Divide-and-Conquer Sequential Monte Carlo

Adam M. Johansen

Joint work with:
John Aston, Alexandre Bouchard-Côté, Ryan Chan, Francesca Crucinio, Brent Kirkpatrick, Juan Kuntz, Fredrik Lindsten, Christian Næsseth, Murray Pollock, Gareth Roberts and Thomas Schön

University of Warwick
a.m.johansen@warwick.ac.uk
http://go.warwick.ac.uk/amjohansen/talks/

CoSinES / Bayes4Health Masterclass on SMC
April 1st, 2022

Outline

- Sequential Importance Sampling / Sequential Monte Carlo (SMC)
- SMC to Divide and Conquer SMC (D&C-SMC; Lindsten et al. (2017))
- Some Theoretical Properties of D&C-SMC (Kuntz et al., 2021)
- Illustrative Application: Hierarchical Fusion (Chan et al., 2021)
- Conclusions and Some (Open) Questions

Part I

Sequential Monte Carlo and Divide-and-Conquer Implementations

See Lindsten et al. (2017)

Essential Problem

The Abstract Problem

- Given a density,

$$\mu(x) = \frac{\rho(x)}{Z},$$

- such that $\rho(x)$ can be evaluated pointwise,
- how can we approximate μ
- and how about Z?
- Can we do so robustly?
- In a distributed setting?

Sequential Importance Resampling

Ingredients:

- Sequence of unnormalized (path space) targets ρ_t on $\mathbf{E}_t = \otimes_{s=0}^{t} E_s$.
- Normalizing constants $Z_t = \rho_t(\mathbf{E}_t)$
- Normalized counterparts $\mu_t = \rho_t / Z_t$.
- Proposals K_t: conditional laws over E_t given $\mathbf{x}_{t-1} \in \mathbf{E}_{t-1}$.
- Importance weights / potential functions:

$$w_t = \frac{d\rho_t}{d\rho_{t-1} \otimes K_t}.$$

Algorithm: iterative importance sampling and resampling.

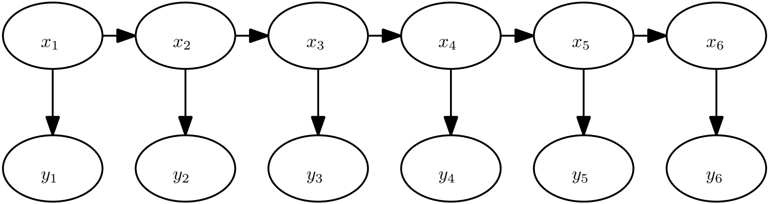
Sequential Importance Resampling

1: *Propose:* for $n \leq N$, draw $\mathbf{X}_0^{n,N}$ independently from K_0.
2: *Correct:* compute $\rho_0^N := N^{-1} \sum_{n=1}^N w_0(\mathbf{X}_0^{n,N})\delta_{\mathbf{X}_0^n}$, where $w_0 := d\rho_0/dK_0$, $Z_0^N = \rho_0^N(\mathbf{E}_0)$ and $\mu_0^N := \rho_0^N/Z_0^N$.
3: **for** $t = 1, \ldots, T$ **do**
4: *Resample:* for $n \leq N$, draw $\mathbf{X}_{t-}^{n,N}$ independently from μ_{t-1}^N.
5: *Mutate:* for $n \leq N$, draw $X_t^{n,N}$ independently from $K_t(\mathbf{X}_{t-}^{n,N}, dx_t)$ and set $\mathbf{X}_t^{n,N} := (X_t^{n,N}, \mathbf{X}_{t-}^{n,N})$.
6: *Correct:* compute

$$\rho_t^N = \frac{Z_{t-1}^N}{N} \sum_{i=1}^N w_t(\mathbf{X}_t^{n,N})\delta_{\mathbf{X}_t^{n,N}},$$

$Z_t^N = \rho_t^N(\mathbf{E}_t)$ and $\mu_t^N := \rho_t^N/Z_t^N$.
7: **end for**

SIR Example: Simple Particle Filters



- Unobserved Markov chain $\{X_n\}$ transition f.
- Observed process $\{Y_n\}$ conditional density g.
- The joint density is available:

$$p(x_{1:n}, y_{1:n}|\theta) = f_1^\theta(x_1) g^\theta(y_1|x_1) \prod_{i=2}^{n} f^\theta(x_i|x_{i-1}) g^\theta(y_i|x_i).$$

- Natural SIR target distributions:

$$\mu_n^\theta(x_{1:n}) := p(x_{1:n}|y_{1:n}, \theta) \propto p(x_{1:n}, y_{1:n}|\theta) =: \rho_n^\theta(x_{1:n})$$

$$Z_n^\theta = \int p(x_{1:n}, y_{1:n}|\theta) dx_{1:n} = p(y_{1:n}|\theta)$$

Bootstrap PFs and Similar

- Choosing

$$\mu_n^\theta(x_{1:n}) := p(x_{1:n}|y_{1:n},\theta) \propto p(x_{1:n},y_{1:n}|\theta) =: \rho_n^\theta(x_{1:n})$$
$$Z_n^\theta = \int p(x_{1:n},y_{1:n}|\theta)dx_{1:n} = p(y_{1:n}|\theta)$$

- and $K_p(x_p|x_{1:p-1}) = f^\theta(x_p|x_{p-1})$ yields the bootstrap particle filter of Gordon et al. (1993),
- whereas $K_p(x_p|x_{1:p-1}) = p(x_p|x_{p-1},y_p,\theta)$ yields the "locally optimal" particle filter.
- Note: Many alternative particle filters are SIR algorithms with other targets. Cf. J. and Doucet (2008); Doucet and J. (2011).

