Sequential Monte Carlo Samplers

Pierre Del Moral (delmoral@cict.fr) L.S.P. Toulouse

Arnaud Doucet (ad2@eng.cam.ac.uk) Cambridge University

Gareth Peters (gwp20@eng.cam.ac.uk) Cambridge University

Organization of the Talk

1. Motivation and Objectives of new methodology : SMC Samplers.

2. Brief Review of Standard SMC Methods.

3. SMC Samplers.

4. Trans-dimensional SMC Samplers.

5. Examples.

Motivation and Objectives of SMC Samplers

• Let $\{\pi_n\}_{n\geq 1}$ be a sequence of probability distributions defined on E such that $\pi_n(dx) = \pi_n(x)\,dx$ and each $\pi_n(x)$ is known up to a normalizing constant, i.e.

$$\pi_n(x) = \underbrace{Z_n^{-1}}_{\text{unknown}} \cdot \underbrace{f_n(x)}_{\text{known}}.$$

• Estimate expectations $\int \varphi(x)\,\pi_n(dx)$ and/or normalizing constants Z_n

sequentially; i.e. first π_1 then π_2 and so on.

• Objectives: Obtain SMC algorithms to do the job instead of MCMC.

$\Rightarrow$ Well-suited to parallel computers, well adapted to multimodal problems, no burn in.

$\Rightarrow$ Complementary approach to MCMC.

Examples relevant to SMC Samplers Framework

- Sequential Bayesian Inference: $\pi_n(x) = p(x|\, y_{1:n})$.

- Global optimization: $\pi_n(x) \propto [\pi(x)]^{\gamma_n}$ with $\{\gamma_n\}$ increasing sequence such that $\gamma_n \to \infty$.

- Sampling from a fixed target π : $\pi_n(x) \propto [\mu_1(x)]^{\gamma_n} [\pi(x)]^{1-\gamma_n}$ where μ_1 easy to sample and $\gamma_1 = 1$, $\gamma_n < \gamma_{n-1}$ and $\gamma_P = 0$.

- Rare event simulation $\pi(A) \ll 1$: $\pi_n(x) \propto \pi(x) 1_{A_n}(x)$ with $A_1 = E$, $A_n \subset A_{n-1}$ and $A_P = A$.

Brief Review of Standard SMC Methods

- $\{E_n\}_{n\geq 1}$ is such that $\dim(E_n) < \dim(E_{n+1})$; e.g. $E_1 = E$ and $E_n = E^n$.

- Let $\{\widetilde{\pi}_n\}_{n\geq 1}$ be a sequence of probability distributions defined on $\{E_n\}_{n\geq 1}$ such that $\widetilde{\pi}_n(dx_{1:n}) = \widetilde{\pi}_n(x_{1:n})\,dx_{1:n}$ and each $\widetilde{\pi}_n(x_{1:n})$ is known up to a normalizing constant, i.e.

$$\widetilde{\pi}_n(x_{1:n}) = \underbrace{Z_n^{-1}}_{\text{unknown}} \cdot \underbrace{f_n(x_{1:n})}_{\text{known}} \text{ where } x_{1:n} \triangleq (x_1, x_2, \ldots, x_n).$$

- Estimate expectations

$$\int \varphi_n(x_{1:n})\,\widetilde{\pi}_n(dx_{1:n}) \text{ and/or normalizing constant } Z_n.$$

- Sequential method: first sample from $\widetilde{\pi}_1$ then $\widetilde{\pi}_2$ and so on.

• SMC Principle: Approximate $\{\widetilde{\pi}_n\}$ by weighted sum of random samples/particles $\left\{X_{1:n}^{(i)}, W_n^{(i)}\right\}$ $\left(W_n^{(i)} > 0,\ \sum_{i=1}^N W_n^{(i)} = 1\right)$; i.e.

$$\widehat{\widetilde{\pi}}_n\left(dx_{1:n}\right) = \sum_{i=1}^N W_n^{(i)} \delta_{X_{1:n}^{(i)}}\left(dx_{1:n}\right) \text{ and } \widehat{\widetilde{\pi}}_n \to \widetilde{\pi}_n \text{ as } N \to \infty$$

• Key elements of SMC methods:

$\Rightarrow$ Sequential Importance Sampling

$\Rightarrow$ If variance of the weights high then Resample.

