Multilevel Sequential Monte Carlo Samplers

Ajay Jasra

National University of Singapore

KAUST

Introduction

Bayesian Inverse Problem

Multilevel Monte Carlo

Multilevel Sequential Monte Carlo

Theoretical Results

Simulations

Summary

Introduction

- In the following talk I will summarize some recent work on the development of multilevel Monte Carlo (MLMC).
- In particular, I detail a sequential Monte Carlo (SMC) extension.
- This approach is developed for Bayesian inverse problems.

- The basic idea of MLMC is in the context of the discretization of a problem, e.g. an SDE or PDE.
- Using the 'ML identity' one can estimate expectations w.r.t. probability laws associated to the discretization.
- Moreover, for a given level of error the amount of work can be reduced.

- The MLMC method assumes one can sample from the associated probability laws exactly.
- For many problems of practical importance, this is not possible.
- We show how (one) SMC version of the ML identity can be used.
- Under assumptions, for a given level of error the amount of work can be reduced.

Bayesian Inverse Problem

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

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

where

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

- 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.

- Define $u = \{u_k\}_{k=1}^K \in E := \prod_{k=1}^K [-1, 1]$, with $u_k \sim U[-1, 1]$ i.i.d. This determines the prior distribution for u.
- So the coefficient in the PDE are random and it is this quantity that we want to infer, on the basis of data.
- The randomness is now an important assumption in applied mathematics, to address possible uncertainties in the assumed PDE.

- Let $p(\cdot; u)$ denote the weak solution for parameter value u, and define

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

- It is assumed that the data take the form

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

- The unnormalized density of $u|y$ is given by

$$\gamma(u) = e^{-\Phi[\mathcal{G}(p(\cdot;u))]} \; ; \quad \Phi(\mathcal{G}) = \tfrac{1}{2} \left| \mathcal{G} - y \right|_{\Gamma}^2 .$$

- The posterior density is

$$\text{TARGET :} \qquad \eta(u) = \frac{\gamma(u)}{Z}, \quad Z = \int_E \gamma(u) du.$$

- For this particular posterior, in order to evaluate it pointwise, one needs to solve a PDE.
- The main problem is that one does not know how to solve the PDE exactly.
- As a result, one can consider the posterior associated to a numerical solution of the PDE.
- We will use the finite element method, which uses a discretization of the problem.

- We denote the 'finest possible' discretization by h_L.
- Finest possible essentially refers to the discretization which has the maximal computational cost that we are prepared (or allowed) to spend.
- We denote the posterior associated to the finite element discretization level h_l as η_l.

Multilevel Monte Carlo

- Aim: Compute $\eta(g) := \mathbb{E}_\eta(g)$ for many $g : E \to \mathbb{R}$.
- There are (at least) two difficulties here:
  - We do not know how to calculate $\eta(u)$.
  - We do not know how to calculate $\eta(g)$.
- Let us assume that we can sample from η_l for any l we desire.
- Then one can use the Monte Carlo method.

- Sample $U_L^{(i)} \sim \eta_L$ i.i.d., and approximate

$$\eta_L(g) := \mathbb{E}_{\eta_L}(g) \approx \hat{Y}_L^{N_L} := \frac{1}{N_L} \sum_{i=1}^{N_L} g(U_L^{(i)}).$$

- Mean square error (MSE) $\mathbb{E}[\{\hat{Y}_L - \mathbb{E}_{\eta_\infty}[g(U)]\}^2]$ splits into

$$\underbrace{\mathbb{E}\{\hat{Y}_L - \mathbb{E}_{\eta_L}[g(U)]\}^2}_{\text{variance}=\mathcal{O}(N_L^{-1})} + \underbrace{\{\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)]\}^2}_{\text{bias}}$$

- Cost to compute MSE$= \mathcal{O}(\varepsilon^2)$ is Cost$(U_L^{(i)}) \times \varepsilon^{-2}$.
- This is not taking into account the bias, which one can only be changed via h_L.
- The bias can be very problem specific. E.g. the behaviour for PDEs/SDEs can be very different, depending on the context.

- Introduce a hierarchy of discretization levels $\{\eta_l\}_{l=1}^{L}$ and define $Y_l = \{\mathbb{E}_{\eta_l}[g(U)] - \mathbb{E}_{\eta_{l-1}}[g(U)]\}$, with $\eta_{-1} := 0$.
- Observe the telescopic sum

$$\mathbb{E}_{\eta_L}[g(U)] = \sum_{l=0}^{L} Y_l$$

- Each term can be unbiasedly approximated by

$$Y_l^{N_l} = \frac{1}{N_l} \sum_{i=1}^{N_l} \{g(U_l^{(i)}) - g(U_{l-1}^{(i)})\}$$

where $g(U_{-1}^{(i)}) := 0$.