Sequential Monte Carlo Samplers: Another SIR Class

Given a sequence of targets $\bar{\mu}_1, \ldots, \bar{\mu}_n$ on *arbitrary* spaces, Del Moral et al. (2006) extend the space:

$$\mu_n(x_{1:n}) = \bar{\mu}_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p)$$

$$\rho_n(x_{1:n}) = \bar{\rho}_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p)$$

$$Z_n = \int \rho_n(x_{1:n}) dx_{1:n}$$

$$= \int \bar{\rho}_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p) dx_{1:n} = \int \bar{\rho}_n(x_n) dx_n = \bar{Z}_n$$

SIR: Theoretical Justification — Some Of

Under regularity conditions we have:

unbiasedness

$$\mathbb{E}[\widehat{Z}_n^N] = Z_n$$

slln

$$\lim_{N\to\infty} \widehat{\pi}_n^N(\varphi) \stackrel{\text{a.s.}}{=} \pi_n(\varphi)$$

clt For a normal random variable W_n of appropriate variance:

$$\lim_{N\to\infty} \sqrt{N}[\widehat{\pi}_n^N(\varphi) - \pi_n(\varphi)] \stackrel{\text{d}}{=} W_n$$

although establishing this requires a little work (cf., e.g. Del Moral (2004).

Auxiliary sequential importance resampling

Ingredients:

- Sequence of unnormalized (path space) targets ρ_t on $\mathbf{E}_t = \otimes_{s=0}^{t} E_s$.
- Sequences of auxiliary targets $\gamma_{t\_}$ and $\gamma_t := \gamma_{t\_} \otimes K_t$.
- Normalizing constants $Z_t = \rho_t(\mathbf{E}_t)$
- Auxiliary normalizing constants $\mathcal{Z}_t = \gamma_t(\mathbf{E}_t)$
- Normalized counterparts $\mu_t = \rho_t / Z_t$.
- Normalized auxiliary targets $\pi_t = \gamma_t / \mathcal{Z}_t$.
- Proposal kernels K_t: conditional laws over E_t given $\mathbf{E}_{t-1}$.
- Importance weights / potential functions:

$$w_t = \frac{d\gamma_{t\_}}{d\gamma_{t-1}}.$$

Algorithm: iterative importance sampling and resampling targeting auxiliary targets and an extra importance sampling correction.

Auxiliary sequential importance resampling

1: *Propose:* for $n \leq N$, draw $\mathbf{X}_0^{n,N}$ independently from K_0.
2: *Compute:* $\gamma_0^N := N^{-1} \sum_{n=1}^N \delta_{\mathbf{X}_0^{n,N}}$.
3: **for** $t = 1, \ldots, T$ **do**
4: *Correct:* compute $\gamma_{t_-}^N(d\mathbf{x}_{t-1}) := w_{t_-}(\mathbf{x}_{t-1})\gamma_{t-1}^N(d\mathbf{x}_{t-1})$ and $\pi_{t_-}^N := \gamma_{t_-}^N / \gamma_{t_-}^N(\mathbf{E}_{t-1})$.
5: *Resample:* for $n \leq N$, draw $\mathbf{X}_{t_-}^{n,N}$ independently from $\pi_{t_-}^N$.
6: *Mutate:* for $n \leq N$, draw $X_t^{n,N}$ independently from $K_t(\mathbf{X}_{t_-}^{n,N}, dx_t)$ and set $\mathbf{X}_t^{n,N} := (X_t^{n,N}, \mathbf{X}_{t_-}^{n,N})$.
7: *Compute:* $\gamma_t^N := \frac{\mathcal{Z}_t^N}{N} \sum_{n=1}^N \delta_{\mathbf{X}_t^{n,N}}$ where $\mathcal{Z}_t^N := \gamma_{t_-}^N(\mathbf{E}_{t-1})$.
8: **end for**

Auxiliary Particle Filters

In the filtering setting, take:

- $\gamma_{t_-}(d\mathbf{x}_{t-1}) = p(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\hat{p}(y_t|x_{t-1})$
- $\pi_{t_-} = \gamma_{t_-}/\gamma_{t_-}(\mathbf{E}_{t-1})$.

and one recovers the auxiliary particle filter of Pitt and Shephard (1999).

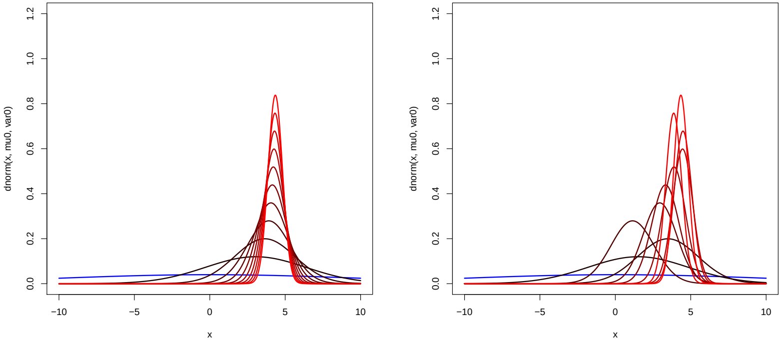
Bayesian Inference via SMC (Chopin, 2001;Del Moral et al., 2006)

In a Bayesian context:

- Given a prior $p(\theta)$ and likelihood $l(\theta; y_{1:m})$
- One could specify:

  Data Tempering $\bar{\rho}_p(\theta) = p(\theta) l(\theta; y_{1:m_p})$ for
  $$m_1 = 0 < m_2 < \cdots < m_T = m$$
  Likelihood Tempering $\bar{\rho}_p(\theta) = p(\theta) l(\theta; y_{1:m})^{\beta_p}$ for
  $$\beta_1 = 0 < \beta_2 < \cdots < \beta_T = 1$$
  Something else?
- Here $Z_T = \int p(\theta) l(\theta; y_{1:n}) d\theta$ and $\bar{\rho}_T(\theta) \propto p(\theta | y_{1:n})$.
- Specifying $(m_1, \ldots, m_T)$, $(\beta_1, \ldots, \beta_T)$ or $(\gamma_1, \ldots, \gamma_T)$ is hard.