$\Rightarrow$ Algorithm of complexity $O\left(N\right)$

• Remark: Memory requirements in $O\left(N\right)$ and does not increase over time if only interested in approximating the marginal $\widetilde{\pi}_n\left(x_n\right)$

Sequential Importance Sampling Resampling

- If $X_{1:n-1}^{(i)} \sim \mu_{n-1}$ and target is $\widetilde{\pi}_{n-1}$ then

$$W_{n-1}^{(i)} \propto \frac{\widetilde{\pi}_{n-1}\left(X_{1:n-1}^{(i)}\right)}{\mu_{n-1}\left(X_{1:n-1}^{(i)}\right)}, \quad \sum_{i=1}^{N} W_{n-1}^{(i)} = 1.$$

- If $X_n^{(i)} \Big| X_{n-1}^{(i)} \sim K_n\left(X_{n-1}^{(i)}, \cdot\right)$ and target is $\widetilde{\pi}_n$ then

$$\begin{aligned} W_n^{(i)} &\propto \frac{\widetilde{\pi}_n\left(X_{1:n-1}^{(i)}\right)}{\mu_{n-1}\left(X_{1:n-1}^{(i)}\right) K_n\left(X_{n-1}^{(i)}, X_n^{(i)}\right)} \\ &= W_{n-1}^{(i)} \frac{\widetilde{\pi}_n\left(X_{1:n}^{(i)}\right)}{\widetilde{\pi}_{n-1}\left(X_{1:n-1}^{(i)}\right) K_n\left(X_{n-1}^{(i)}, X_n^{(i)}\right)} \end{aligned}$$

New : Sequential Monte Carlo Samplers

- Standard SMC Methods do NOT apply to our problem; i.e.

one needs $E_{n-1} \subset E_n$ whereas we have $E_n = E$.

- "Idea": Consider a new sequence of artificial distributions $\{\widetilde{\pi}_n\}_{n \geq 1}$ defined on $E_n = E^n$ such that

$$\int \widetilde{\pi}_n \left(x_{1:n-1}, x_n\right) dx_{1:n-1} = \pi_n \left(x_n\right)$$

and apply standard SMC.

- Example:

$$\widetilde{\pi}_n \left(x_{1:n}\right) = \pi_n \left(x_n\right) \prod_{k=1}^{n-1} L_k \left(x_{k+1}, x_k\right).$$

where $\{L_k\}$ is arbitrary sequence of Markov kernels.

SMC Samplers

At time n; $n \geq 1$.

Sampling step. For $i = 1, ..., N$, sample $X_n^{(i)} \sim K_n\left(X_{n-1}^{(i)}, \cdot\right)$ and set

$$W_n^{(i)} \propto W_{n-1}^{(i)} \frac{\pi_n\left(X_n^{(i)}\right) L_{n-1}\left(X_n^{(i)}, X_{n-1}^{(i)}\right)}{\pi_{n-1}\left(X_{n-1}^{(i)}\right) K_n\left(X_{n-1}^{(i)}, X_n^{(i)}\right)}.$$

Resampling step. If ESS<Threshold then resample particles $\left\{W_n^{(i)}, X_n^{(i)}\right\}$ to obtain N new particles $\left\{N^{-1}, X_n^{(i)}\right\}$.

- Monte Carlo approximation

$$\widehat{\pi}_n\left(dx_n\right) = \sum_{i=1}^{N} W_n^{(i)} \delta_{\widetilde{X}_n^{(i)}}\left(dx_n\right).$$

- Ratio of normalizing constants

$$\widehat{\frac{Z_n}{Z_{n-1}}} = \sum_{i=1}^{N} W_{n-1}^{(i)} \widetilde{w}_n\left(X_{n-1:n}^{(i)}\right)$$

where

$$\widetilde{w}_n\left(x_{n-1:n}\right) = \frac{f_n\left(x_n\right) L_{n-1}\left(x_n, x_{n-1}\right)}{f_{n-1}\left(x_{n-1}\right) K_n\left(x_{n-1}, x_n\right)}.$$

Algorithm Settings - Selection of artificial kernels

• (Too) many degrees of freedom! Typically $\{\pi_n\}$ given by problems but what about $\{K_n\}$ and the artificial sequence $\{L_n\}$???

