Divide-and-Conquer Sequential Monte Carlo

Adam M. Johansen

Joint work with:
John Aston, Alexandre Bouchard-Côté, Brent Kirkpatrick,
Fredrik Lindsten, Christian Næsseth, Thomas Schön

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

OxWaSP

OxWaSP Lunch Talk
October 20th, 2016

Outline

- Importance Sampling to Sequential Monte Carlo (SMC)
- SMC to Divide and Conquer SMC (DC-SMC)
- Some Theoretical Properties of DC-SMC
- Illustrative Applications
- Conclusions and Some (Open) Questions

Essential Problem

The Abstract Problem

- Given a density,
$$\pi(x) = \frac{\gamma(x)}{Z},$$
- such that $\gamma(x)$ can be evaluated pointwise,
- how can we approximate π
- and how about Z?
- Can we do so robustly?
- In a distributed setting?

Importance Sampling

- Simple identity: provided $\gamma \ll \mu$:

$$Z = \int \gamma(x)dx = \int \frac{\gamma(x)}{\mu(x)}\mu(x)dx$$

- So, if $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mu$, then:

unbiasedness $$\forall N : \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\gamma(X_i)}{\mu(X_i)}\right] = Z$$

slln $$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\gamma(X_i)}{\mu(X_i)}\varphi(X_i) \overset{\text{a.s.}}{\to} \gamma(\varphi)$$

clt $$\lim_{N\to\infty}\sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\gamma(X_i)}{\mu(X_i)}\varphi(X_i) - \gamma(\varphi)\right] \overset{\text{d}}{\to} W$$

where $W \sim \mathcal{N}\left(0, \mathrm{Var}\left[\frac{\gamma(X_i)}{\mu(X_1)}\varphi(X_1)\right]\right)$.

Sequential Importance Sampling

- Write
$$\gamma(x_{1:n}) = \gamma(x_1) \prod_{p=2}^{n} \gamma(x_p|x_{1:p-1}),$$
- define, for $p = 1, \ldots, n$
$$\gamma_p(x_{1:p}) = \gamma_1(x_1) \prod_{q=2}^{p} \gamma(x_q|x_{1:q-1}),$$
- then
$$\underbrace{\frac{\gamma(x_{1:n})}{\mu(x_{1:n})}}_{W_n(x_{1:n})} = \underbrace{\frac{\gamma_1(x_1)}{\mu_1(x_1)}}_{w_1(x_1)} \prod_{p=2}^{n} \underbrace{\frac{\gamma_p(x_{1:p})}{\gamma_{p-1}(x_{1:p-1})\mu_p(x_p|x_{1:p-1})}}_{w_p(x_{1:p})},$$
- and we can *sequentially* approximate $Z_p = \int \gamma_p(x_{1:p}) dx_{1:p}$.

Sequential Importance *Resampling* (SIR)

Given a sequence $\gamma_1(x_1), \gamma_2(x_{1:2}), \ldots$:

Initialisation, $n = 1$:

- Sample $X_1^1, \ldots, X_1^N \overset{\text{iid}}{\sim} \mu_1$
- Compute

$$W_1^i = \frac{\gamma_1(X_1^i)}{\mu_1^i(X_1^i)}$$

- Obtain $\widehat{Z}_1^N = \frac{1}{N} \sum_{i=1}^N W_1^i \qquad \widehat{\pi}_1^N = \frac{\sum_{i=1}^N W_1^i \delta_{X_1^i}}{\sum_{j=1}^N W_1^j}$

[This is *just* (self-normalized) importance sampling.]

Iteration, $n \leftarrow n+1$:

- Resample: sample $(X^1_{n,1:n-1}, \ldots, X^N_{n,1:n-1}) \overset{\text{iid}}{\sim} \sum_{i=1}^N \delta_{X^i_{n-1}}$
- Sample $X^i_{n,n} \sim q_n(\cdot|X^i_{n,1:n-1})$
- Compute

$$W^i_n = \frac{\gamma_n(X^i_{n,1:n})}{\gamma_{n-1}(X^i_{n,1:n-1}) \cdot q_n(X^i_{n,n}|X^i_{n,1:n-1})}.$$