- Sample $U_{l,l-1}^{(i)} \sim \eta_{l,l-1}$ i.i.d., and approximate

$$\eta_L(g) \approx \hat{Y}_{L,\text{Multi}} := \sum_{l=0}^{L} Y_l^{N_l} \ .$$

- Mean square error (MSE) given by

$$\mathbb{E}\{\hat{Y}_{L,\text{Multi}} - \mathbb{E}_{\eta_\infty}[g(U)]\}^2 =$$
$$\underbrace{\mathbb{E}\{\hat{Y}_{L,\text{Multi}} - \mathbb{E}_{\eta_L}[g(U)]\}^2}_{\text{variance}=\sum_{l=0}^{L} V_l/N_l} + \underbrace{\{\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)]\}^2}_{\text{bias}} \ .$$

- Fix bias by choosing L. Minimize variance as a function of $\{N_l\}_{l=0}^L$ for a fixed Cost$= \sum_{l=0}^L C_l N_l \Rightarrow N_l \propto \sqrt{V_l/C_l}$.
- Assume $h_l = M^{-l}$ and there are α, and $\beta > \zeta$ such that
  - (i) weak error $|\mathbb{E}[g(U_l) - g(U)]| = \mathcal{O}(h_l^\alpha)$.
  - (ii) strong error $\mathbb{E}|g(U_l) - g(U)|^2 = \mathcal{O}(h_l^\beta) \Rightarrow V_l = \mathcal{O}(h_l^\beta)$,
  - (iii) computational cost for a realisation of $g(U_l) - g(U_{l-1})$, $C_l = \mathcal{O}(h_l^{-\zeta})$.

- In both cases, require $h_L^\alpha = \mathcal{O}(\varepsilon) \Rightarrow L \propto |\log \varepsilon|$.
- Single level cost $C = \mathcal{O}(\varepsilon^{-\zeta/\alpha-2})$: cost per sample is $C_L \propto \varepsilon^{-\zeta/\alpha}$, and $V_L \propto \varepsilon^2 \Rightarrow N_L \propto \varepsilon^{-2}$.
- Multilevel cost $C_{\mathrm{ML}} = \mathcal{O}(\varepsilon^{-2})$: $N_l \propto \varepsilon^{-2} h_l^{(\beta+\zeta)/2}$, so $V \propto \epsilon^2 \sum_{l=0}^{L} h_l^{(\beta-\zeta)/2} \propto \varepsilon^2$ (Giles, 2008) – cost of simulating a scalar random variable.
- Example: Milstein solution of SDE

$$C = \mathcal{O}(\varepsilon^{-3}) \quad \text{vs.} \quad C_{\mathrm{ML}} = \mathcal{O}(\varepsilon^{-2}).$$

- The main point is that one can achieve the same error for less work, in some contexts.
- The problem is that for our Bayesian inverse problem, one cannot achieve an i.i.d. sampling from the $(\eta_l)_{0\leq l\leq L}$.
- For future reference:

$$\eta_l(u) = \frac{\gamma_l(u)}{Z_l}.$$

Multilevel Sequential Monte Carlo

- We now present a method that can:
  - Implement a version of ML identity.
  - Deal with the fact that one cannot achieve an i.i.d. sampling from the $(\eta_l)_{0\leq l\leq L}$.
  - Still achieve the same error for less work versus i.i.d. sampling from η_L (even though the latter is not possible).

- To present our approach, we need some notations.
- $(E, \mathcal{E})$ is a measurable space.
- $\mathcal{B}_b(E)$ is the class of bounded and measurable real-valued functions, with $\|f\|_\infty := \sup_{u \in E} |f(u)|$.

- Consider $K : E \times \mathcal{E} \to \mathbb{R}_+$.
- For a finite measure μ on $(E, \mathcal{E})$ and $f : E \to \mathbb{R}$ use notations

$$\mu K : A \mapsto \int K(u, A)\, \mu(du)\ ; \quad Kf : u \mapsto \int f(v)\, K(u, dv).$$

- Also write $\mu(f) = \int f(u)\mu(du)$.

- Distributions η_l dictated by an accuracy parameter h_l (here FEM mesh diameter) $\infty > h_0 > h_1 \cdots > h_\infty = 0$.
- Recall, we want to approximate $\mathbb{E}_{\eta_L}[g(U)] = \eta_L(g) = \int_E g(u)\eta_L(u)du$.
- Idea: combine sequential importance resampling (selection) along the hierarchy, and mutation by MCMC kernels.
- This is a version of SMC samplers (Del Moral et al. 2006).

