Limit theorems for sequential MCMC methods

Axel Finke\textsuperscript{1} \quad Arnaud Doucet\textsuperscript{2} \quad Adam M. Johansen\textsuperscript{3}

\textsuperscript{1}DSAP, NUS \quad \textsuperscript{2}Dept of Statistics, University of Oxford \quad \textsuperscript{3}Dept of Statistics, University of Warwick; The Alan Turing Institute

4th September 2018
Standard PFs and MCMC-PFs

Convergence analysis of MCMC-PFs

Application to state-space models

Numerical illustrations
Summary

- **Sequential MCMC methods a.k.a. MCMC-PFs:**
  - proposed in Berzuini et al. (1997), extended in Septier et al. (2009); Septier and Peters (2016); Finke et al. (2016).

- **Difference with (standard) particle filters (PFs):**
  - PFs sample/resample particles conditionally independently,
  - MCMC-PFs sample/resample particles jointly according to a Markov chain.

- **This work:**
  - convergence analysis of MCMC-PFs;
  - guidance on when to use standard PFs/MCMC-PFs
Path-space Feynman–Kac model

Setup & notation:

- **path-space** formulation:
  \( x_n := x_{1:n} = (x_{n-1}, x_n) \in E_n := E_{n-1} \times E, \)
- mutation kernels: \( M_n(x_{n-1}, dx_n), \)
- bounded potential functions: \( G_n(x_n) \in (0, 1]. \)

Goal: approximate distributions \((\eta_n)_{n\geq 1}\) on \((E_n)_{n\geq 1}\):

\[
\eta_n(dx_n) \propto \gamma_n(dx_n) := \eta_1 Q_{1,n}(dx_n),
\]

\[
Q_{p,n}(dx_n)(x_p) := \prod_{q=p}^n G_{q-1}(x_{q-1}) M_q(x_{q-1}, dx_q),
\]

- unknown normalising constant: \( Z_n := \gamma_n(1), \)
- recursive definition: \( \eta_n = \Phi_n^{\eta_{n-1}}, \) where

\[
\Phi_n^\mu(dx_n) := \frac{G_{n-1}(x_{n-1})}{\mu(G_{n-1})} [\mu \otimes M_n](dx_n).
\]
Example: Bootstrap PF flow I

State-space model:

- Bivariate Markov chain \((X_n, Y_n)_{n \in \mathbb{N}}\),
- with transition kernel \(f(dx_n|x_{n-1})g(y_n|x_n)dy_n\),
- only \(Y_n = y_n\) is observed; \(X_n\) is latent.

\[
\begin{align*}
Y_{n-1} & \quad \xleftarrow{g} \quad Y_n \\
X_{n-2} \quad f \quad X_{n-1} \quad f \quad X_n \quad f \quad X_{n+1}
\end{align*}
\]
Example: Bootstrap PF flow II

Take

\[ G_n(x_n) := g(y_n|x_n), \]
\[ M_n(x_{n-1}, dx_n) := f(dx_n|x_{n-1}). \]

Then

\[ \eta_n(dx_n) = p(dx_{1:n}|y_{1:n-1}) \]
\[ = \Phi_n^{\eta_{n-1}}(dx_n) \]
\[ = \frac{\eta_{n-1}(dx_{n-1})G_{n-1}(x_{n-1})}{\eta_{n-1}(G_{n-1})}M_n(x_{n-1}, dx_n) \]
\[ = \frac{p(dx_{1:n-1}|y_{1:n-2})g(y_{n-1}|x_{n-1})}{\int p(dx'_{1:n-1}|y_{1:n-2})g(y_{n-1}|x'_{n-1})f(dx_n|x_{n-1})}f(dx_n|x_{n-1}), \]

as well as \( \mathcal{Z}_n = p(y_{1:n-1}) \).
• Problem: $\eta_n$ is intractable.
• Idea: recursively construct approximation $\eta_n^N$ of $\eta_n = \Phi^{\eta_{n-1}}_n$.
  1. given $\eta_{n-1}^N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_{n-1}^i}$, obtain the mixture
     $$
     \Phi^{\eta_{n-1}^N}_n = \sum_{i=1}^{N} \frac{G_{n-1}(\xi_{n-1}^i)}{\sum_{j=1}^{N} G_{n-1}(\xi_{n-1}^j)} \left[ \delta_{\xi_{n-1}^i} \otimes M_n \right],
     $$
  2. sample $N$ particles $\xi_1^N \ldots, \xi_N^N$ (approximately) from $\Phi^{\eta_{n-1}^N}_n$,
  3. approximate $\eta_n$ by $\eta_{n}^N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i^n}$.

Algorithm (PF). In Step 2, sample $\xi_1^N \ldots, \xi_N^N \overset{iid}{\sim} \Phi^{\eta_{n-1}^N}_n$.