Illustrative Sequences of Targets



One Adaptive Scheme (Zhou, J. & Aston, 2016)+Refs

Resample When $\mathrm{ESS}(W_n^{1:N}) = \left(\sum_{i=1}^N (W_n^i)^2\right)^{-1}$ is below a threshold.

Likelihood Tempering At iteration n: Set β_n such that:

$$\frac{N(\sum_{j=1}^N W_{n-1}^{(j)} w_n^{(j)})^2}{\sum_{k=1}^N W_{n-1}^{(k)} (w_n^{(k)})^2} = \mathrm{CESS}_\star$$

which controls χ^2-discrepancy between successive distributions.

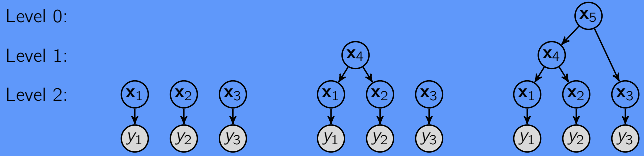
Proposals Follow (Jasra et al., 2010): adapt to keep acceptance rate about right.

Question

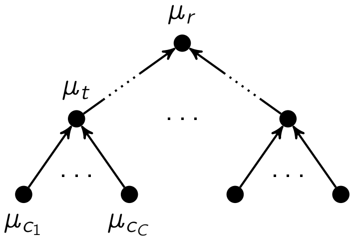
Are there better, practical approaches to specifying a sequence of distributions?

Divide-and-Conquer (Lindsten et al., 2017)

Many models admit natural decompositions:

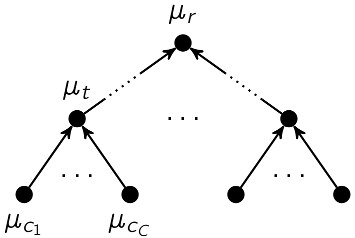


To which we can apply a divide-and-conquer strategy:



A few formalities...

- Use a tree, $\mathbb{T}$ of models (with rootward variable inclusion):



- Let $t \in \mathbb{T}$ denote a node; $r \in \mathbb{T}$ is the root.
- Let $\mathcal{C}_t = \{c_1, \ldots, c_C\}$ denote the children of t.
- Let E_t denote the space of variables included in t but *not* its children.
- Let $\mathbf{E}_t = E_t \times \bigotimes_{c \in C(t)} \mathbf{E}_c$ be the space of all variables included in $\mathbb{T}_t$: the subtree rooted at t.
- dc-smc can be viewed as a recursion over this tree.
- NB. The tree of models can be constructed even for models which are not tree-like.

dac_smc(u) for u in $\mathbb{T}$.

if u is a leaf (i.e. $u \in \mathbb{T}^{\partial}$) **then**
    Propose: for $n \leq N$, draw $\mathbf{X}_u^{n,N}$ independently from K_u.
    Return: $\gamma_u^N := N^{-1} \sum_{n=1}^{N} \delta_{\mathbf{X}_u^{n,N}}$.
else
    for v in $\mathcal{C}_u$ **do**
        Recurse: set $\gamma_v^N :=$ dac_smc(v).
    end for
    Correct: compute $\gamma_{u_-}^N$ and $\pi_{u_-}^N := \gamma_{u_-}^N / \gamma_{u_-}^N(\mathbf{E}_{\mathcal{C}_u})$.
    Resample: for $n \leq N$, draw $\mathbf{X}_{u_-}^{n,N}$ independently from $\pi_{u_-}^N$.
    Mutate: for $n \leq N$, draw $X_u^{n,N}$ independently from $K_u(\mathbf{X}_{u_-}^{n,N}, dx_u)$ and set $\mathbf{X}_u^{n,N} := (X_u^{n,N}, \mathbf{X}_{u_-}^{n,N})$.
    Return: $\gamma_u^N := N^{-1} \mathcal{Z}_u^N \sum_{n=1}^{N} \delta_{\mathbf{X}_u^{n,N}}$ where $\mathcal{Z}_u^N := \gamma_{u_-}^N(\mathbf{E}_{\mathcal{C}_u})$.
end if

Part II

Theoretical Properties
See Kuntz et al. (2021)

Theoretical Properties: Regularity Assumptions I

Assumption (1. Absolute Continuity)

For all u in $\mathbb{T}$ and v in $\mathbb{T}^{\partial}$, ρ_u is absolutely continuous w.r.t. γ_u, γ_{v_-} is absolutely continuous w.r.t. $\gamma_{\mathcal{C}_v}$, and the Radon-Nikodym derivatives $w_u := d\rho_u/d\gamma_u$ and $w_{v_-} := d\gamma_{v_-}/d\gamma_{\mathcal{C}_v}$ are positive everywhere.

Assumption (2. Boundedness)

For all u in $\mathbb{T}^{\partial}$ and v in $\mathbb{T}$, $w_{u_-} = d\gamma_{u_-}/d\gamma_{\mathcal{C}_u}$ and $w_v = d\rho_v/d\gamma_v$ are bounded: $||w_{u_-}||_\infty < \infty$ and $||w_v||_\infty < \infty$.

