MultiLevel Sequential Monte Carlo Samplers

Alexandros Beskos

University College London

Athens, 9th May 2015

Joint work with:
Ajay Jasra, Yang Zhou NUS, Kody Law, Raul Tempone KAUST

- Research funded by Leverhulme Trust Prize

Outline

Outline

Introduction

- We look at the Multilevel Monte Carlo (MLMC) algorithm (Giles, 08).
- Another important MC technique is Sequential Monte Carlo (SMC) (Del Moral, 04).
- We show that SMC and Multilevel can be naturally combined, thus develop a new methodology: MLSMC.
- Main Reference:
  'MultiLevel Sequential Monte Carlo Samplers', arxiv preprint.

Outline

Multilevel Monte Carlo

- MLMC is relevant e.g. in the context of the discretization of a mathematical model, like an SDE or PDE.
- Numerical approximation gives rise to a resolution parameter $h_\ell = 2^{-\ell-\ell_0}$, $\ell \geq 0$, and sequence of distributions:

$$\eta_0, \eta_1, \ldots, \eta_L \to \eta_\infty$$

- The objective is to apply Monte Carlo approach to estimate expectations $E_{\eta_\infty}[g(U)]$.
- Standard MC chooses large enough L, and averages over iid samples from η_L.
- MLMC combines samples from all members $\eta_0, \eta_1, \ldots, \eta_L$ to get better estimates.

Example: SDE

- Consider the SDE:

$$dX_t = \alpha(X_t)dt + dW_t$$

- For sampling purposes one discretises (Euler-Maruyama):

$$X_{t+h_\ell} \approx X_t + \alpha(X_t)h_\ell + N(0, h_\ell)$$

  for resolution level $h_\ell = 2^{-\ell-\ell_0}$.

- This gives rise to $\eta_0, \eta_1, \ldots, \eta_L \to \eta_\infty$.
- We are interested in estimating e.g. $\mathbb{E}_{\eta_\infty}[g(X_T)]$.

Standard Monte Carlo

- Standard MC for estimating $\mathbb{E}_{\eta_\infty}[g(U)]$:
  1. Sample independently $U_L^{(i)} \sim \eta_L$, for $1 \leq i \leq N_L$.
  2. Monte-Carlo average:

$$\eta_L^{N_L}(g) := \frac{1}{N_L}\sum_{i=1}^{N_L} g(U_L^{(i)})$$

- Error/Cost analysis:

$$MSE_L = \mathbb{E}\big[\{\eta_L^{N_L}(g) - \eta_\infty(g)\}^2\big] = \mathcal{O}(h_L^{2\alpha}) + \mathcal{O}(\tfrac{1}{N_L})$$

$$CC_L = N_L \cdot h_L^{-\zeta}$$

- In the SDE example, $\alpha = \zeta = 1$. Thus, for $MSE_L = \mathcal{O}(\epsilon^2)$ one (optimally) needs $h_L = \mathcal{O}(\epsilon)$, $N_L = \mathcal{O}(\epsilon^{-2})$, thus computations of $CC_L = \mathcal{O}(\epsilon^{-3})$.

MultiLevel Monte Carlo (Giles, 08)

- Apply a telescopic sum:

$$\mathbb{E}_{\eta_L}[g(U)] = \sum_{\ell=0}^{L} \{ \mathbb{E}_{\eta_\ell}[g(U)] - \mathbb{E}_{\eta_{\ell-1}}[g(U)] \} =: \sum_{\ell=0}^{L} Y_\ell$$

- Notice that $Y_\ell \to 0$, as $\ell \to \infty$.
- Each Y_ℓ term can be unbiasedly approximated by:

$$Y_\ell^{N_\ell} = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \{ g(U_\ell^{(i)}) - g(U_{\ell-1}^{(i)}) \}$$

- Overall estimate:

$$\widehat{Y}_{ML} = \sum_{l=0}^{L} Y_\ell^{N_\ell}$$

Error Analysis, MLMC

- Recall that: $\widehat{Y}_{ML} = \sum_{\ell=0}^{L} Y_\ell^{N_\ell}$.
- The idea is that for increasing ℓ, the cost per sample increases, but one needs fewer samples to control errors. Optimal allocation of N_ℓ's can give benefit vs standard MC.
- Error/cost analysis:

$$VAR = \sum_{\ell=0}^{L} V_\ell N_\ell^{-1}\ , \quad CC = \sum_{\ell=0}^{L} N_\ell C_\ell$$

- Constrained optimisation gives optimal allocation of resources: $N_\ell = \mathcal{O}(c(\epsilon, L)\sqrt{V_\ell / C_\ell})$, $0 \leq \ell \leq L$.
- Replacing $V_\ell = h_\ell^\beta$, $C_\ell = h_\ell^{-\zeta}$, $b_L = h_L^\alpha$ we give concrete results.