• To understand how to select $\{L_n\}$, the key is to remember

$$X_{1:n}^{(i)} \sim \mu_n(\cdot) \text{ where } \mu_n(x_{1:n}) = \mu_1(x_1) \prod_{k=2}^{n} K_k(x_{k-1}, x_k)$$

• Discrepancy between $\pi_n(x_n)$ and $\mu_n(x_n) \Rightarrow$ IS correction

$$w_n(x_n) = \frac{\pi_n(x_n)}{\mu_n(x_n)}$$

$\Rightarrow$ Impossible to implement as $\mu_n(x_n)$ admits no closed-form expression.

$$\mu_n(x_n) = \begin{cases} \mu_1 K_{2:n}(x_n) & \text{if no resampling before } n \\ \pi_s K_{s:n}(x_n) & \text{last resampled at } s \end{cases}$$

- Introduction of kernels $\{L_n\}$

$$w_n(x_{1:n}) = \frac{\pi_n(x_n)\prod_{k=1}^{n-1} L_k(x_{k+1}, x_k)}{\mu_1(x_1)\prod_{k=2}^{n} K_k(x_{k-1}, x_k)} = w_{n-1}(x_{1:n-1})\frac{\pi_n(x_n)}{\pi_{n-1}(x_n)}\frac{L_{n-1}(x_n, x_{n-1})}{K_n(x_{n-1}, x_n)}.$$

$\Rightarrow$ Possible to compute consistent IS weight without having to compute $\mu_n(x_n)$.

$\Rightarrow$ No free lunch theorem: price to pay is variance increase as now weights are evaluated on E^n as opposed to E.

- Optimal sequence of kernels $\left\{L_n^{\mathsf{opt}}\right\}$ brings you back to IS on E

$$L_{k-1}^{\mathsf{opt}}(x_k, x_{k-1}) = \frac{\mu_{k-1}(x_{k-1})K_k(x_{k-1}, x_k)}{\mu_k(x_k)}.$$

Remark. It follows straightforwardly from forward-backward formula

$$\mu_1(x_1)\prod_{k=2}^{n} K_k(x_{k-1}, x_k) = \mu_n(x_n)\prod_{k=2}^{n} L_{k-1}^{\mathsf{opt}}(x_k, x_{k-1}).$$

Selection of Artificial Kernels

• Hard to use $\left\{L_{k-1}^{\text{opt}}\right\}$ as do not know $\mu_n(x_n)$ but can be approximated.

$\Rightarrow$ Suboptimal but asymptoticaly consistent strategy consists of using the approximation of L_{n-1}^{opt} given below

$$L_{n-1}(x_n, x_{n-1}) = \frac{\pi_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n)}{\pi_{n-1} K_n(x_n)} \Rightarrow w_n(x_{n-1:n}) = \frac{\pi_n(x_n)}{\pi_{n-1} K_n(x_n)}$$

• If K_n is an MCMC kernel of invariant distribution π_n and $\pi_{n-1} \simeq \pi_n$ then a convenient choice is

$$L_{n-1}(x_n, x_{n-1}) = \frac{\pi_n(x_{n-1}) K_n(x_{n-1}, x_n)}{\pi_n(x_n)} \Rightarrow w_n(x_{n-1:n}) = \frac{\pi_n(x_{n-1})}{\pi_{n-1}(x_{n-1})}.$$

$\Rightarrow$ this choice of L_n kernel is a good approximation of L_{n-1}^{opt} when $\pi_{n-1} \approx \pi_n$

• Remark. Aestethic but generally bad idea: $L_{n-1} = K_n$.

Selection of initial distribution and transition kernels

• μ_1 should be selected so that it is easy to sample and IS weights $\frac{\pi_1}{\mu_1}$ are upper bounded over E and almost uniform

• If $\pi_{n-1} \simeq \pi_n$, can select K_n as an MCMC kernel of invariant distribution π_n.

• If π_{n-1} and π_n vary significantly (e.g. sequential Bayes), possibility to use artificial dynamics.

• To sample from high dimensional distributions, a SMC sampler can update the components of x via subblocks.

Convergence Results

• Convergence results follow from general results on Feynman-Kac formula (see Del Moral, 2004).