Theoretical Properties I

L_p Error Bounds (Kuntz et al., 2021, Theorem 5)

If Assumptions 1–2 hold, then, for each $p \geq 1$ and u in $\mathbb{T}$, then there exist constants $C_u^\rho, C_u^\mu < \infty$ such that

$$\mathbb{E}\left|\rho_u^N(\varphi) - \rho(\varphi)\right|^{p\,\frac{1}{p}} \leq \frac{C_u^\rho ||\varphi||_\infty}{N^{1/2}},$$

$$\mathbb{E}\left|\mu_u^N(\varphi) - \mu_u(\varphi)\right|^{p\,\frac{1}{p}} \leq \frac{C_u^\mu ||\varphi||_\infty}{N^{1/2}},$$

for all $N > 0$ and φ in $\mathcal{B}_b(\mathbf{E}_u)$. In particular,

$$\mathbb{E}[|Z_u^N - Z_u|^p]^{1/p} \leq C_u^\rho / N^{1/2}$$

for all $N > 0$.

Theoretical Properties II

Strong Law of Large Numbers (Kuntz et al., 2021, Theorem 1)

If Assumptions 1–2 are satisfied, u belongs to $\mathbb{T}$, and φ belongs to $\mathcal{B}_b(\mathbf{E}_u)$, then

$$\lim_{N\to\infty} \rho_u^N(\varphi) = \rho_u(\varphi), \quad \lim_{N\to\infty} \mu_u^N(\varphi) = \mu(\varphi), \quad \lim_{N\to\infty} Z_u^N = Z_u,$$

almost surely.

Strong Law of Large Numbers (Kuntz et al., 2021, Theorem 2)

If, in addition to Assumptions 1–2, the spaces $(E_u)_{u\in\mathbb{T}}$ are Polish and $(\mathcal{E}_u)_{u\in\mathbb{T}}$ are the corresponding Borel sigma algebras, then

$$\rho_u^N \rightharpoonup \rho_u, \quad \mu_u^N \rightharpoonup \mu_u, \quad \text{almost surely},$$

for each u in $\mathbb{T}$, where $\rightharpoonup$ denotes weak convergence as $N \to \infty$.

Theoretical Properties III

Central Limit theorem (Kuntz et al., 2021, Theorem 6)

If Assumptions 1–2 hold, then, as $N \to \infty$,

$$N^{1/2}\left(\rho_u^N(\varphi) - \rho_u(\varphi)\right) \Rightarrow \mathcal{N}(0, \sigma^2_{\rho_u}(\varphi)),$$
$$N^{1/2}\left(\mu_u^N(\varphi) - \mu_u(\varphi)\right) \Rightarrow \mathcal{N}(0, \sigma^2_{\mu_u}(\varphi)),$$

for any given u in $\mathbb{T}$ and φ in $\mathcal{B}_b(\mathbf{E}_u)$, where $\Rightarrow$ denotes convergence in distribution,

$$\sigma^2_{\rho_u}(\varphi) := \sum_{v \in \mathbb{T}_u} \pi_v([\mathcal{Z}_v \Gamma_{v,u}[w_u \varphi] - \rho_u(\varphi)]^2),$$
$$\sigma^2_{\mu_u}(\varphi) := \sum_{v \in \mathbb{T}_u} \pi_v([\mathcal{Z}_v \Gamma_{v,u}[w_u Z_u^{-1}[\varphi - \mu_u(\varphi)]]]^2).$$

Theoretical Properties IV

More on the CLT

In particular, $N^{1/2}\left(Z_u^N - Z_u\right) \Rightarrow \mathcal{N}(0, \sigma^2_{Z_u})$ as $N \to \infty$ with

$$\sigma^2_{Z_u} := Z_u^2 \sum_{v \in \mathbb{T}_u} \pi_v \left(\left[\frac{d\mu_u^v}{d\pi_v} - 1 \right]^2 \right), \tag{1}$$

where μ_u^v denotes the $\mathbf{E}_v$-marginal of μ_u (i.e. $\mu_u^v(A) := \mu_u(A \times E_{\mathbb{T}_u \setminus \mathbb{T}_v})$ for all A in $\mathcal{E}_v$).

Unbiasedness of NC Estimates (Kuntz et al., 2021, Theorem 3)

If Assumptions 1–2 hold, then for all $u \in \mathbb{T}$:

$$\mathbb{E}\left[\rho_u^N(\varphi)\right] = \rho_u(\varphi), \quad \mathbb{E}\left[Z_u^N\right] = Z_u, \quad \forall N > 0, \ \ \varphi \in \mathcal{B}_b(\mathbf{E}_u).$$

Theoretical Properties V

One Key Ingredient: Multinomial Expansion

Fix any u in $\mathbb{T}^{\partial}$ and φ in $\mathcal{B}_b(\mathbf{E}_{\mathcal{C}_u})$. Note that,

$$\gamma_{\mathcal{C}_u}^N - \gamma_{\mathcal{C}_u} = \prod_{v \in \mathcal{C}_u} [\gamma_v^N - \gamma_v + \gamma_v] - \gamma_{\mathcal{C}_u} = \sum_{\emptyset \neq A \subseteq \mathcal{C}_u} \Delta_A^N \times \gamma_{\mathcal{C}_u}^{\not A}, \quad (2)$$

where $\Delta_A^N := \prod_{v \in A}(\gamma_v^N - \gamma_v)$ and $\gamma_{\mathcal{C}_u}^{\not A} := \gamma_{\mathcal{C}_u \setminus A}$ for all subsets A of $\mathcal{C}_u$.