- Initialize i.i.d. $U_0^i \sim \eta_0, i = 1, \ldots N$.
- Resample $\{\widehat{U}_l^i\}_{i=1}^N$ according to the weights $\{G_l(U_l^i) = (\gamma_{l+1}/\gamma_l)(U_l^i)\}_{i=1}^N$.
- Draw $U_{l+1}^i \sim M_{l+1}(\widehat{U}_l^i, \cdot)$, where M_{l+1} is an MCMC kernel such that $\eta_{l+1} M_{l+1} = \eta_{l+1}$. Repeat for $l \in \{0, \ldots, L-2\}$.
- For $\varphi : E \to \mathbb{R}$, $l \in \{0, \ldots, L\}$, we have the following estimators

$$\mathbb{E}_{\eta_l}[\varphi(U)] \approx \eta_l^N(\varphi) := \frac{1}{N} \sum_{i=1}^N \varphi(U_l^i) \ .$$

- Observe the decomposition

$$\mathbb{E}_{\eta_L}[g(U)] = \mathbb{E}_{\eta_0}[g(U)] + \sum_{l=1}^{L} \left\{ \mathbb{E}_{\eta_l}[g(U)] - \mathbb{E}_{\eta_{l-1}}[g(U)] \right\}$$

$$= \mathbb{E}_{\eta_0}[g(U)] + \sum_{l=1}^{L} \mathbb{E}_{\eta_{l-1}} \left[\left(\frac{\gamma_l(U) Z_{l-1}}{\gamma_{l-1}(U) Z_l} - 1 \right) g(U) \right]. \quad \dagger$$

- We want to approximate $\dagger$ using SMC samplers.

- Sample $(U_0^{1:N_0}, \ldots, U_{L-1}^{1:N_{L-1}})$, with $+\infty > N_0 \geq N_1 \geq N_{L-1} \geq 1$ as for SMC samplers.
- The joint probability distribution is

$$\prod_{i=1}^{N_0} \eta_0(du_0^i) \prod_{l=1}^{L-1} \prod_{i=1}^{N_l} \frac{\eta_{l-1}^{N_{l-1}}(G_{l-1} M_l(du_l^i))}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} .$$

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

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

- Using standard theory for SMC (e.g. Del Moral (2004)) one can show that this estimator is consistent.

Theoretical Results

- We now show that the estimator given previously can be better than just using the samples at the end.
- This will require quite substantial efforts as we will now describe.
- There are two main points to note.

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

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

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

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

$$\begin{aligned} \sup_{1 \leq l \leq L} \sup_{u \in E} G_l(u) &\leq \overline{C} \\ \inf_{1 \leq l \leq L} \inf_{u \in E} G_l(u) &\geq \underline{C}. \end{aligned}$$

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

$$\int_A M_p(u, du') \geq \rho \int_A M_p(v, dv').$$

Main result

Theorem (BJLTZ15)

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

$$\mathbb{E}\left[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2\right] = \frac{V}{2}$$

$$\lesssim \frac{1}{N_0} + \sum_{l=1}^{L}\left(\frac{h_l^\beta}{N_l} + \left(\frac{h_l^\beta}{N_l}\right)^{1/2}\sum_{q=l+1}^{L}\frac{h_q^{\beta/2}}{N_q}\right).$$

In particular, for $\beta > \zeta$, L and $\{N_l\}$ can be chosen such that MSE$= \mathcal{O}(\varepsilon^2)$ for computational cost$= \mathcal{O}(\varepsilon^{-2})$.

- The proof is particularly challenging.
- It requires one to understand the auto-covariance of SMC approximations.
- In particular one must prove results of the form, for $n > p \geq 0$:

$$|\mathbb{E}[(\eta_p^{N_p} - \eta_p](\varphi_p))(\eta_n^{N_n} - \eta_n](\varphi_n))]| \leq \frac{C\|\varphi_p\|_\infty \|\varphi_n\|_\infty \kappa^{n-p}}{N_p}$$

where $\kappa \in (0,1)$, C do not depend on n, p.

- The strategy is to consider the 'magic' decomposition of Del Moral et al. (2012):

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

where $D_{p,n}$ are some semi-groups $V_p^{N_p}$ are local errors which become asymptotically independent w.r.t time and $R_{p+1}^{N_p}$ are some remainders.