Trans-dimensional SMC Samplers

• Consider case : $E = \cup_{k=0}^{\infty} \{k\} \times \vartheta_k$; i.e. traditionaly where Trans-dimensional MCMC are used in a batch setting (RJMCMC : Green, 1995).

• Now using the SMC sampler methodology, TD-SMC may also be used for this batch analysis situation.

• In this case, a mixture of moves is useful

$$K_n\left(x, x'\right) = \sum_{m=1}^{M} \alpha_{n,m}\left(x\right) K_{n,m}\left(x, x'\right)$$

where $\alpha_{n,m}\left(x\right) > 0$, $\sum_{m=1}^{M} \alpha_{n,m}\left(x\right) = 1$ and $\{K_{n,m}\}$ is a collection of transition kernels.

• Incremental weight can be computed by standard formula but too expensive if M is large.

- Alternative importance sampling on joint space

$$\frac{\pi_n\left(X_n^{(i)}\right)\beta_{n-1,M_n^{(i)}}\left(X_n^{(i)}\right)L_{n-1,M_n^{(i)}}\left(X_n^{(i)},X_{n-1}^{(i)}\right)}{\pi_{n-1}\left(X_{n-1}^{(i)}\right)\alpha_{n,M_n^{(i)}}\left(X_{n-1}^{(i)}\right)K_{n,M_n^{(i)}}\left(X_{n-1}^{(i)},X_n^{(i)}\right)}$$

where $\Pr\left(M_n^{(i)}=m\right)=\alpha_{n,m}\left(X_{n-1}^{(i)}\right)$ and $X_n^{(i)}\sim K_{n,M_n^{(i)}}\left(X_{n-1}^{(i)},\cdot\right)$.

- $\beta_{n-1,m}(x)>0$, $\sum_{m=1}^{M}\beta_{n-1,m}(x)=1$ and $\left\{L_{n-1,m}\right\}$ artificial transition kernels.

- Optimal choice follows straightforwardly

$$\beta_{n-1,m}^{opt}\left(x'\right)L_{n-1,m}^{opt}\left(x',x\right)=\frac{\pi_{n-1}(x)\,\alpha_{n,m}(x)\,K_{n,m}(x,x')}{\sum_{m=1}^{M}\int\pi_{n-1}(x)\,\alpha_{n,m}(x)\,K_{n,m}(x,x')\,dx}$$

• In this case, birth/death, split/merge moves can be developed

$\Rightarrow$ No reversibility constraints.

• Standard SMC framework can be re-used here to design optimal moves; e.g. birth step corresponds to increase dimension.

• Numerous applications to sequential Bayesian estimation; e.g. "on line" sequential analysis of model selection and parameter/state estimation, batch analysis typically associated with RJMCMC .

SMC Samplers Examples : Bayesian variable selection

- Model

$$Y = \sum_{k=1}^{M} I_k \beta_k \Psi_k (X) + V; \, V \sim \mathcal{N}\left(0, \sigma^2\right).$$

where $I_k \in \{0,1\}$ is such that $\beta_k = 0$ if $I_k = 0$ and $\beta_k \neq 0$ if $I_k = 1$; i.e. 2^M different models for the regression function.

- Standard conjugate priors leads to marginal posterior distribution

$$\begin{aligned} p\left(i_{1:M} | \, x_{1:T}, y_{1:T}\right) \; \propto \; & \left(\nu_0 + y_{1:T}^{\mathsf{T}} P\left(i_{1:M}\right) y_{1:T}\right)^{T/2+\frac{\gamma_0}{2}} \\ & \times \left(1+\delta^2\right)^{-l(i_{1:M})/2} l\left(i_{1:M}\right)! \left(T - l\left(i_{1:M}\right)\right)! \end{aligned}$$

where

$$P\left(i_{1:M}\right) = I_{l(i_{1:M})} - \left(1+\delta^{-2}\right)^{-1} D\left(i_{1:M}\right) \left(D^{\mathsf{T}}\left(i_{1:M}\right) D\left(i_{1:M}\right)\right)^{-1} D^{\mathsf{T}}\left(i_{1:M}\right).$$

Algorithm Settings

- To sample/maximize $p\left(i_{1:M}|\, y_{1:T}, x_{1:T}\right)$, consider $\pi_n\left(i_{1:M}\right) \propto \left[p\left(i_{1:M}|\, x_{1:T}, y_{1:T}\right)\right]^{\gamma_n}$

where $n \in \{1, \ldots, P\}$, $M = 50$, $P \in \{250, 500, 1250, 2500, 5000\}$ and $\gamma_1 = 0$.