- Obtain

$$\widehat{Z}^N_n = \widehat{Z}^N_{n-1} \cdot \frac{1}{N}\sum_{i=1}^N W^i_n \qquad \widehat{\pi}^N_n = \frac{\sum_{i=1}^N W^i_n \delta_{X^i_n}}{\sum_{j=1}^N W^j_n}.$$

SIR: Theoretical Justification

Under regularity conditions we still have:

unbiasedness

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

slln

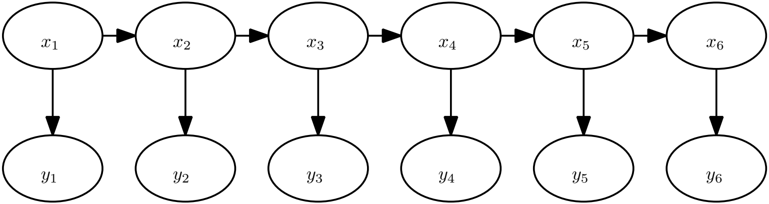
$$\lim_{N\to\infty} \widehat{\pi}_n^N(\varphi) \stackrel{\text{a.s.}}{=} \pi_n(\varphi)$$

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 becomes a little harder (cf., e.g. Del Moral (2004), Andrieu et al. 2010).

Simple Particle Filters: One Family of SIR Algorithms



- 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:

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

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

Bootstrap PFs and Similar

- Choosing

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

- and $q_p(x_p|x_{1:p-1}) = f^\theta(x_p|x_{p-1})$ yields the bootstrap particle filter of Gordon et al. (1993),
- whereas $q_p(x_p|x_{1:p-1}) = p(x_p|x_{p-1}, y_p, \theta)$ 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 $\pi_1, \ldots, \pi_n$ on *arbitrary* spaces, Del Moral et al. (2006) extend the space:

$$\tilde{\pi}_n(x_{1:n}) = \pi_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p)$$

$$\tilde{\gamma}_n(x_{1:n}) = \gamma_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p)$$

$$\tilde{Z}_n = \int \tilde{\gamma}_n(x_{1:n}) dx_{1:n}$$

$$= \int \gamma_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p) dx_{1:n} = \int \gamma_n(x_n) dx_n = Z_n$$

A Simple SMC Sampler

Given $\gamma_1, \ldots, \gamma_n$, on $(E, \mathcal{E})$, for $i = 1, \ldots, N$

- Sample $X_1^i \overset{\text{iid}}{\sim} \mu_1$ compute $W_1^i = \frac{\gamma_1(X_1^i)}{\mu_1(X_1^i)}$ and $\widehat{Z}_1^N = \frac{1}{N}\sum_{i=1}^N W_1^i$
- For $p = 2, \ldots, n$
  - Resample: $X_{n,n-1}^{1:N} \overset{\text{iid}}{\sim} \sum_{i=1}^N W_{n-1}^i \delta_{X_{n-1}}$.
  - Sample: $X_n^i \sim K_n(X_{n,1:n-1}^i, \cdot)$, where $\pi_n K_n = \pi_n$.
  - Compute: $W_n^i = \frac{\gamma_n(X_{n,n-1}^i)}{\gamma_{n-1}(X_{n,n-1}^i)}$.
  - Then $\widehat{Z}_n^N = \widehat{Z}_{n-1}^N \cdot \frac{1}{N}\sum_{i=1}^N W_n^i$,
  - and $\pi_n^N = \frac{\sum_{i=1}^n W_n^i \delta_{X_n^i}}{\sum_{j=1}^n W_n^j}$.

Bayesian Inference (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 $\gamma_p(\theta) = p(\theta) l(\theta; y_{1:m_p})$ for

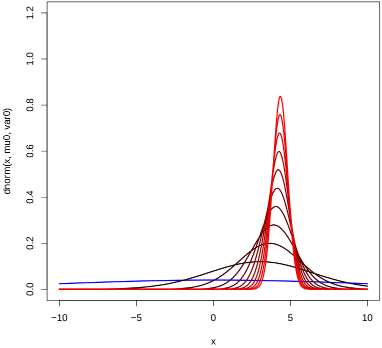
  $$m_1 = 0 < m_2 < \cdots < m_T = m$$

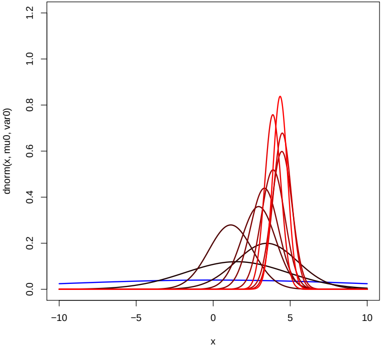
  Likelihood Tempering $\gamma_p(\theta) = p(\theta) l(\theta; y_{1:m})^{\beta_p}$ for

  $$\beta_1 = 0 < \beta_2 < \cdots < \beta_T = 1$$

  Something else?
- Here $Z_T = \int p(\theta) l(\theta; y_{1:n}) d\theta$ and $\gamma_T(\theta) \propto p(\theta | y_{1:n})$.
- 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.