MLMC: Final Statements

- Bias is $h_L^\alpha \propto \{2^{-L}\}^\alpha$: set $L = \mathcal{O}|\log \epsilon|$, to get $b_L^2 = \mathcal{O}(\epsilon^2)$.
- Under optimal allocation of N_ℓ's:

$$VAR_\ell = \mathcal{O}(c^{-1}(\epsilon, L)\sqrt{V_\ell C_\ell}) = \mathcal{O}(c^{-1}(\epsilon, L)\, h_\ell^{(\beta-\zeta)/2})$$
$$C_\ell = \mathcal{O}(c(\epsilon, L)\sqrt{V_\ell C_\ell}) = \mathcal{O}(c(\epsilon, L)\, h_\ell^{(\beta-\zeta)/2})$$

- So we have the cases:
  1. $\beta = \zeta$: choosing $c(\epsilon, L) = L\epsilon^{-2} = \mathcal{O}(|\log \epsilon|\epsilon)^{-2}$ gives:
  $$MSE = \mathcal{O}(\epsilon^2)\ ; \quad CC = \mathcal{O}(|\log^2 \epsilon|\epsilon^{-2})$$
  2. $\beta > \zeta$: choosing $c(\epsilon, L) = \mathcal{O}(\epsilon^{-2})$ gives:
  $$MSE = \mathcal{O}(\epsilon^2)\ ; \quad CC = \mathcal{O}(\epsilon^{-2})$$
  3. $\beta < \zeta$: Other considerations ...

Limitations of Analysis

- Thus, with MLMC one can achieve the same error as with standard MC, but faster.
- Unfortunately, the analysis in Giles (08) assumes iid sampling at each level, which is unrealistic for numerous potential applications.
- Not surprisingly, several recent works have tried to move beyond original set-up.

Outline

PDE Model

- We first describe an inference problem that will be used to illustrate the methodology.
- Let $\Omega \subset \mathbb{R}^d$, we consider the following elliptic PDE:

$$
\begin{aligned}
-\nabla \cdot (\widehat{u} \nabla p) = f \,, &\quad on \;\; \Omega \\
p = 0 \,, &\quad on \;\; \partial\Omega,
\end{aligned}
$$

where

$$\widehat{u}(x) = \bar{u}(x) + \sum_{k=1}^{K} u_k \sigma_k \phi_k(x)$$

PDE Model

- The first equation expresses continuity of mass and here f is assumed to be known and characterizes the source/sink configuration.
- These type of equations can be used for problems in hydrology.
- That is, the estimation of subsurface flow from measurements of the pressure at certain locations in Ω.
- Some technical conditions required to ensure that the solution to the PDE exists.

Numerical Solution of PDE

- Typically, one can only solve the PDE numerically.
- We look at the finite element method, which defines a discretization of the problem, and gives rise to the diameter of the employed triangulation, $h > 0$.
- That is, given PDE coefficient $u = \{u_1, u_2, \ldots, u_K\}$, and given triangulation $\mathcal{T}_h$, we have a numerical solution:

$$p_h(\cdot\,;\,u) \to p(\cdot\,;u)\ , \quad h \to 0$$

- PDE theory provides error analysis for the numerical solver.

Bayesian Inverse Problem

- We have unknown $u = \{u_k\}_{k=1}^{K} \in E := \prod_{k=1}^{K}[-1, 1]$, and assume a prior distribution $u_k \sim U[-1, 1]$ iid.
- That is, the coefficient in the PDE is random and it is this quantity that we want to infer, on the basis of data.
- Such randomness is now an important assumption in applied mathematics, to address possible uncertainties in the assumed PDE.

