# On the forgetting of particle filters

Matti Vihola (Jyväskylä)

Third workshop on Monte Carlo methods in Warsaw
15 Dec 2023

Joint work with: Joona Karjalainen (Jyväskylä)
Anthony Lee (Bristol)
Sumeetpal S. Singh (Wollongong)

FINNISH CENTRE OF EXCELLENCE IN RANDOMNESS AND STRUCTURES 2022–2029

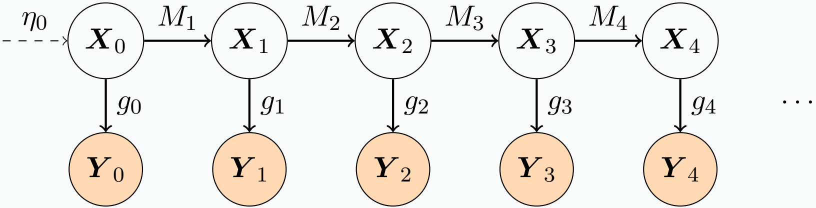
Research Council of Finland

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS AND STATISTICS

# Outline

# Introduction: Hidden Markov Model and filtering

# Hidden Markov model (a.k.a. 'state-space model')




- Hidden (unobserved) Markov state process $(\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_3, \ldots)$:
  - Initial density $\eta_0(\boldsymbol{x}_0)$
  - Transition densities $M_t(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$
- Observations (or measurements) $(\boldsymbol{Y}_0, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots)$:
  - Conditionally independent given $(\boldsymbol{X}_t)_{t\geq 0}$
  - Observation densities $g_t(\boldsymbol{y}_t \mid \boldsymbol{x}_t)$
  - Observed values $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots \rightsquigarrow$ potentials $G_t(\boldsymbol{x}_t) = g_k(\boldsymbol{y}_t \mid \boldsymbol{x}_t)$

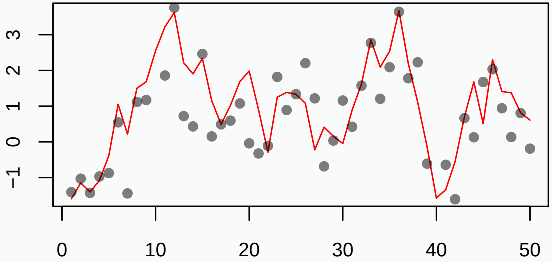
# Running example: Noisy AR(1)

$X_{0:T}$ stationary AR(1) process:

- $X_1 \sim N(0, \sigma_\eta^2/(1-\phi^2))$.
- $X_k = \phi X_{k-1} + \eta_k;\ \eta_k \sim N(0, \sigma_\eta^2)$

$Y_{0:T-1}$ noisy observations of $X_{1:T-1}$:

- $Y_k \sim N(X_k, \sigma_Y^2)$
- $G_k(x_k) = c_Y \exp\left(-\frac{1}{2\sigma_Y^2}(x_k - y_k)^2\right)$



# The filtering problem

- We are interested in:
  - predictive distributions $\eta_t = \mathrm{Law}(\boldsymbol{X}_t \mid \boldsymbol{y}_{0:t-1})$
  - filtering distributions $\pi_t = \mathrm{Law}(\boldsymbol{X}_t \mid \boldsymbol{y}_{0:t})$
- Practical example: GPS navigation
  - varying quality GPS observations $\rightsquigarrow G_t$
  - simple model of movement such as Brownian velocity $\rightsquigarrow M_t$
- Cannot determine $\eta_t$ and $\pi_t$ in a closed form

  (essentially unless $M_t$ and $G_t$ linear-Gaussian or state-space finite)
- We focus on Monte Carlo approximations, that is, sampling number of *particles* $\boldsymbol{X}_t^{1:N}$ "approximately from $\eta_t$", iteratively in $t = 0, 1, 2, \ldots$