$\Rightarrow$ Case 1 : To sample from π, $\gamma_n = a \log(n) + b$ with $\gamma_P = 1$.

$\Rightarrow$ Case 2 : To maximize π, $\gamma_n = a \log(n) + b$ with $\gamma_P = 10$.

- K_n deterministic scan Gibbs sampler invariant distribution π_n (1 site per n)
- L_n kernels : approximation of L_{n-1}^{opt} and the Annealed Importance Sampling choice of (Neal,2001)

$L_{n-1}\left(i_{1:M}^{(n)}, i_{1:M}^{(n-1)}\right)$	A $\frac{\pi_{n-1}\left(i_{1:M}^{(n-1)}\right) K_n\left(i_{1:M}^{(n-1)}, i_{1:M}^{(n)}\right)}{\pi_{n-1} K_n\left(i_{1:M}^{(n)}\right)}$	B $\frac{\pi_n\left(i_{1:M}^{(n-1)}\right) K_n\left(i_{1:M}^{(n-1)}, i_{1:M}^{(n)}\right)}{\pi_n\left(i_{1:M}^{(n)}\right)}$
$w_n\left(i_{1:M}^{(n-1)}, i_{1:M}^{(n)}\right)$	$\frac{\pi_n\left(i_{1:M}^{(n-1)}\right) + \pi_n\left(i_{1:M}^{(n)}\right)}{\pi_{n-1}\left(i_{1:M}^{(n-1)}\right) + \pi_{n-1}\left(i_{1:M}^{(n)}\right)}$	$\frac{\pi_n\left(i_{1:M}^{(n-1)}\right)}{\pi_{n-1}\left(i_{1:M}^{(n-1)}\right)}$

Simulation Results : Sampling $\pi_n(i_{1:M}) \propto [p(i_{1:M}|\, x_{1:T}, y_{1:T})]^{\gamma_n}$

• $y_{1:T}$ noisy samples from a sinc function, basis functions $\Psi_k(x)$ were Gaussians.

Case 1 : (50 simulations) sample from π

• Compared SMC samplers using (A), Annealed Importance Sampling using (A) and (B) and an MCMC Gibbs sampler of chain length PN.

• Remark : AIS does not have a resampling step.

• Results demonstrated in all simulations that resampling utilised by SMC samplers produced a reduction in the MSE's variance, cheaply.

Case 2 : (50 simulations) maximize π

• Compared SMC samplers using (A), Simulated Annealing and N (non interacting) Parallel Annealing runs.

• SMC Samplers out performed both S.A. and Parallel Annealing with respect to maximum posterior mode estimate obtained in each simulation.

• This further emphasises gains that can be made through introduction of resampling which allows the Simulated Annealing chains to interact in a principled manner.

• SMC Samplers performed significantly better than the compared algorithms when P is small.

• Intuitively when P is small there are less intermediate distributions hence π_{s-1} and π_s can be significantly different.

• As expected the number of resampling steps carried out, decreased as P increased

SMC Samplers Examples : Sequential Bayesian Estimation

- At time t, time occurrences assumed to follow an inhomogeneous Poisson process of intensity $\lambda : R^+ \to R^+$

$$p_t\left(y_{1:l_t} \middle| \{\lambda(u)\}_{u \leq t}\right) = \exp\left(-\int_0^t \lambda(u)\, du\right) \prod_{l=1}^{l_t} \lambda(y_l)\,.$$

- We want to estimate unknown intensity $\lambda(t)$ sequentially in time.
- Simple piecewise constant model for $\lambda(t)$

$$\lambda(t) = \sum_{m=1}^{k} \lambda_m 1_{[\tau_{m-1},\tau_m)}(t) + \lambda_{k+1} 1_{[\tau_m,\infty)}(t)\,.$$

- The number of steps k, their amplitudes $\lambda_{1:k+1}$ and the knot points $\tau_{1:k}$ are assumed unknown