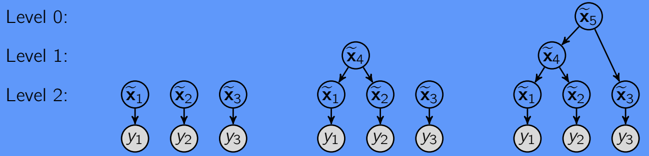
Proposals: Follow (Jasra et al., 2010): adapt to keep acceptance rate about right.

Question

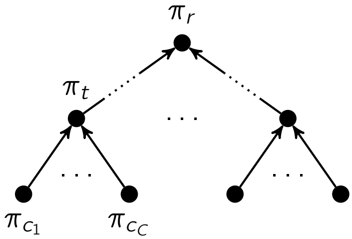
Are there better, practical approaches to specifying a sequence of distributions?

Divide-and-Conquer (Lindsten, J. et al., 2016)

Many models admit natural decompositions:

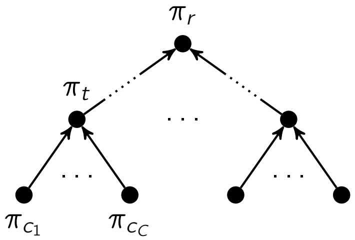


To which we can apply a divide-and-conquer strategy:



A few formalities...

- Use a tree, T of models (with rootward variable inclusion):



- Let $t \in T$ denote a node; $r \in T$ is the root.
- Let $\mathcal{C}(t) = \{c_1, \ldots, c_C\}$ denote the children of t.
- Let $\widetilde{\mathcal{X}}_t$ denote the space of variables included in t but *not* its children.
- dc-smc can be viewed as a recursion over this tree.

dc-smc(t) an extension of SIR

1. For $c \in \mathcal{C}(t)$:
   1.1 $(\{X_c^i, W_c^i\}_{i=1}^N, \widehat{Z}_c^N) \leftarrow$ dc-smc(c).
   1.2 Resample $\{\mathbf{x}_c^i, \mathbf{w}_c^i\}_{i=1}^N$ to obtain the equally weighted particle system $\{\hat{\mathbf{x}}_c^i, 1\}_{i=1}^N$.
2. For particle $i = 1 : N$:
   2.1 If $\tilde{\mathcal{X}}_t \neq \emptyset$, simulate $\widetilde{\mathbf{x}}_t^i \sim q_t(\cdot \mid \hat{\mathbf{x}}_{c_1}^i, \ldots, \hat{\mathbf{x}}_{c_C}^i)$, where $(c_1, c_2, \ldots, c_C) = \mathcal{C}(t)$;
   else $\widetilde{\mathbf{x}}_t^i \leftarrow \emptyset$.
   2.2 Set $\mathbf{x}_t^i = (\hat{\mathbf{x}}_{c_1}^i, \ldots, \hat{\mathbf{x}}_{c_C}^i, \widetilde{\mathbf{x}}_t^i)$.
   2.3 Compute $\mathbf{w}_t^i = \frac{\gamma_t(\mathrm{x}_t^i)}{\prod_{c \in \mathcal{C}(t)} \gamma_c(\hat{\mathrm{x}}_c^i)} \frac{1}{q_t(\widetilde{\mathrm{x}}_t^i | \hat{\mathrm{x}}_{c_1}^i, \ldots, \hat{\mathrm{x}}_{c_C}^i)}$.
3. Compute $\widehat{Z}_t^N = \left\{ \frac{1}{N} \sum_{i=1}^N \mathbf{w}_t^i \right\} \prod_{c \in \mathcal{C}(t)} \widehat{Z}_c^N$.
4. Return $(\{\mathbf{x}_t^i, \mathbf{w}_t^i\}_{i=1}^N, \widehat{Z}_t^N)$.