Some (Importance) Extensions

1. (Lightweight) Mixture Resampling [with Rejection Sampling]
2. Tempering (Del Moral et al, 2006)
3. Adaptation (Zhou, J. and Aston, 2016)

Part III

Illustrative Application: Hierarchical Monte Carlo Fusion
See Chan et al. (2021)

Hierarchical Monte Carlo Fusion (Chan et al., 2001) I

Objective: combine approximations of "subposteriors":

$$f(\mathbf{x}) \propto \prod_{c \in \mathcal{C}} f_c(\mathbf{x}), \tag{3}$$

Proposition (Dai et al. (2019))

Suppose that p_c is the transition density of a Markov chain on $\mathbb{R}^d$ with a stationary probability density proportional to f_c^2. Then the $(|\mathcal{C}|+1)d$-dimensional probability density proportional to the integrable function

$$g_{\mathcal{C}}\left(\vec{\mathbf{x}}^{(\mathcal{C})}, \mathbf{y}^{(\mathcal{C})}\right) := \prod_{c \in \mathcal{C}} \left[f_c^2\left(\mathbf{x}^{(c)}\right) \cdot p_c\left(\mathbf{y}^{(\mathcal{C})} \middle| \mathbf{x}^{(c)}\right) \cdot \frac{1}{f_c(\mathbf{y}^{(\mathcal{C})})} \right], \tag{4}$$

admits marginal density $f^{(\mathcal{C})} \propto \prod_{c \in \mathcal{C}} f_c$ over $\mathbf{y}^{(\mathcal{C})} \in \mathbb{R}^d$.

Hierarchical Monte Carlo Fusion (Chan et al., 2001) II

This can be exploited by taking a proposal distribution proportional to:

$$h_{\mathcal{C}}\left(\vec{\mathbf{x}}^{(\mathcal{C})}, \mathbf{y}^{(\mathcal{C})}\right) := \prod_{c \in \mathcal{C}} f_c\left(\mathbf{x}^{(c)}\right) \cdot \exp\left\{-\frac{(\mathbf{y}^{(\mathcal{C})} - \tilde{\mathbf{x}}^{(\mathcal{C})})^{\intercal} \boldsymbol{\Lambda}_{\mathcal{C}}^{-1} (\mathbf{y}^{(\mathcal{C})} - \tilde{\mathbf{x}}^{(\mathcal{C})})}{2T}\right\}$$

where

$$\tilde{\mathbf{x}}^{(\mathcal{C})} := \left(\sum_{c \in \mathcal{C}} \boldsymbol{\Lambda}_c^{-1}\right)^{-1} \left(\sum_{c \in \mathcal{C}} \boldsymbol{\Lambda}_c^{-1} \mathbf{x}^{(c)}\right), \qquad \boldsymbol{\Lambda}_{\mathcal{C}}^{-1} := \sum_{c \in \mathcal{C}} \boldsymbol{\Lambda}_c^{-1}.$$

Hierarchical Monte Carlo Fusion (Chan et al., 2001) III

Proposition

If $p_c\left(\mathbf{y}^{(\mathcal{C})}|\mathbf{x}^{(c)}\right)$ is the transition density of a suitable Langevin diffusion

$$\frac{g_{\mathcal{C}}(\vec{\mathbf{x}}^{(\mathcal{C})},\mathbf{y}^{(\mathcal{C})})}{h_{\mathcal{C}}(\vec{\mathbf{x}}^{(\mathcal{C})},\mathbf{y}^{(\mathcal{C})})} \propto \rho_0(\vec{\mathbf{x}}^{(\mathcal{C})}) \cdot \rho_1(\vec{\mathbf{x}}^{(\mathcal{C})},\mathbf{y}^{(\mathcal{C})}),$$

$$\rho_0(\vec{\mathbf{x}}^{(\mathcal{C})}) := \exp\left\{-\sum_{c\in\mathcal{C}} \frac{(\tilde{\mathbf{x}}^{(\mathcal{C})}-\mathbf{x}^{(c)})^\intercal \boldsymbol{\Lambda}_c^{-1}(\tilde{\mathbf{x}}^{(\mathcal{C})}-\mathbf{x}^{(c)})}{2T}\right\},$$

$$\rho_1(\vec{\mathbf{x}}^{(\mathcal{C})},\mathbf{y}^{(\mathcal{C})}) := \prod_{c\in\mathcal{C}} \mathbb{E}_{\mathbb{W}_{\Lambda_c}}\left[\exp\left\{-\int_0^T \phi_c\left(\mathbf{X}_t^{(c)}\right) dt\right\}\right],$$

$$\phi_c(\mathbf{x}) := \frac{1}{2}\left(\nabla \log f_c(\mathbf{x})^\intercal \boldsymbol{\Lambda}_c \nabla \log f_c(\mathbf{x}) + Tr(\boldsymbol{\Lambda}_c \nabla^2 \log f_c(\mathbf{x}))\right),$$

where $Tr(\cdot)$ denotes the trace of a matrix, and $\mathbb{W}_{\Lambda_c}$ denotes the law of a Brownian bridge $\{\mathbf{X}_t^{(c)}, t\in[0,T]\}$ with $\mathbf{X}_0^{(c)} := \mathbf{x}^{(c)}$, $\mathbf{X}_T^{(c)} := \mathbf{y}^{(\mathcal{C})}$ and covariance matrix $\boldsymbol{\Lambda}_c$.

D&C Fusion I

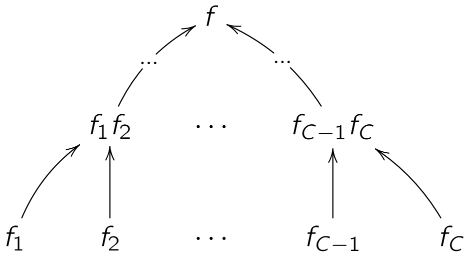
`general.fusion`$(\mathcal{C}, \{\{\mathbf{x}_{0,i}^{(c)}, w_i^{(c)}\}_{i=1}^{M}, \boldsymbol{\Lambda}_c\}_{c\in\mathcal{C}}, N, T)$

Input: Samples $\{\mathbf{x}_{0,i}^{(c)}, w_i^{(c)}\}_{i=1}^{M}$ for $c \in \mathcal{C}$, matrices, $\{\boldsymbol{\Lambda}_c : c \in \mathcal{C}\}$, particle count, N, and time horizon, $T > 0$.