$\Rightarrow$ Set following time-dependent prior distribution

$$p_t\left(k, \lambda_{1:k+1}, \tau_{1:k}\right) = p_t\left(k\right) p\left(\lambda_{1:k+1} \middle| k\right) p_t\left(\tau_{1:k} \middle| k\right)$$

where $p_t\left(k\right)$ Poisson $\lambda_q t$, $p_t\left(\tau_{1:k} \middle| k\right)$ uniform order statistics on $[0, t]$ and $\lambda_1 \sim G\left(\alpha, \beta\right)$ and $\lambda_l \middle| \lambda_{l-1} \sim G\left(\lambda_{l-1}^2/\chi; \lambda_{l-1}/\chi\right)$.

- Sequential estimation of posterior distributions over times $n\Delta T$

$$\pi_n\left(k, \lambda_{1:k+1}, \tau_{1:k}\right) = p_{n\Delta T}\left(k, \lambda_{1:k+1}, \tau_{1:k} \middle| y_{1:l_{n\Delta T}}\right)$$

where ΔT is a time interval defined by the user.

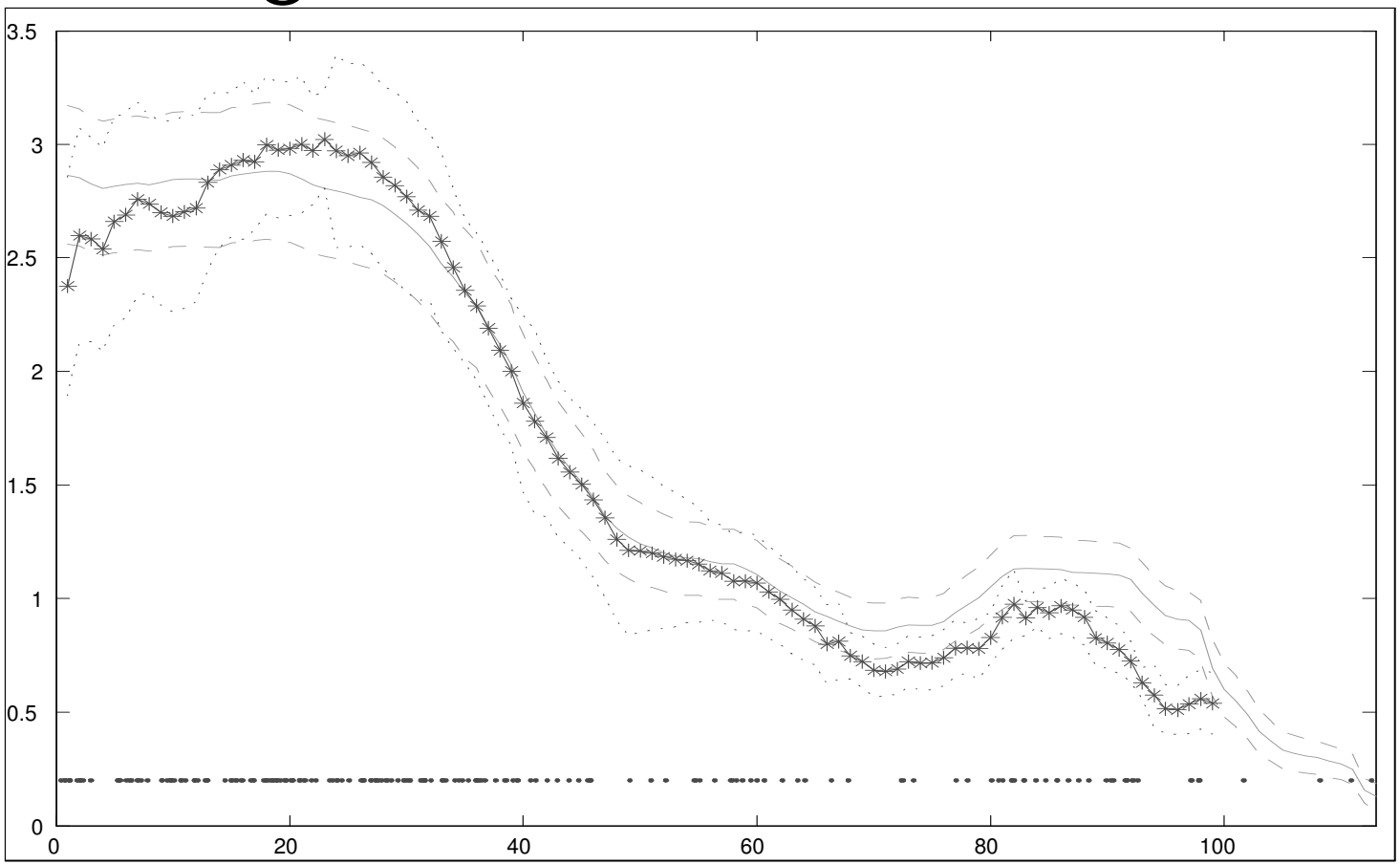
- These distributions are defined on $E = \cup_{k=0}^{\infty} \{k\} \times \vartheta_k$ where $\vartheta_k = \left\{\tau_{1:k} \in \left(R^+\right)^k; 0 < \tau_1 < \ldots < \tau_k\right\} \times \left(R^+\right)^{k+1}$, the support of π_n being reduced to $\left\{\tau_{1:k} \in \left(R^+\right)^k; 0 < \tau_1 < \ldots < \tau_k < n\Delta T\right\} \times \left(R^+\right)^{k+1}$.