# The ideal filter

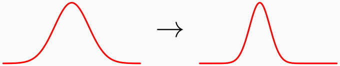
- Update $\color{red}{\eta_t} \to \color{blue}{\pi_t}$ by weighting with $G_t$:

$$\pi_t(\boldsymbol{x}_t) = \Psi_t(\eta_t)(\boldsymbol{x}_t) \qquad \text{where} \qquad \Psi_t(\mu)(\boldsymbol{x}) = \frac{\mu(\boldsymbol{x})G_t(\boldsymbol{x})}{\int G_t(\boldsymbol{z})\mu(\boldsymbol{z})\mathrm{d}\boldsymbol{z}}$$

- Mutate $\color{blue}{\pi_t} \to \color{red}{\eta_{t+1}}$ by pushing through $M_{t+1}$:

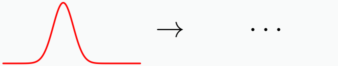
$$\eta_{t+1}(\boldsymbol{x}_{t+1}) = (\pi_t M_{t+1})(\boldsymbol{x}_{t+1}) = \int M_{t+1}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t)\pi_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t$$

- We denote the composition of the above by $\Phi_t$:



$$\eta_{t+1} = \Phi_{t+1}(\eta_t) = \Psi_t(\eta_t)M_{t+1}$$

and compositions of these by $\Phi_{t,u}$:



$$\eta_u = \Phi_{t,u}(\eta_t) = \Phi_u \circ \cdots \circ \Phi_{t+1}(\eta_t)$$

# Particle filter

# Particle filter algorithm

Gordon, Salmond and Smith (*IEE Proc. F*, 1993)

PF($\eta_0, (M_t)_{t\geq 1}, (G_t)_{t\geq 0}, N$)

$X_0^i \sim \eta_0(\,\cdot\,)$    for $i \in \{1{:}N\}$

for $t = 1, 2, \ldots$ do

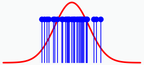
$\quad W_{t-1}^i = \dfrac{G_{t-1}(\boldsymbol{X}_{t-1}^i)}{\sum_{j=1}^N G_{t-1}(\boldsymbol{X}_{t-1}^j)}$    for $i \in \{1{:}N\}$

$\quad A_{t-1}^i \sim \mathrm{Categorical}\big(W_{t-1}^{1:N}\big)$    for $i \in \{1{:}N\}$

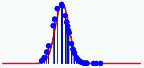
$\quad \boldsymbol{X}_t^i \sim M_t(\,\cdot \mid \boldsymbol{X}_{t-1}^{A_{t-1}^i})$    for $i \in \{1{:}N\}$

end for

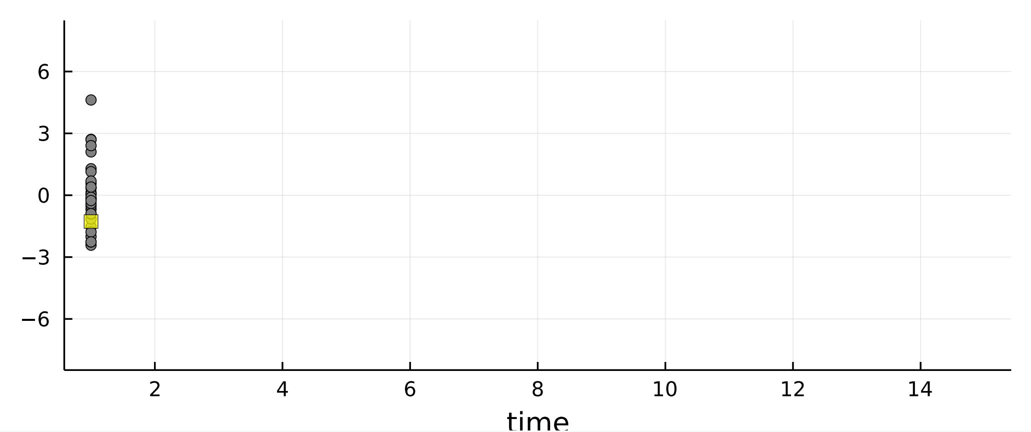
Produces empirical approximations of $\eta_t$ and $\pi_t$:



$$\eta_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{X}_t^i} \qquad \text{and} \qquad \pi_t^N = \Psi_t(\eta_t^N) = \sum_{i=1}^N W_t^i \delta_{\boldsymbol{X}_t^i}$$