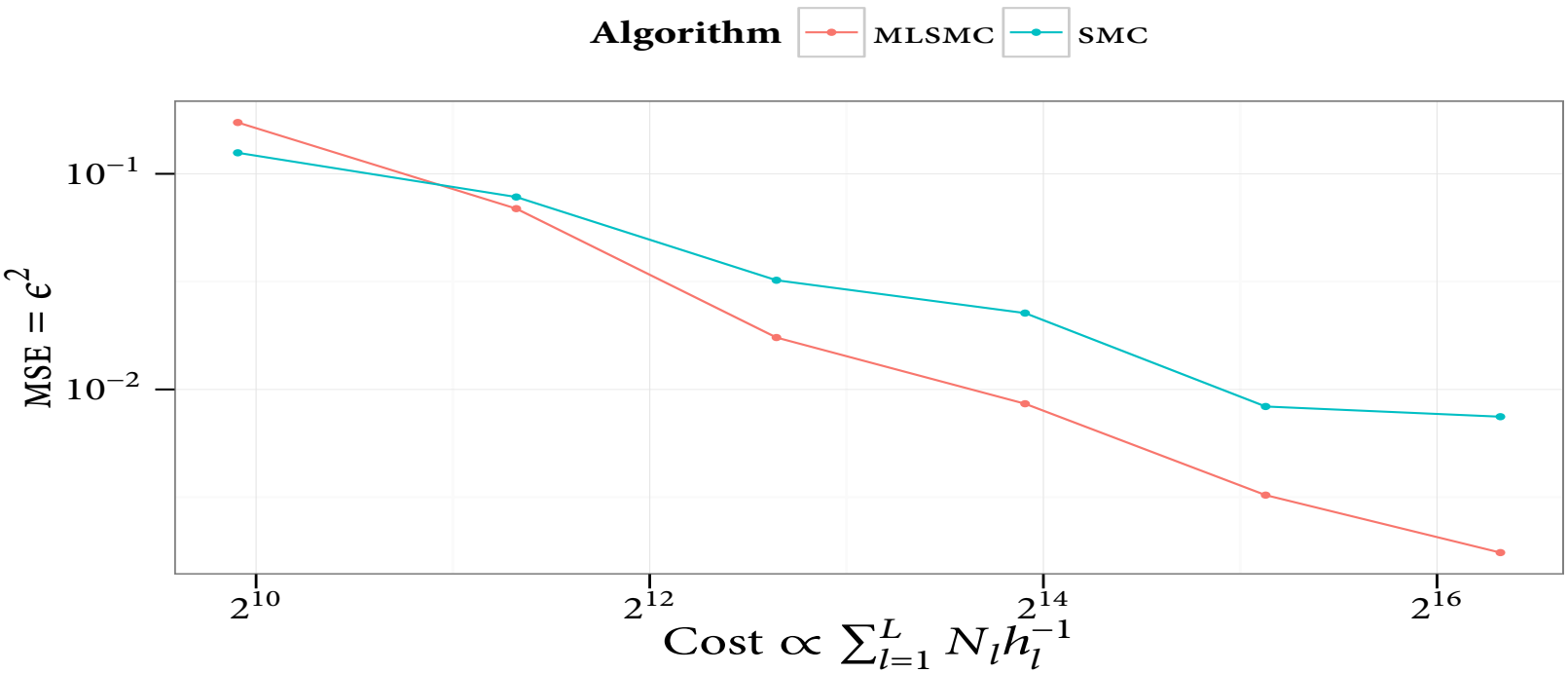
- The key point is then that the auto-covariance of the remainders decays much faster than the those of the local errors.
- As the local errors are asymptotically independent, one has to characterise how fast this happens w.r.t. time - it is geometric under our assumptions.

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 $\phi_2(x) = \cos(2\pi x)$.

- The forward problem at level l is solved using piecewise linear shape functions on a uniform mesh with mesh width $h_l = 2^{-(l+k)}$, with $k = 3$.
- 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 a function of runtime



Summary

- We have developed an SMC approach to implement the ML identity.
- Alternative procedures possible.
- Applicability beyond Bayesian inverse problems.

- Multilevel Sequential Monte Carlo sampler (MLSMC) can perform asymptotically as well as MLMC.
- Cost-to-ε asymptotically the same as for a scalar random variable!
- Results assume $\beta > \zeta$. If $\beta \leq \zeta$, cost is somewhat higher, analogous to standard MLMC.
- If $\zeta > 2\alpha$ then the optimal cost is $\varepsilon^{-\zeta/\alpha}$, the cost of a single simulation at the finest level.
- Many extensions under investigation.

- Joint work with: Kody Law (KAUST), Yan Zhou (NUS), Alex Beskos (UCL) & Raul Tempone (KAUST).
- Paper: Beskos, A., Jasra A., Law, K . J. H., Tempone, R. & Zhou, Y. (2015). Multi-Level SMC samplers. arXiv preprint.

References

- **[BJLTZ15]**: Beskos, Jasra, Law, Tempone, Zhou. "Multilevel SMC Samplers." arXiv (2015).
- **[G08]**: Giles. "Multilevel Monte Carlo." Op. Res., 56, 607-617 (2008).
- **[DDJ06]**: Del Moral, Doucet, Jasra. "SMC samplers." J. R. Statist. Soc. B, 68, 411-436 (2006).
- **[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).