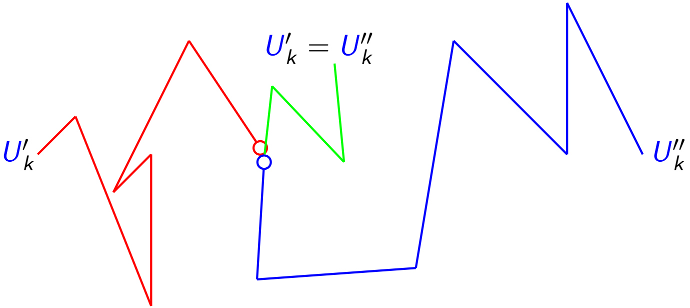
ie. The optimal PF forgets its initialization (in a weak sense) exponentially quickly.

Proof

A coupling (U'_k, U''_k) is any joint distribution whose the marginals of the law of (U'_k, U''_k) are $p^k(U'_0, \cdot)$ and $p^k(U''_0, \cdot)$ respectively.

We consider the coupling (U'_k, U''_k) defined in such a way that $U'_k = U''_k$ for all $k \geq k^*$ where k^* is the random time $k^* = \inf\{k : U'_k = U''_k\}$.

Let A_k be the event that $k^* > k$.



Proof

By definition of the TV metric

$$
\begin{aligned}
& d_{TV}(p^k(U_0', \cdot), p^k(U_0'', \cdot)) \\
& \quad = \frac{1}{2} \sup_{|f| \leq 1} \left| \mathbf{E} f(U_k') - \mathbf{E} f(U_k'') \right| \\
& \quad = \frac{1}{2} \sup_{|f| \leq 1} \left| \mathbf{E}(f(U_k') - \mathbf{E} f(U_k'')) I_{A_k} + \mathbf{E}(f(U_k') - \mathbf{E} f(U_k'')) I_{A_k^c} \right| \\
& \quad = \frac{1}{2} \sup_{|f| \leq 1} \left| \mathbf{E}(f(U_k') - \mathbf{E} f(U_k'')) I_{A_k} \right| \leq \mathbf{P}(A_k)
\end{aligned}
$$

So we want to construct a coupling (U_k', U_k'') that couples quickly (probability of not yet coupling decays rapidly).

Proof

Assume first that we have a minorization condition for the kernel $p_k(U, \cdot)$: there exists a probability measure ν and constant $\varepsilon_k \in (0, 1)$ such that

$$p_k(U, \cdot) \geq \varepsilon_k \nu(\cdot)$$

for all U and each k.

Quite easy to verify that (given some natural assumptions on ψ) the optimal PF satisfies this condition with Gaussian ν and ε_k depending on d, N, Y_k.

Proof

The minorization condition allows us to build a Markov chain $\tilde{U}_k$ with kernel

$$\tilde{p}_k(U, \cdot) = (1 - \varepsilon_k)^{-1}(p_k(U, \cdot) - \varepsilon_k \nu(\cdot))$$

We can now represent the Markov chain U_k in the following split chain sense:

$$U_k = \begin{cases} \tilde{U}_k & \text{with probability } 1 - \varepsilon_k \\ \xi & \text{with probability } \varepsilon_k \end{cases}$$

where $\xi \sim \nu(\cdot)$.

On can easily check (for instance by evaluating $\mathbf{E}f(U_k)$) that this does indeed yield a copy of the optimal PF Markov chain.

We define **the coupling** (U'_k, U''_k) using independent copies of $\tilde{U}_k$ but the same ε_k-coin and identical copies of ξ in the split-chain representation.

When the coin lands $(1-\varepsilon_k)$, the two chains evolve independently, but as soon as the coin lands ε_k we have $U'_k = U''_k$.

It follows that

$$d_{TV}(p^k(u'_0, \cdot), p^k(u''_0, \cdot)) \leq \mathbf{P}(A_k) = \Pi_{i=1}^{k}(1-\varepsilon_i)$$

(After filling in a few details) □

References

D. Kelly & A. Stuart. *Ergodicity and Accuracy of Optimal Particle Filters for Bayesian Data Assimilation*. **arXiv** (2016).

All my slides are on my website (www.dtbkelly.com) **Thank you**!