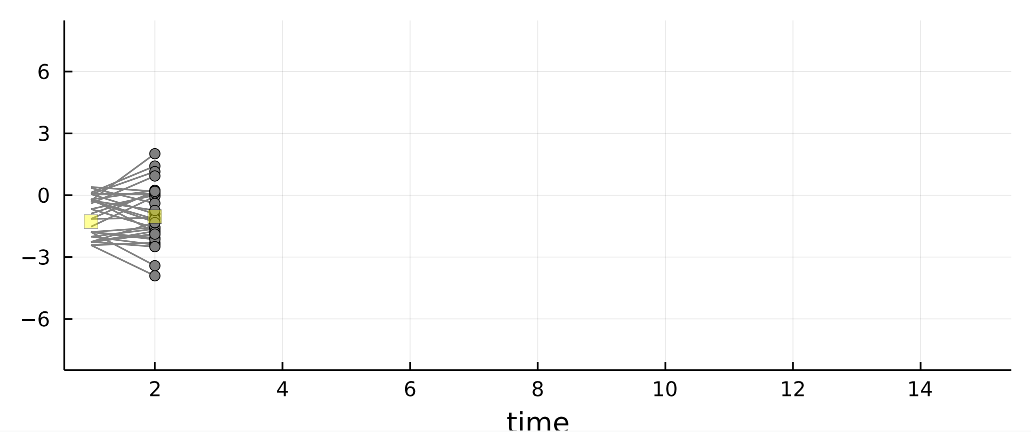


# Particle filter on noisy AR(1): Initialise



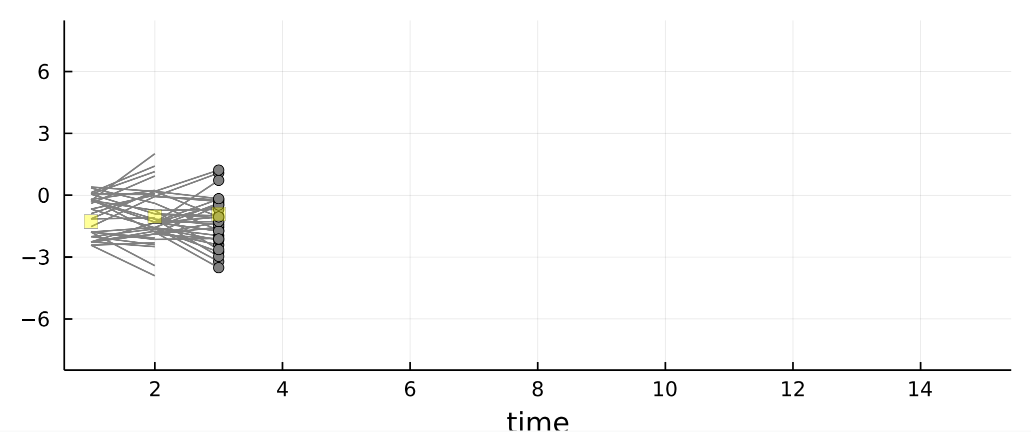

- ● Particles $X_1^{1:N} \sim M_1(\,\cdot\,)$
- ■ Observation $y_1$

# Particle filter on noisy AR(1): Resample and propagate



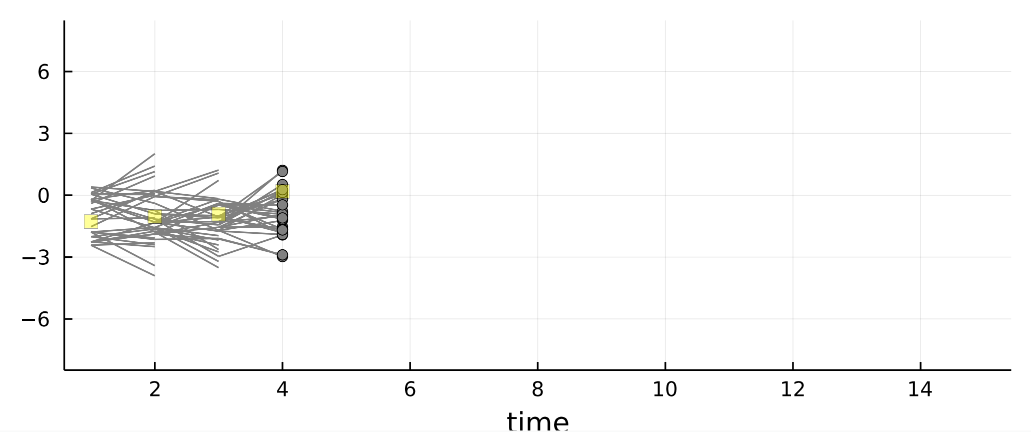

- ● Particles $X_t^i \sim M_t(X_{t-1}^{A_{t-1}^i}, \cdot\,)$
- ■ Observations $y_{0:t}$

# Particle filter on noisy AR(1): Resample and propagate



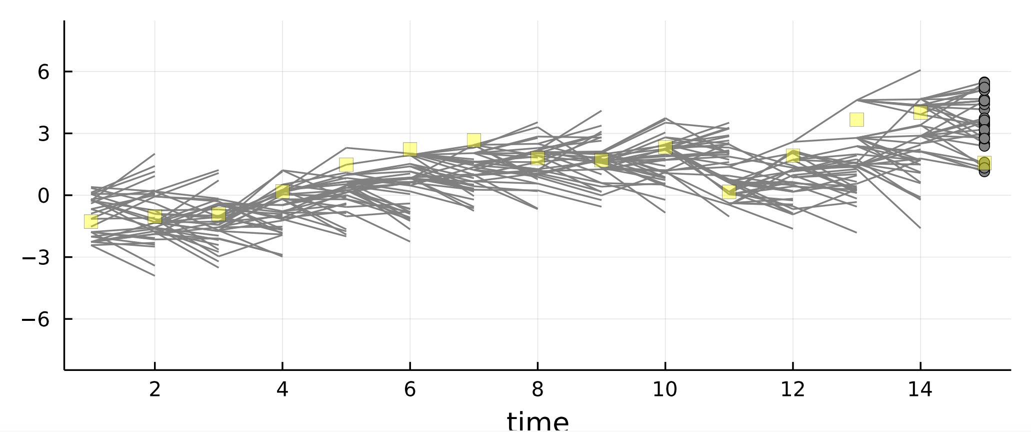

- Particles $X_t^i \sim M_t(X_{t-1}^{A_{t-1}^i}, \cdot\,)$
- Observations $y_{0:t}$

# Particle filter on noisy AR(1): Resample and propagate




- ● Particles $X_t^i \sim M_t(X_{t-1}^{A_{t-1}^i}, \cdot\,)$
- ■ Observations $y_{0:t}$

# Particle filter on noisy AR(1): $t = 15$




- ● Particles $X_{15}^{1:N}$
- ■ Observations $y_{0:15}$

# Forgetting

# Strong mixing condition

We assume the following which is common in (quantitative) particle filter theory:

**Assumption: Strong mixing**

There exist $0 < \underline{M} \leq \overline{M} < \infty$ and $0 < \underline{G} \leq \overline{G} < \infty$ such that $\forall \boldsymbol{x}, \boldsymbol{x}', t$:

- $\underline{M} \leq M_t(\boldsymbol{x}, \boldsymbol{x}') \leq \overline{M}$
- $\underline{G} \leq G_t(\boldsymbol{x}) \leq \overline{G}$

- Typically holds only in a compact state space
- (Variants exist which e.g. extend the requirement for iterates of $M_t$)

# Ideal filter forgetting

**Theorem (Del Moral 2004, Proposition 4.3.6)**

For all probability measures $\mu$ and $\nu$, $t \geq 0$ and $k \geq 0$:

$$\sup_{\mu,\nu} \|\Phi_{t,t+k}(\mu) - \Phi_{t,t+k}(\nu)\|_{\mathrm{TV}} \leq \beta^k \qquad \text{where} \qquad \beta = 1 - \left(\frac{\underline{\underline{M}}}{\overline{\overline{M}}}\right)^2$$

- Ideal filter forgets at exponential rate
- NB: $\Phi_t$ is non-linear and generally not contracting: we might well have

$$\|\Phi_t(\mu) - \Phi_t(\nu)\|_{\mathrm{TV}} > \|\mu - \nu\|_{\mathrm{TV}}$$

# Time-uniform $L^p$ errors

**Theorem (Del Moral 2004, Theorem 7.4.4)**

For all $N \geq 2$, $n \geq 0$ and $p \geq 1$, $\operatorname{osc}(\phi) \leq 1$:

$$\|\eta_t^N(\phi) - \eta_t(\phi)\|_p \leq \frac{c}{\sqrt{N}} \qquad \text{where} \qquad c_p = 2d_p^{1/p}\left(\frac{\overline{M}}{\underline{M}}\right)^3 \frac{\overline{G}}{\underline{G}}$$

where $d_p$ has been defined in Del Moral (2004), and in particular, $d_2 = 1$.

- $\eta_t^N(\phi) = N^{-1}\sum_{i=1}^N \phi(\boldsymbol{X}_t^i) \approx \eta_t(\phi) = \int \phi(\boldsymbol{x}_t)\eta_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t$ for large $N$ uniform in $t$ (in $L^p$ sense)

$\therefore$ Monte Carlo errors do not accumulate $\rightsquigarrow$ stability

# Forgetting of the particle filter?

In summary:

- Ideal filter is exponentially forgetting
- Particle filter is increasingly accurate approximation of the ideal filter...
- ...in a time-uniform manner

So the particle filter must also be exponentially forgetting, at least if $N$ is large enough?

# Particle filter as a Markov chain

- Unlike the ideal filter, particle filter defines a Markov chain $(\boldsymbol{X}_t^{1:N})_{t\geq 0}$
- Denote its Markov transition

$$\mathbf{M}_t(\boldsymbol{x}_{t-1}^{1:N}, \,\cdot\,) = \left( \sum_{i=1}^{N} \frac{G_{t-1}(\boldsymbol{x}_{t-1}^i)}{\sum_{j=1}^N G_{t-1}(\boldsymbol{x}_{t-1}^j)} M_t(\,\cdot\, \mid \boldsymbol{x}_{t-1}^i) \right)^{\otimes N}$$

- Is $\mathbf{M}_t$ contracting in Dobrushin sense:

$$\beta_{\mathrm{TV}}(\mathbf{M}_t) = \sup_{\boldsymbol{x}^{1:N}, \tilde{\boldsymbol{x}}^{1:N}} \|\mathbf{M}_t(\boldsymbol{x}^{1:N}, \,\cdot\,) - \mathbf{M}_t(\tilde{\boldsymbol{x}}^{1:N}, \,\cdot\,)\|_{\mathrm{TV}} < 1?$$

# An earlier forgetting result

Yes, $\mathbf{M}_t$ are contracting:

**Lemma (Tadić & Doucet, 2021)**

*For all $N \geq 1$ and $t \geq 0$:*

$$\beta_{\mathrm{TV}}(\mathbf{M}_t) \leq 1 - \epsilon^N, \qquad \textit{where} \qquad \epsilon = \left(\frac{\underline{M}}{\overline{\overline{M}}}\right)^2$$

Direct corollary of the above:

$$\beta_{\mathrm{TV}}\left(\mathbf{M}_{t,t+k}\right) \leq (1 - \epsilon^N)^k \qquad \text{where} \qquad \mathbf{M}_{t,t+k} = \mathbf{M}_{t+1}\mathbf{M}_{t+2}\cdots\mathbf{M}_{t+k}$$

$\rightsquigarrow$ forgetting in $k = O(e^N)$ time 🤔

# Forgetting result

# Particle filter forgetting

**Theorem (Karjalainen, Lee, Singh & V (2023))**

*For all $k \geq 1, t \geq 0, N \geq 2$,*

$$\beta_{\mathrm{TV}}(\mathbf{M}_{t,t+k}) \leq (1-\varepsilon)^{\lfloor k/(c \log N) \rfloor},$$

*where $\varepsilon \in (0,1)$ and $c < \infty$ only depend on the strong mixing constants.*

$\rightsquigarrow$ PF forgets in $k = O(\log N)$ time 😀

- Seems like the right order: a specific example where forgetting $\Omega(\log N)$...

# Proof sketch 1: Hellinger distance

## Definition

The squared Hellinger distance between two probability measures $P$ and $Q$ having densities $p$ and $q$ with respect to a common dominating measure $\lambda$ is

$$H^2(P,Q) = \frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \lambda(\mathrm{d}x) = 1 - \int \sqrt{p(x)q(x)}\lambda(\mathrm{d}x)$$

## Lemma (Tensorisation)

$$1 - H^2(P^{\otimes N}, Q^{\otimes N}) = \left(1 - H^2(P,Q)\right)^N$$

# Proof sketch 2: Total variation & random measures

**Lemma (Le Cam's inequality)**

*For any probability measures $\mu$ and $\nu$ it holds that*

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \sqrt{1 - (1 - H^2(\mu,\nu))^2}$$

Using tensorisation & Jensen's inequality:

**Corollary**

*For random product measures*

$$\|\mathbb{E}\mu^{\otimes N} - \mathbb{E}\nu^{\otimes N}\|_{\mathrm{TV}} \leq \sqrt{1 - \left(1 - \mathbb{E}H^2(\mu,\nu)\right)^{2N}}$$

# Proof sketch 3: Bounding expected Hellinger with $L^2$

**Lemma**

*Let $\mu$ and $\nu$ be two random probability measures, then*

$$\mathbb{E}H^2(\mu M, \nu M) \leq c' \sup_{\operatorname{osc}(\phi)\leq 1} \mathbb{E}(|\mu(\phi) - \nu(\phi)|^2) \qquad \textit{where} \qquad c' = \frac{1}{8}\left(\frac{\overline{M}}{\underline{M}}\right)^2$$

# Proof sketch 4: Putting things together

Let $\pi_t^N$ and $\tilde{\pi}_t^N$ be particle filters with same $M_t, G_t$ but with initial $\boldsymbol{X}_0^{1:N}$ and $\tilde{\boldsymbol{X}}_0^{1:N}$

$$\begin{aligned}\beta_{\mathrm{TV}}(\mathbf{M}_{0,k}) &= \inf_{\boldsymbol{X}_0^{1:N}, \tilde{\boldsymbol{X}}_0^{1:N}} \|\mathbb{E}(\pi_{k-1}^N M_k)^{\otimes N} - \mathbb{E}(\pi_{k-1}^N M_k)^{\otimes N}\|_{\mathrm{TV}} \\ &\leq \left(1 - \left(1 - c' \sup_{\mathrm{osc}(\phi)\leq 1} \|\pi_{k-1}^N(\phi) - \tilde{\pi}_{k-1}^N(\phi)\|_2^2\right)^{2N}\right)^{1/2}\end{aligned}$$