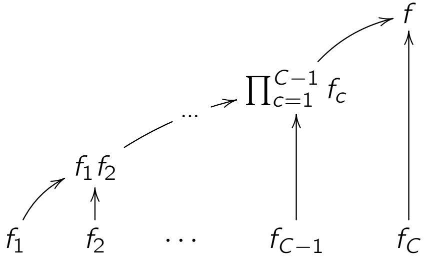
1. **Partial proposal:** Compose samples $\{\vec{\mathbf{x}}_{0,j}^{(\mathcal{C})}, \vec{w}_j\}_{j=1}^{M}$ where $\vec{w}_j := \left(\prod_{c\in\mathcal{C}} w_j^{(c)}\right) \cdot \rho_0(\vec{\mathbf{x}}_{0,j}^{(\mathcal{C})})$ for $j \in \{1, \ldots, M\}$.
2. For i in 1 to N,
   - 2.1 $\vec{\mathbf{x}}_{0,i}^{(\mathcal{C})}$**:** Sample $I \sim \text{categorical}(\vec{w}_{1:M})$ and set $\vec{\mathbf{x}}_{0,i}^{(\mathcal{C})} := \vec{\mathbf{x}}_{0,I}^{(\mathcal{C})}$.
   - 2.2 **Complete proposal:** Simulate $\mathbf{y}_i^{(\mathcal{C})} \sim \mathcal{N}_d\left(\tilde{\mathbf{x}}_i^{(\mathcal{C})}, T\boldsymbol{\Lambda}_{\mathcal{C}}\right)$.
   - 2.3 $\tilde{\rho}_{1,i}^{(\mathcal{C})}$**:** Compute importance weight $\tilde{\rho}_{1,i}^{(\mathcal{C})} := \tilde{\rho}_1^{(b)}\left(\vec{\mathbf{x}}_{0,i}^{(\mathcal{C})}, \mathbf{y}_i^{(\mathcal{C})}\right)$.
3. For i in 1 to N compute $w_i^{(\mathcal{C})} = \tilde{\rho}_{1,i}^{(\mathcal{C})} / \sum_{k=1}^{N} \tilde{\rho}_{1,k}^{(\mathcal{C})}$.

Output: $\left\{\vec{\mathbf{x}}_{0,i}^{(\mathcal{C})}, \mathbf{y}_i^{(\mathcal{C})}, w_i^{(\mathcal{C})}\right\}_{i=1}^{N}$.

D&C Fusion II



A balanced-binary tree.



A progressive tree.

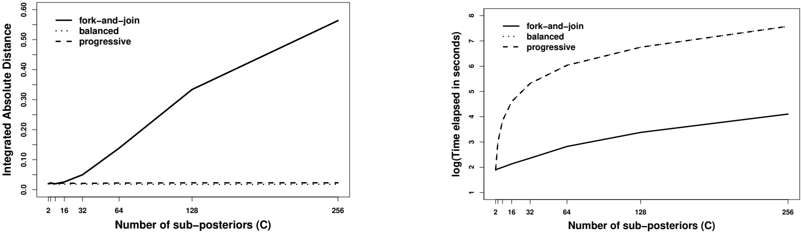
D&C Fusion III

`d&c.fusion`(v, N, T)

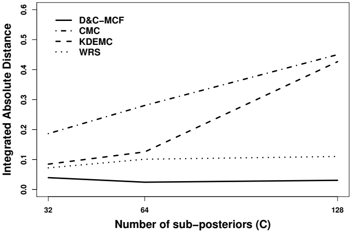
Given: Sub-posteriors, $\{f_u\}_{u \in \mathrm{Leaf}(\mathbb{T})}$, and preconditioning matrices $\{\mathbf{\Lambda}_u\}_{u \in \mathbb{T}}$.

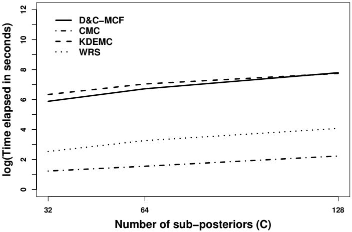
Input: Node in tree, v, the number of particles N, and time horizon $T > 0$.

1. For $u \in \mathrm{Ch}(v)$,
   1.1 $\left\{\mathbf{x}_i^{(u)}, \mathbf{y}_i^{(u)}, w_i^{(u)}\right\}_{i=1}^{N} \leftarrow$ `d&c.fusion`(u, N, T).
2. If $v \in \mathrm{Leaf}(\mathbb{T})$,
   2.1 For $i = 1, \dots, N$, sample $\mathbf{y}_i^{(v)} \sim f_v(\mathbf{y})$.
   2.2 **Output:** $\{\emptyset, \mathbf{y}_i^{(v)}, \frac{1}{N}\}_{i=1}^{N}$.
3. If $v \notin \mathrm{Leaf}(\mathbb{T})$,
   3.1 **Output:** Call `general.fusion`$(\mathrm{Ch}(v), \{\{\mathbf{y}_i^{(u)}, w_i^{(u)}\}_{i=1}^{N}, \mathbf{\Lambda}_u\}_{u \in \mathrm{Ch}(v)}, N, T)$.



Illustrative comparison of the effect of using different hierarchies, with $f \propto \prod_{c=1}^{C} f_c$, where $f_c \sim \mathcal{N}(0, C)$ for $c = 1, \ldots, C$ (averaged over 50 runs).





Comparison of methods [CMC=Consensus Monte Carlo; KDEMC=kernel density averaging approach of Neiswanger et al. (2014); WRS=Weierstrass Rejection Sampler] applied to a logistic regression problem with credit card data*.