Posterior Distribution

- For $p(\cdot; u)$ being the weak solution for parameter value u, we assume data:

$$\textit{DATA}: \qquad y = \mathcal{G}(p) + \xi\ , \quad \xi \sim N(0, \Gamma)\ , \quad \xi \perp u$$

for the linear observation operator:

$$\mathcal{G}(p) = [g_1(p), \cdots, g_M(p)]^\top$$

- The unnormalized posterior of $u|y$ is given by:

$$\gamma(u) = \exp\{-\Phi[\mathcal{G}(p(\cdot; u))]\}\ ; \quad \Phi(\mathcal{G}) = \frac{1}{2}\,|\mathcal{G} - y|_\Gamma^2$$

- The normalised posterior density is:

$$\textit{TARGET}: \qquad \eta(u) = \gamma(u)/Z\ , \quad Z = \int_E \gamma(u)du$$

Sequence of Posteriors

- In practice, one solves the PDE numerically, for a choice of diameter h.
- Assuming a sequence of choices $h_\ell = 2^{-\ell-\ell_0}$, we have a sequence of solutions of increasing accuracy, $p_{h_l}(\cdot; u)$.
- And, a corresponding sequence of posteriors:

$$\eta_\ell(u) \propto \exp\left\{ -\Phi[\mathcal{G}(p_{h_\ell}(\cdot; u))] \right\}$$

- Thus, we have a sequence $\eta_0, \eta_1, \ldots, \eta_L \to \eta_\infty$, but cannot sample iid from any of these.
- Objective: Estimate $\mathbb{E}_{\eta_\infty}[g(U)]$ as fast as possible.

Multi-Level Principle (again)

- Recall that $\eta_\ell(u) = \gamma_\ell(u)/Z, \qquad Z_\ell = \int_E \gamma_\ell(u)du$.
- We use ML identity:

$$\mathbb{E}_{\eta_L}[g(U)] = \mathbb{E}_{\eta_0}[g(U)] + \sum_{\ell=1}^{L}\Big\{\mathbb{E}_{\eta_\ell}[g(U)] - \mathbb{E}_{\eta_{\ell-1}}[g(U)]\Big\}$$

$$= \mathbb{E}_{\eta_0}[g(U)] + \sum_{\ell=1}^{L}\mathbb{E}_{\eta_{\ell-1}}\Big[\Big(\frac{\gamma_\ell(U)Z_{\ell-1}}{\gamma_{\ell-1}(U)Z_\ell} - 1\Big)g(U)\Big]$$

- Notice that the summands diminish in the limit as

$$(\gamma_\ell/\gamma_{\ell-1})(u) = \exp\big\{ -\Phi[\mathcal{G}(p_{h_\ell}(\,\cdot\,;u))] + \Phi[\mathcal{G}(p_{h_{\ell-1}}(\,\cdot\,;u))]\,\big\}$$

converges to 1 as $\ell \to \infty$.

Sampling from Sequence of Distributions

- It is impossible to get iid sampling and carry out related error analysis.
- Instead, SMC samplers provide a natural mechanism for getting correlated samples from sequence of slowly changing distributions and obtaining estimates of normalising constants.
- Coming up: SMC Samplers.

Outline

Sequential Monte Carlo

- Assume a sequence of distributions

$$\eta_0, \eta_1, \ldots, \eta_L$$

- SMC provides an iterative procedure that provides an evolving, consistent particle population from the laws in the sequence.
- A lot of theory has been developed the last 10-15 year on the accuracy of the method.
- For instance, one can get several asymptotic, non-asymptotic results using the Feynman-Kac interpretation of such methods.
- See Del Moral, 04.

SMC Sampler, a Version of

- The Algorithm:

0. Sample $U_0^{(1:N_0)} \sim \eta_0$. Set $\ell = 0$.
1. Given $U_\ell^{(1:N_\ell)} \sim \eta_\ell$:
   - Assign weights $G_\ell^{(1:N_\ell)} = (\gamma_{\ell+1}/\gamma_\ell)(U_\ell^{1:N_\ell})$
   - Resample $N_{\ell+1}$ particles iid from approximation:

$$\sum_{i=1}^{N_\ell} \delta_{U_\ell^{(i)}}(dU) \cdot G_\ell^{(i)}$$

   - Mutate: 'disperse' each resampled particle according to a Markov kernel $M_{\ell+1}(U, dU')$ that preserves $\eta_{\ell+1}$.
2. We have now obtained $U_{\ell+1}^{(1:N_{\ell+1})} \sim \eta_{\ell+1}$.
3. Set $\ell = \ell + 1$. Return to Step 1.