Theoretical Properties I

Unbiasedness of Normalising Constant Estimates

Provided that $\gamma_t \ll \otimes_{c \in \mathcal{C}(t)} \gamma_c \otimes q_t$ for every $t \in T$ and an unbiased, exchangeable resampling scheme is applied to every population at every iteration, we have for any $N \geq 1$:

$$\mathbb{E}[\widehat{Z}_r^N] = Z_r = \int \gamma_r(\mathbf{x}_r) d\mathbf{x}_r.$$

Strong Law of Large Numbers

Under regularity conditions the weighted particle system $(\mathbf{x}^i_{r,N}, \mathbf{w}^i_{r,N})_{i=1}^N$ generated by dc-smc(r) is consistent in that for all functions $f : \mathcal{Z} \to \mathbb{R}$ satisfying certain assumptions:

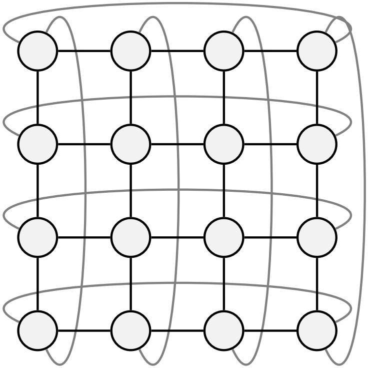
$$\sum_{i=1}^{N} \frac{\mathbf{w}_r^{N,i}}{\sum_{j=1}^{N} \mathbf{w}_r^{N,j}} f(\mathbf{x}^i_{r,N}) \overset{\text{a.s.}}{\to} \int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \qquad \text{as } N \to \infty.$$

Some (Importance) Extensions

1. Mixture Resampling
2. Tempering (Del Moral et al, 2006)
3. Adaptation (Zhou, J. and Aston, 2016)

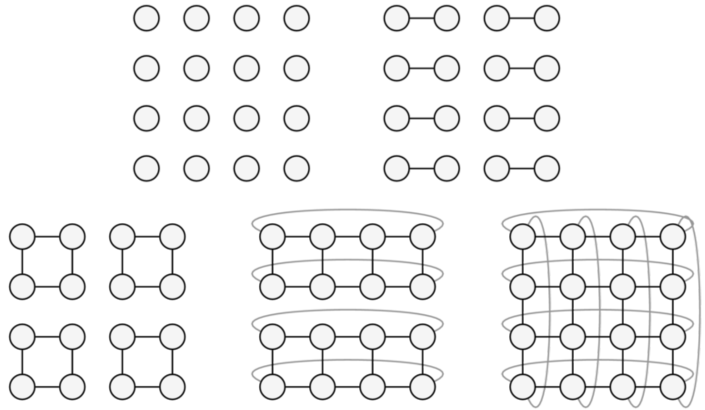
An Ising Model

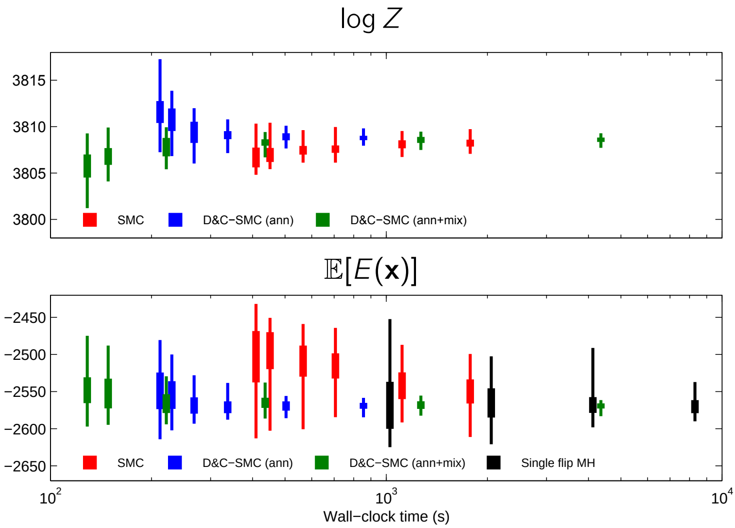
- k indexes $\mathcal{V} \in \mathcal{V} \subset \mathbb{Z}^2$
- $x_k \in \{-1, 1\}$
- $p(\mathbf{z}) \propto e^{-\beta E(\mathrm{z})}$, $\beta \geq 0$
- $E(\mathbf{z}) = -\sum_{(k,l)\in\mathcal{E}} x_k x_l$



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.