Algorithm (MCMC-PF). In Step 2,
- initialise $\xi_1^n \sim \kappa_{n-1}^{\eta_n^N} \approx \Phi^{\eta_{n-1}^N}_n$,
- sample $\xi_i^n \sim K_{n-1}^{\eta_n^N}(\xi_{i-1}^N, \cdot)$, for $i = 2, \ldots, N$.

\(\Phi^{\eta_{n-1}^N}_n\)-invariant MCMC kernel
Standard PFs and MCMC-PFs

Convergence analysis of MCMC-PFs

Application to state-space models

Numerical illustrations
• Recall: $\eta_n(dx_n) = \gamma_n(dx_n)/\mathcal{Z}_n$.

• Usual estimates of $\gamma_n(\varphi_n)$ and $\mathcal{Z}_n$:

\[
\gamma^N_n(\varphi_n) := \eta^N_n(\varphi_n) \prod_{p=1}^{n-1} \eta^N_p(G_p),
\]
\[
\mathcal{Z}^N_n := \gamma^N_n(1) = \prod_{p=1}^{n-1} \frac{1}{N} \sum_{i=1}^{N} G_p(\xi^i_p).
\]

**Proposition (unbiasedness).** For any $n \geq 1$, $N \geq 1$ and $\varphi_n \in \mathcal{B}(E_n)$, if the chains are initialised from stationarity, i.e. if $\kappa^\mu_p = \Phi^\mu_p$ for $1 \leq p \leq n$,

1. $\mathbb{E}[\gamma^N_n(\varphi_n)] = \gamma_n(\varphi_n)$,
2. $\mathbb{E}[\mathcal{Z}^N_n] = \mathcal{Z}_n$. 

Assumptions

A1 For any \( n \geq 1 \), there exists \( i_n \in \mathbb{N} \) such that
\[
\sup_{\mu \in \mathcal{P}(E_{n-1})} \beta((K_n^\mu)^{i_n}) < 1,
\]
where \( \beta(K) := \sup_{x,y} \|K(x, \cdot) - K(y, \cdot)\| \).

A2 For any \( n \geq 1 \), there exists a constant \( \bar{\Gamma}_n < \infty \) and a family of bounded integral operators \((\Gamma_n^\mu)_{\mu \in \mathcal{P}(E_{n-1})}\) from \( \mathcal{B}(E_{n-1}) \) to \( \mathcal{B}(E_n) \) s.t. for any \((\mu, \nu) \in \mathcal{P}(E_{n-1})^2\) and any \( f_n \in \mathcal{B}(E_n) \),
\[
\|[K_n^\mu - K_n^\nu](f_n)\| \leq \int_{\mathcal{B}(E_{n-1})} |[\mu - \nu](g)| \Gamma_n^\mu(f_n, dg)
\]
and
\[
\int_{\mathcal{B}(E_{n-1})} g \|\Gamma_n^\mu(f_n, dg) \leq \|f_n\| \bar{\Gamma}_n.
\]

- strong but similar to assumptions in Del Moral and Doucet (2010); Brockwell et al. (2010); Bercu et al. (2012),
- satisfied, e.g. if \( K_n^\mu \) is an independent MH kernel & \( E \) finite.
Proposition ($\mathbb{L}_p$-error bound). Under $A1$, for any $n, p \geq 1$, there exist $a_n, b_p < \infty$ such that for any $\varphi_n \in B(E_n)$ and any $N \geq 1$,

$$
\mathbb{E}
\left[
\left[
\left|
\eta_n^N - \eta_n
\right|
\left|
\varphi_n
\right|
\right]^p
\right]^{\frac{1}{p}} \leq \frac{a_n b_p}{\sqrt{N}} \|\varphi_n\|.
$$

- Under strong mixing assumptions and if $\varphi_n(x_{1:n}) = \varphi(x_n)$,
  $$\sup_{n \geq 1} a_n < \infty.$$

Corollary (strong law of large numbers). Under $A1$, for any $n \geq 1$ and $\varphi_n \in B(E_n)$, as $N \to \infty$,

1. $\gamma_n^N(\varphi_n) \to \text{a.s.} \gamma_n(\varphi_n)$,
2. $\eta_n^N(\varphi_n) \to \text{a.s.} \eta_n(\varphi_n)$.
For any $\nu$-invariant Markov kernel $K$, define the integrated autocorrelation time

$$iact_K[\varphi] := 1 + 2 \sum_{l=1}^{\infty} \frac{\text{cov}_\nu[\varphi, K^l(\varphi)]}{\text{var}_\nu[\varphi]}.$$ 

Proposition (central limit theorem). Under A1–A2, for any $n \geq 1$ and any $\varphi_n \in \mathcal{B}(E_n)$, as $N \to \infty$,

1. $\sqrt{N}[\gamma_n^N/\gamma_n(1) - \eta_n](\varphi_n) \to_d N(0, \sigma_n^2[\varphi_n])$,

2. $\sqrt{N}[\eta_n^N - \eta_n](\varphi_n) \to_d N(0, \sigma_n^2[\varphi_n - \eta_n(\varphi_n)])$,

with asymptotic variance

$\sigma_n^2[\cdot] := \sum_{p=1}^{n} \text{var}_{\eta_p}[\tilde{Q}_{p,n}(\cdot)] \times iact_{K_{\eta p}^{-1}}[\tilde{Q}_{p,n}(\cdot)].$ 

Here, $\tilde{Q}_{p,n} := \frac{\gamma_n(1)}{\gamma_p(1)} Q_{p,n}$ satisfies $\eta_p \tilde{Q}_{p,n} = \eta_n$.