Feynman Kac Models

- The above corresponds to a particular case of a Monte-Carlo approximation of a Feynman Kac model.
- Different choice of potential $G_\ell(u)$ and Markov kernel $M_\ell(u, du')$ will give for instance the particle filter for hidden Markov models.
- Del Moral (04,11) provides a rigorous error analysis in a general setting.
- Average of weights gives (unbiased) estimate of ratio of normalising constants:

$$\frac{1}{N_\ell}\sum_{i=1}^{N_\ell} G_\ell^{(i)} \approx \frac{Z_{\ell+1}}{Z_\ell}$$

Outline

MultiLevel Sequential Monte Carlo

- We now develop a new algorithm that:
  - Applies the ML identity.
  - Deals with the fact that one cannot achieve an iid sampling from the $(\eta_\ell)_{0 \le \ell \le L}$.
  - Still achieve the same error for less work versus iid sampling from η_ℓ (even though the latter is not possible).
- Main Challenges:
  - Extend standard SMC error analysis.
  - Obtain MC error bounds that can then be treated by PDE error analysis ...

The Algorithm (a)

- Recall the ML identity:

$$\mathbb{E}_{\eta_L}[g(U)] = \mathbb{E}_{\eta_0}[g(U)] + \sum_{\ell=1}^{L} \Big\{ \mathbb{E}_{\eta_\ell}[g(U)] - \mathbb{E}_{\eta_{\ell-1}}[g(U)] \Big\}$$
$$= \mathbb{E}_{\eta_0}[g(U)] + \sum_{\ell=1}^{L} \mathbb{E}_{\eta_{\ell-1}} \Big[\Big(\frac{\gamma_\ell(U) Z_{\ell-1}}{\gamma_{\ell-1}(U) Z_\ell} - 1 \Big) g(U) \Big].$$

- Given unnormalised sequence $\eta_0, \eta_1, \ldots, \eta_L$, SMC provides population of particles:

$$U_0^{1:N_0} \sim \eta_0\,, \quad U_1^{1:N_1} \sim \eta_1\,, \ldots\,, U_L^{1:N_L} \sim \eta_L$$

and estimates of ratio of normalising constants $Z_\ell / Z_{\ell-1}$.

The Algorithm (b)

- The MLSMC estimator of $\eta_L(g)$ is given by

$$\widehat{Y} := \eta_0^{N_0}(g) + \sum_{\ell=1}^{L} \left\{ \frac{\eta_{\ell-1}^{N_{\ell-1}}(gG_{\ell-1})}{\eta_{\ell-1}^{N_{\ell-1}}(G_{\ell-1})} - \eta_{\ell-1}^{N_{\ell-1}}(g) \right\}.$$

- Using standard theory for SMC (e.g. Del Moral (04)) one can show that this estimator is consistent.

Error Analysis

i) the $L+1$ terms above are not unbiased estimates of $\mathbb{E}_{\eta_\ell}[g(U)] - \mathbb{E}_{\eta_{\ell-1}}[g(U)]$, so decompose MSE as:

$$\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_\infty}[g(U)]\}^2\big] \leq$$
$$2\,\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2\big] + 2\,\{\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)]\}^2$$

ii) the same $L+1$ estimates are not independent, so a more complex error analysis will be required to characterise $\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2\big]$.