and

$$\begin{aligned}\|\pi_{k-1}^N(\phi) - \tilde{\pi}_{k-1}^N(\phi)\|_2 \leq & \|\pi_{k-1}^N(\phi) - \pi_{k-1}(\phi)\|_2 && \leq cN^{-1/2} \\ & + |\pi_{k-1}(\phi) - \tilde{\pi}_{k-1}(\phi)| && \leq \beta^{k-2} \leq cN^{-1/2} \text{ if } k \geq c\log N \\ & + \|\tilde{\pi}_{t-1}(\phi) - \tilde{\pi}_{t-1}^N(\phi)\|_2 && \leq cN^{-1/2}\end{aligned}$$

in which case

$$\beta_{\mathrm{TV}}(\mathbf{M}_{0,k}) \leq \left(1 - \left(1 - \frac{c}{N}\right)^{2N}\right)^{1/2} \approx \left(1 - e^{-2c}\right)^{1/2} \leq 1 - \epsilon \qquad \square$$

# Conditional particle filter

# Conditional particle filter algorithm

Andrieu, Doucet & Holenstein (*JRSS B*, 2010)

**CPF**$(\eta_0, (M_t)_{t\geq 1}, (G_t)_{t\geq 0}, (\boldsymbol{x}_t^*)_{t\geq 0}, N)$

1: $\boldsymbol{X}_t^0 = \boldsymbol{x}_t^*$, for $t \geq 0$
2: $\boldsymbol{X}_0^i \sim \eta_0(\,\cdot\,)$ for $i \in \{1{:}N\}$
3: **for** $t = 1, 2, \ldots$ **do**
4: $\quad W_{t-1}^i = \dfrac{G_{t-1}(\boldsymbol{X}_{t-1}^i)}{\sum_{j=1}^N G_{t-1}(\boldsymbol{X}_{t-1}^j)}$ for $i \in \{0{:}N\}$
5: $\quad A_{t-1}^i \sim \mathrm{Categorical}\big(W_{t-1}^{0:N}\big)$ for $i \in \{1{:}N\}$
6: $\quad \boldsymbol{X}_t^i \sim M_t(\,\cdot \mid \boldsymbol{X}_{t-1}^{A_{t-1}^i})$ for $i \in \{1{:}N\}$
7: **end for**

- Used for smoothing, but we think of CPF as a perturbed particle filter

⇝ Empirical approximations of $\eta_t$ and $\pi_t$:

$$\hat{\eta}_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{X}_t^i} \qquad \text{and} \qquad \hat{\pi}_t^N = \sum_{i=1}^N W_t^i \delta_{\boldsymbol{X}_t^i}$$

# Time-uniform $L^p$ errors for CPF

**Theorem (Karjalainen, Lee, Singh & V, 2023)**

*For every $p \geq 1$, there exists a constant $c = c(p)$ depending on strong mixing constants, such that for all $\mathrm{osc}(\phi) \leq 1$, $t \geq 0$, $N \geq 1$ and $x_t^*$,*

$$\left\| \hat{\pi}_t^N(\phi) - \pi_t(\phi) \right\|_p \leq \frac{c}{\sqrt{N}}$$

- Monte Carlo errors do not accumulate
- The effect of reference $x_t^*$ remains limited, too

# Forgetting of CPF

**Theorem**

*For all* $k \geq 1, t \geq 0, N \geq 2$, *and references* $x^* = (x_t^*)_{t \geq 0}$

$$\beta_{\mathrm{TV}}(\mathbf{M}_{t,t+k}^{x^*}) \leq (1 - \varepsilon)^{\lfloor k/(c \log N) \rfloor},$$