- Under strong mixing assumptions and if $\varphi_n(x_{1:n}) = \varphi(x_n)$, $\sup_{n \geq 1} \sigma_n^2[\varphi_n - \eta_n(\varphi_n)] < \infty$. 

Standard PFs and MCMC-PFs

Convergence analysis of MCMC-PFs

Application to state-space models

Numerical illustrations
State-space model

- Bivariate Markov chain \((X_n, Y_n)_{n \in \mathbb{N}}\),
- with transition kernel \(f(dx_n|x_{n-1})g(y_n|x_n)dy_n\),
- only \(Y_n = y_n\) is observed; \(X_n\) is latent.
Bootstrap PF (BPF)-type flow

Example (BPF flow).

\[
G_{n-1}(x_{n-1}) := g(y_{n-1}|x_{n-1}), \\
M_n(x_{n-1}, dx_n) := f(dx_n|x_{n-1}).
\]

In this case, \( \eta_n(dx_n) = p(dx_{1:n}|y_{1:n-1}) \), \( \mathcal{Z}_n = p(y_{1:n-1}) \) and

\[
\Phi_{n}^{\eta_n^{N}}(dx_n) = \sum_{i=1}^{N} \frac{g(y_{n-1}|\xi^{i}_{n-1})}{\sum_{j=1}^{N} g(y_{n-1}|\xi^{j}_{n-1})} \delta_{\xi^{i}_{n-1}}(dx_{n-1}) f(dx_n|\xi^{i}_{n-1}).
\]

⇒ can typically implement both BPF and MCMC-BPF.
Example (FA-APF flow).

\[ G_{n-1}(x_{n-1}) := p(y_n | x_{n-1}) = \int g(y_n | x_n) f(dx_n | x_{n-1}) \]

\[ M_n(x_{n-1}, dx_n) := p(dx_n | y_n, x_{n-1}) := \frac{g(y_n | x_n) f(dx_n | x_{n-1})}{p(y_n | x_{n-1})}. \]

In this case, \( \eta_n(dx_n) = p(dx_{1:n} | y_{1:n}) \), \( Z_n = p(y_{1:n}) \) and

\[ \Phi_{n}^{\eta_{n-1}}(dx_n) = \sum_{i=1}^{N} \frac{p(y_n | \xi^i_{n-1})}{\sum_{j=1}^{N} p(y_n | \xi^j_{n-1})} \delta_{\xi^i_{n-1}}(x_{n-1})p(dx_n | y_n, \xi^i_{n-1}) \]

\[ \propto \sum_{i=1}^{N} g(y_n | x_n) \delta_{\xi^i_{n-1}}(dx_{n-1}) f(dx_n | \xi^i_{n-1}). \]

\( \Rightarrow \) can typically implement MCMC-FA-APF but not FA-APF.
### Variance–variance trade-off

<table>
<thead>
<tr>
<th></th>
<th>BPF flow</th>
<th>FA-APF flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard PF</td>
<td>BPF</td>
<td>MCMC-FA-APF (usually intractable)</td>
</tr>
<tr>
<td>MCMC-PF</td>
<td>MCMC-BPF</td>
<td>MCMC-FA-APF</td>
</tr>
<tr>
<td></td>
<td>(not very useful)</td>
<td></td>
</tr>
</tbody>
</table>

- PFs preferable if they target the same distribution flow.
- MCMC-PFs preferable if
  - they can target a more efficient distribution flow,
  - the MCMC kernels do not mix too poorly.

**Trade-off:** variance due to importance-sampling vs. variance due to additional particle (auto-)correlation.
Standard PFs and MCMC-PFs

Convergence analysis of MCMC-PFs

Application to state-space models

Numerical illustrations
Binary state-space model

Asymptotic variances (relative to the asymptotic variance of the BPF).
Estimates of the marginal likelihood (relative to the true marginal likelihood) using $N = 10,000$ particles.
Ongoing work

With Alex Thiery:

- additional dependence between particles may be more useful within ‘conditional’ SMC algorithms,
- permits ‘local’ conditional SMC algorithms,
  - better scaling in high dimensions,
  - example: embedded hidden Markov model method (Shestopaloff and Neal, 2018) which is the conditional SMC version of MCMC-PFs.
- more on this in my talk next week.