PDE Error Analysis

- We can get the following results (under conditions):
  - $\|p_{h_\ell}(\,\cdot\,;u) - p(\,\cdot\,;u)\|_V \leq C \cdot h_\ell$
  - $1 - \mathcal{O}(h_\ell) \leq e^{-\Phi[\mathcal{G}(p_{h_\ell}(\cdot;u))] + \Phi[\mathcal{G}(p(\cdot;u))]} \leq 1 + \mathcal{O}(h_l)$
  - Recall that:

$$
\begin{aligned}
G_{\ell-1}(u) &= (\gamma_\ell/\gamma_{\ell-1})(u) \\
&= \exp\Big\{ -\Phi[\mathcal{G}(p_{h_\ell}(\cdot;u))] + \Phi[\mathcal{G}(p_{h_{\ell-1}}(\cdot;u))] \Big\}
\end{aligned}
$$

  - $1 - \mathcal{O}(h_\ell) \leq (\gamma_\ell/\gamma_{\ell-1})(u) \leq 1 + \mathcal{O}(h_\ell)$
  - From the above, one can obtain the bias:

$$
|\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)]| \leq \int |g(u)| \left| \tfrac{\gamma_L}{\gamma_\infty}(u) \tfrac{Z_\infty}{Z_L} - 1 \right| \leq C \cdot h_L
$$

Conditions

- The following conditions imply stability of SMC sampler:

(A1) There exist $0 < \underline{C} < \overline{C} < +\infty$ such that

$$\sup_{1\leq\ell}\sup_{u\in E} G_\ell(u) \leq \overline{C}$$
$$\inf_{1\leq\ell}\inf_{u\in E} G_\ell(u) \geq \underline{C}$$

(A2) There exist a $\rho \in (0,1)$ such that for any $1 \leq \ell$, $(u,v) \in E^2$, $A \in \mathcal{E}$

$$\int_A M_\ell(u, du') \geq \rho \int_A M_\ell(v, dv')$$

SMC Error Analysis

- Recall that:

$$\widehat{Y} = \eta_0^{N_0}(g) + \sum_{\ell=1}^{L} \Bigg\{ \underbrace{\frac{\eta_{\ell-1}^{N_{\ell-1}}(gG_{\ell-1})}{\eta_{\ell-1}^{N_{\ell-1}}(G_{\ell-1})} - \eta_{\ell-1}^{N_{\ell-1}}(g)}_{Y_{\ell-1}^{N_{\ell-1}}} \Bigg\}$$

- We want to compare the above with:

$$\mathbb{E}_{\eta_L}[g(U)] = \eta_0(g) + \sum_{\ell=1}^{N} \Bigg\{ \underbrace{\frac{\eta_{\ell-1}(gG_{\ell-1})}{\eta_{\ell-1}(G_{\ell-1})} - \eta_{\ell-1}(g)}_{Y_{\ell-1}} \Bigg\}$$

- We want to bound:

$$\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2\big]$$

.

SMC Error Analysis

- Standard results (Del Moral, 04) give:

$$\mathbb{E}\Big[\{Y_{\ell-1}^{N_{\ell-1}} - Y_{\ell-1}\}^2\Big] \leq \frac{C\,\|\frac{Z_\ell}{Z_{\ell-1}}G_{\ell-1} - 1\|_\infty^2}{N_{\ell-1}}$$

- We also need to control cross-population terms:

$$\sum_{2\leq\ell<q\leq L} \mathbb{E}\Big[(Y_{\ell-1}^{N_{\ell-1}} - Y_{\ell-1})(Y_{q-1}^{N_{q-1}} - Y_{q-1})\Big]$$

- Cauchy-Scwartz would give upper bound:

$$\sum_{2\leq\ell\leq q\leq L} \|\frac{Z_\ell}{Z_{\ell-1}}G_{\ell-1} - 1\|_\infty^2 \frac{1}{N_{q-1}}$$

but is not enough; we need to look deeper into SMC error analysis to better control cross-population terms.

Error Analysis

- The strategy is to consider a powerful decomposition of Del Moral et al. (12):

$$[\eta_n^{N_n} - \eta_n](\varphi) = \sum_{p=0}^{n} \frac{V_p^{N_p}(D_{p,n}(\varphi))}{\sqrt{N_p}} + \sum_{p=0}^{n-1} R_{p+1}^{N_p}(D_{p,n}(\varphi))$$

- Here, $V_p^{N_p}$ are local errors due to sampling, and $R_{p+1}^{N_p}$ are remainders.
- We can obtain bounds for all above quantities, e.g.:
  - $\|D_{p,n}(\varphi)\|_\infty \leq C\kappa^{n-p}\|\varphi\|_\infty$
- We then combine with Cauchy-Scwartz.