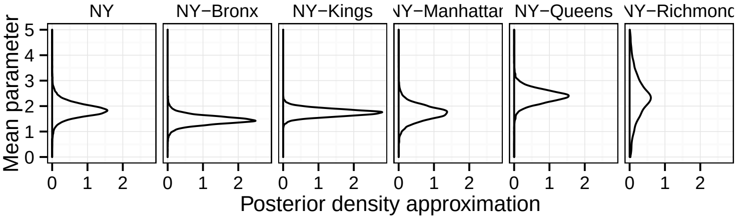
New York Schools Maths Test: Bayesian Model

- 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`

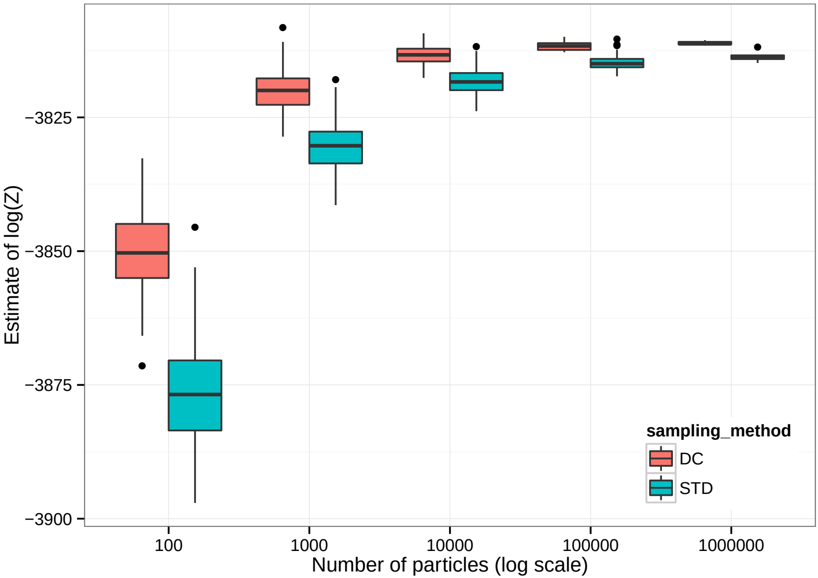
New York Schools Maths Test: Results



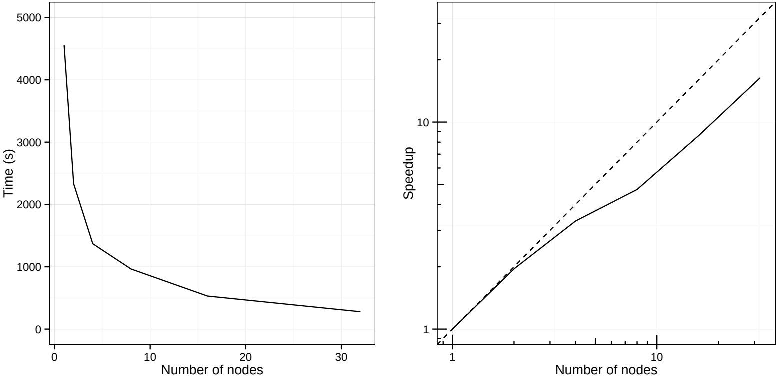
- DC with 10 000 particles.
- Bronx County has the highest fraction (41%) of children (under 18) living below poverty level.[1]
- Queens has the second lowest (19.7%),
- after Richmond (Staten Island, 16.7%).
- Staten Island contains a single school district so the posterior distribution is much flatter 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
- Some questions remain unanswered:
  - How can we construct (near) optimal tree-decompositions?
  - How much standard SMC theory can be extended to this setting?
- Some application areas are appealing:
  - Inference for phylogenetic trees in linguistics.
  - Principled aggregration of "mass univariate" analyses from neuroimaging.

References

[1] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo. *Journal of the Royal Statistical Society B*, 72(3):269–342, 2010.

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

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

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

[5] 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.

[6] 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.

[7] 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, Dec. 2010.

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

[9] 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*, 2016. In press.

[10] 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.