*where* $\varepsilon$ *and* $c$ *only depend on the strong mixing constants.*

- Similar proof as for PF, using the time-uniform $L^p$ error result

# Concluding remarks

# Maximal coupling of particle filters

- Coupled particle filters (and CPFs) have been used recently for multilevel Monte Carlo (and unbiased estimation)
- Let $\mu_k = \Phi_t(\eta_k^N)$ and $\tilde{\mu}_k = \Phi_t(\tilde{\eta}_k^N)$ stand for the conditional distributions of $X_t^{1:N}$ and $\tilde{X}_t^{1:N}$, respectively.
- Jasra & Yu (2020) suggested an 'independent maximal coupling' (IMC) algorithm:

$$(X_k^i, \tilde{X}_k^i) \sim \textsc{MaxCouple}(\mu_k, \tilde{\mu}_k), \quad \text{independently for } i = 1, \ldots, N$$

- Our analysis suggests another 'joint maximal coupling' (JMC) algorithm:

$$(X_k^{1:N}, \tilde{X}_k^{1:N}) \sim \textsc{MaxCouple}(\mu_k^{\otimes N}, \tilde{\mu}_k^{\otimes N})$$

- Both are implementable, but have $O(N^2)$ complexity

# Coupling probability when $N$ increases

- Suppose $X_t^{1:N}$ and $\tilde{X}_t^{1:N}$ follow same dynamics, but have $\eta_0 \neq \tilde{\eta}_0$
- Suppose that $k > c \log N$
- Our analysis says that with JMC, for all $N \geq 1$:

$$\mathbb{P}(X_t^{1:N} \neq \tilde{X}_t^{1:N}) = \mathbb{E}\|\mu_k - \tilde{\mu}_k\|_{\mathrm{TV}} \leq \left(1 - \left(1 - \frac{c}{N}\right)^{2N}\right)^{1/2} \leq \epsilon$$

$\rightsquigarrow$ filters fully coupled after $O(\log N)$ iterations.

- In contrast, with IMC, we can only guarantee:

$$\mathbb{P}(X_t^{1:N} = \tilde{X}_t^{1:N}) = \mathbb{E}\left[(1 - \|\mu_k - \tilde{\mu}_k\|_{\mathrm{TV}})^N\right] \geq \left(1 - \frac{c}{\sqrt{N}}\right)^N,$$

which vanishes as $N \to \infty$

# Discussion

- Dobrushin forgetting is a strong form of stability
  - Complementary to $L^p$ uniform stability, which is about averages
  - Could prove useful in some PF/CPF analysis
- Forgetting entails $\log N$ ‘penalty’ term
  - Seems unavoidable
  - Drowned by $N^{-p}$ where $p > 0$
- Implications to CPF out soon...

# References


- **C. Andrieu, A. Doucet and R. Holenstein.**
  Particle Markov chain Monte Carlo methods.
  *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 2010.
- **P. Del Moral.**
  *Feynman-Kac formulae: Genealogical and Interacting Particle Systems with Applications*, volume 88.
  Springer, 2004.
- **N. J. Gordon, D. J. Salmond, and A. F. M. Smith.**
  Novel approach to nonlinear/non-Gaussian Bayesian state estimation.
  *IEE Proceedings-F*, 140(2):107–113, 1993.
- **A. Jasra and F. Yu.**
  Central limit theorems for coupled particle filters.
  *Adv. in Appl. Probab.*, 52:942–1001, 2020.
- **V. Z. Tadić and A. Doucet.**
  Asymptotic properties of recursive particle maximum likelihood estimation.
  *IEEE Trans. Inform. Theory*, 67(3):1825–1848, 2021.
- **J. Karjalainen, A. Lee, S. S. Singh and M. Vihola.**
  On the forgetting of particle filters
  *arXiv:2309.08517*, 2023.


Thank you!