Main result

Theorem (BJLTZ15)

Assume (A1-2). For any $g \in \mathcal{B}_b(E)$, *with* $\|g\|_\infty = 1$,

$$\mathbb{E}[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2] \lesssim \frac{1}{N_0} + \sum_{\ell=1}^{L-1}\Big\{\frac{V_\ell}{N_{\ell-1}} + \Big(\frac{V_\ell}{N_{\ell-1}}\Big)^{1/2}\sum_{q=\ell+1}^{L-1}\frac{V_q^{1/2}}{N_q}\Big\}$$

where we have set $V_\ell = \|\frac{Z_{\ell-1}}{Z_\ell}G_{\ell-1} - 1\|_\infty^2$.

- Computational cost is $\sum_{\ell=0}^{L-1} C_\ell N_\ell$
- As in standard ML, we apply constrained optimisation for error (ignoring the “difficult” term in the Theorem).

Complete Picture

- Final result depends on the constants α, β, ζ characterising:
  1. $b_L = \mathcal{O}(h_L^{\alpha})$ (we have seen, for PDE $\alpha = 1$).
  2. computational cost, $C_l = \mathcal{O}(h_l^{\zeta})$ (for PDE, in 2D, $\zeta = 2$)
  3. "variance" $V_l = \mathcal{O}(h_l^{\beta})$ (for PDE, $\beta = 2$).
- Thus, in the PDE application in 1D, one can obtain MSE of size $\mathcal{O}(\epsilon^2)$ with $\mathcal{O}(\epsilon^{-2})$ computations.

Outline

Simulations

- We now illustrate the method in the context of our Bayesian inverse problem.
- Let $\Omega = [0, 1]$, and let $f(x) = 100x$.
- $K = 2$, $\bar{u} = 0.15$, $\sigma_1 = 0.1$, $\sigma_2 = 0.025$, $\phi_1(x) = \sin(\pi x)$, and $\phi_2(x) = \cos(2\pi x)$.

Numerical Details

- The forward problem at level l is solved using piecewise linear shape functions on a uniform mesh with mesh width $h_l = 2^{-(l+l_0)}$, with $l_0 = 3$.
- g is the solution of the forward problem at the midpoint of the domain $g(u) = p(0.5; u)$, the observation operator is $\mathcal{G}(u) = [p(0.25), p(0.75)]^\top$, and the observational noise is taken to be $\Gamma = 0.25^2 I$.

Error as Function of Runtime

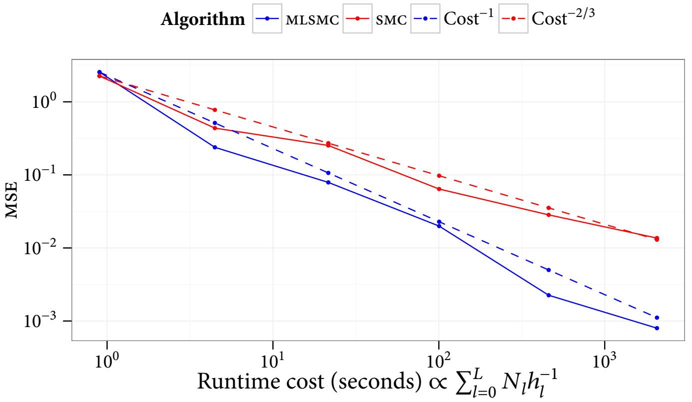

Figure: MSE vs computational time, for SMC and MLSMC.

Outline

Takehome Messages

- We have combined two popular MC techniques, to develop new methodology.
- Multilevel Sequential Monte Carlo sampler (MLSMC) can perform asymptotically as well as MLMC.
- Cost-to-ε could be asymptotically the same as for exact sampling! (When $\beta > \zeta$.)
- If $\beta = \zeta$, cost is somewhat higher, analogous to standard MLMC.

References

- **[BJLTZ15]**: Beskos, Jasra, Law, Tempone, Zhou. "Multilevel SMC Samplers." arXiv (2015).
- **[G08]**: Giles. "Multilevel Monte Carlo." Op. Res., 56, 607-617 (2008).
- **[DDJ12]**: Del Moral, Doucet, Jasra. "On adaptive resampling procedures for SMC." Bernoulli, 18, 252-272 (2012).
- **[D04]**: Del Moral. "Feynman-Kac Formulae." Springer: New York (2004).