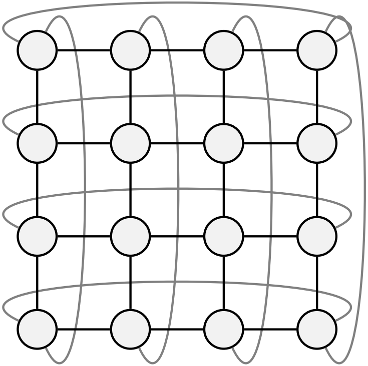
* The *'Default of credit card clients'* data set available from https://archive.ics.uci.edu/ml/datasets. The data set comprised $m = 30000$ records of **response:** whether a default had occurred and binary covariates **Gender** and **Education**.

Part IIIb

Some Other Examples

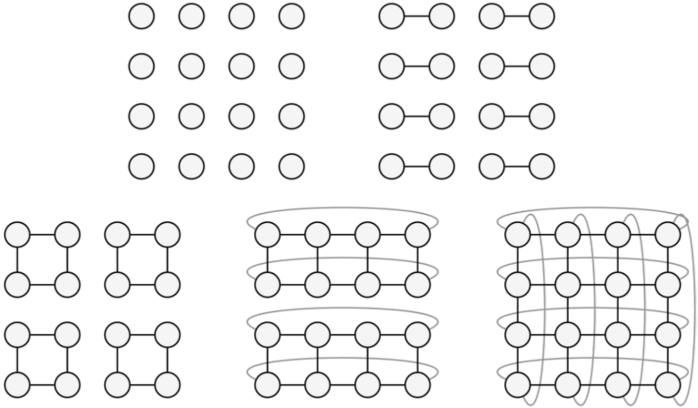
An Ising Model

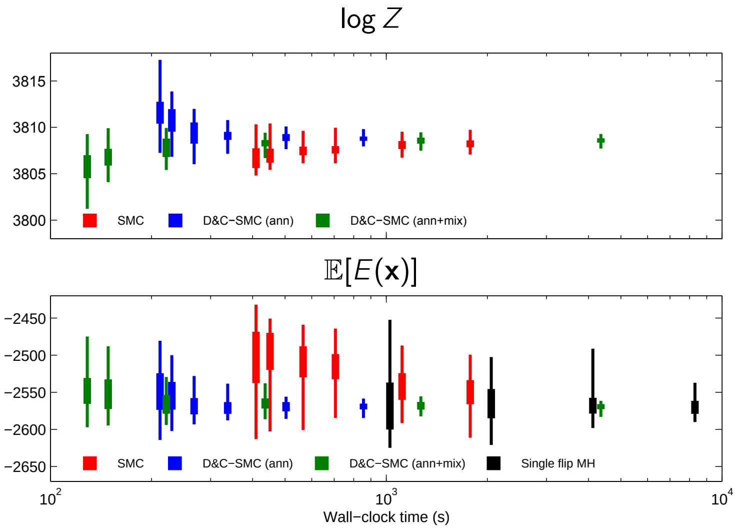
- k indexes $\mathcal{V} \in \mathcal{V} \subset \mathbb{Z}^2$
- $x_k \in \{-1, 1\}$
- $p(\mathbf{z}) \propto e^{-\beta E(\mathbf{z})}$, $\beta \geq 0$
- $E(\mathbf{z}) = -\sum_{(k,l)\in\mathcal{E}} x_k x_l$



We consider a grid of size 64×64 with $\beta = 0.4407$ (the critical temperature).

A sequence of decompositions





Summaries over 50 independent runs of each algorithm.

New York Schools Maths Test: data

- Data organised into a tree T.
- A root-to-leaf path is: NYC (the root, denoted by $r \in T$), borough, school district, school, year.
- Each leaf $t \in T$ comes with an observation of m_t exam successes out of M_t trials.
- Total of 278 399 test instances
- five borough (Manhattan, The Bronx, Brooklyn, Queens, Staten Island),
- 32 distinct districts,
- 710 distinct schools.

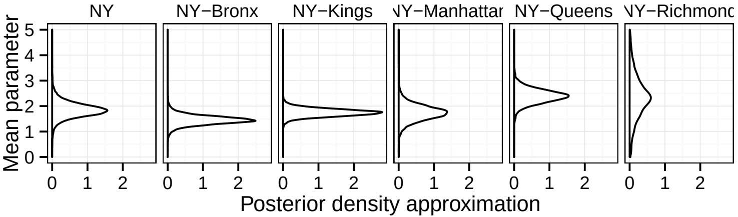
New York Schools Maths Test: Bayesian Model

- Number of successes m_t at a leaf t is $\mathrm{Bin}(M_t, p_t)$.
- where $p_t = \mathrm{logistic}(\theta_t)$, where θ_t is a latent parameter.
- internal nodes of the tree also have a latent θ_t
- model the difference in θ_t along $e = (t \to t')$ as $\theta_{t'} = \theta_t + \Delta_e$,
- where, $\Delta_e \sim \mathrm{N}(0, \sigma_e^2)$.
- We put an improper prior (uniform on $(-\infty, \infty)$) on θ_r.
- We also make the variance random, but shared across siblings, $\sigma_t^2 \sim \mathrm{Exp}(1)$.

New York Schools Maths Test: Implementation

- The basic SIR-implementation of dc-smc.
- Using the natural hierarchical structure provided by the model.
- Given σ_t^2 and the θ_t at the leaves, the other random variables are multivariate normal.
- We instantiate values for θ_t only at the leaves.
- At internal node t', sample only $\sigma_{t'}^2$ and marginalize out $\theta_{t'}$.
- Each step of dc-smc therefore is either:
    i. At leaves sample $p_t \sim \mathrm{Beta}(1 + m_t, 1 + M_t - m_t)$ and set $\theta_t = \mathrm{logit}(p_t)$.
    ii. At internal nodes sample $\sigma_t^2 \sim \mathrm{Exp}(1)$.
- Java implementation:
  `https://github.com/alexandrebouchard/multilevelSMC`

New York Schools Maths Test: Results

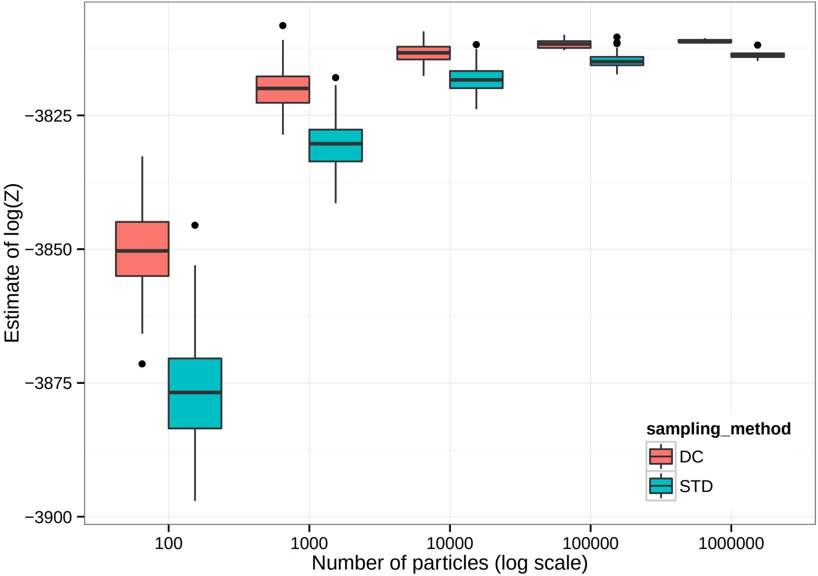


- D&C with 10 000 particles.
- Bronx County has the highest fraction (41%) of children (under 18) living below poverty level.[1]
- Queens has the second lowest (19.7%),
- after Richmond (Staten Island, 16.7%).
- Staten Island contains a single school district so the posterior distribution is much flatter for this borough.

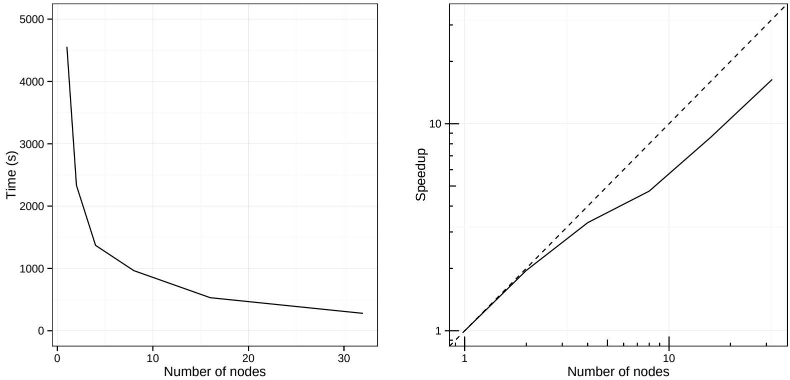
[1] Statistics from the New York State Poverty Report 2013,

`http://ams.nyscommunityaction.org/Resources/Documents/News/NYSCAAs_2013_Poverty_Report.pdf`

Normalising Constant Estimates



Distributed Implementation



Xeon X5650 2.66GHz processors connected by a non-blocking Infiniband 4X QDR network

Conclusions

- SMC $\approx$ SIR
- D&C-SMC $\approx$ SIR + Coalescence
- Distributed implementation is straightforward
- D&C strategy can improve even serial performance
- D&C-SMC inherits many theoretical guarantees from SMC
- Some questions remain unanswered:
  - How can we construct (near) optimal tree-decompositions?
- Some other interesting applications:
  - Parallel (in time) Smoothing (Ding and Gandy, 2018; Corneflos et al., 2022)
  - High-dimensional Filtering (Crucinio and Johansen, 2022)

References I

R. Chan, M. Pollock, A. M. Johansen, and G. O. Roberts. Divide-and-conquer Monte Carlo fusion. e-print 2110.07265, arXiv, 2021.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–551, 2002.

A. Corneflos, N. Chopin, and S. Särkkä. De-sequentialized monte carlo: a parallel-in-time particle smoother. *arXiv preprint arXiv:2202.02264*, 2022.

F. Crucinio and A. M. Johansen. A divide-and-conquer sequential monte carlo approach to high dimensional filtering. In preparation., 2022.

H. Dai, M. Pollock, and G. O. Roberts. Monte Carlo Fusion. *Journal of Applied Probability*, 56(1):174–191, 2019.

P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Probability and Its Applications. Springer Verlag, New York, 2004.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo methods for Bayesian Computation. In *Bayesian Statistics 8*. Oxford University Press, 2006.

Dong Ding and Axel Gandy. Tree-based particle smoothing algorithms in a hidden markov model. eprint 1808.08400, ArXiv Mathematics e-prints, 2018.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fiteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, 2011.

N. J. Gordon, S. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, April 1993.

A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagaris. Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, December 2010.

A. M. Johansen and A. Doucet. A note on the auxiliary particle filter. *Statistics and Probability Letters*, 78(12): 1498–1504, September 2008. doi: 10.1016/j.spl.2008.01.032.

J. Kuntz, F. R. Crucinio, and A. M. Johansen. Divide-and-conquer sequential Monte Carlo: Properties and limit theorems. e-print 2110.15782, arXiv, 2021.

F. Lindsten, A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. Schön, J. A. D. Aston, and A. Bouchard-Côté. Divide and conquer with sequential Monte Carlo samplers. *Journal of Computational and Graphical Statistics*, 26(2):445–458, 2017. doi: 10.1080/10618600.2016.1237363.

M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

Y. Zhou, A. M. Johansen, and J. A. D. Aston. Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016. doi: 10.1080/10618600.2015.1060885.