Simulation Results

• TD-SMC with Birth, Death, Update and Adjustment moves was applied to the Coal Mining Data set (Green, 1995).

• Comparison between RJMCMC and TD-SMC was obtained. The plot of the estimated intensity of the inhomogeneous Poisson process is provided next.

Coal mining data: TD-SMC vs RJMCMC



Coal Mine disaster data, 1851-1962 : Solid line : RJMCMC estimate, Dashed lines : RJMCMC estimate ± 3 std., Star : TD-SMC estimate $E\left[\lambda\left(t\right)|y_{1:l_{n\Delta T+14\Delta T}}\right]$, Dotted lines : TDSMC estimate ± 3 std

Conclusion and Discussion

- SMC samplers provides principled set of new algorithms.

$\Rightarrow$ Interacting parallel MCMC runs (EASY TO CODE); allows to correct for burn-in through importance sampling: complementary to MCMC.

$\Rightarrow$ Global optimization, Sequential Bayesian estimation, Rare event simulation.

- Many methodological problems remain open

$\Rightarrow$ Optimization N versus P for fixed computational complexity.

$\Rightarrow$ Optimal path from π_1 to $\pi_P = \pi$.

References

• P. Del Moral, "Feynman-Kac Formulae : Genealogical and Interacting Particle Systems with Applications", Springer, 2004.

• P. Del Moral and A. Doucet, "On a Class of Genealogical and Interacting Metropolis Models", Séminaire de Probabilités XXXVII, Ed. J. Azéma,

• M. Emery, M. Ledoux & M. Yor, Lecture Notes in Mathematics, Berlin: Springer-Verlag, 2003.

• P. Del Moral, A. Doucet and G.W. Peters, "Sequential Monte Carlo Samplers", in revision J. Royal Stat. Soc. B.

• A. Doucet, N. de Freitas, and N.J. Gordon (editors), Sequential Monte Carlo Methods in Practice, New York: Springer-Verlag, 2001.

Extra Comments : Convergence Results

- Convergence results follow from general results on Feynman-Kac formula (see Del Moral, 2004).
- When no resampling is performed, one has

$$\sqrt{N}\left(\widehat{E}_{\pi_n}(\varphi) - E_{\pi_n}(\varphi)\right) \Rightarrow \mathcal{N}\left(0, \int \frac{\widetilde{\pi}_n^2(x_{1:n})}{\mu_n(x_{1:n})}\left(\varphi(x_n) - E_{\pi_n}(\varphi)\right)^2 dx_{1:n}\right)$$

- When multinomial resampling is used at each iteration, one has

$$\sqrt{N}\left(\widehat{E}_{\pi_n}(\varphi) - E_{\pi_n}(\varphi)\right) \Rightarrow \mathcal{N}\left(0, \sigma^2_{SMC,n}(\varphi)\right),$$

$$\begin{aligned}
\sigma^2_{SMC,n}(\varphi) &= \int \frac{\widetilde{\pi}_n^2(x_1)}{\mu_1(x_1)}\left(\int \varphi(x_n)\,\widetilde{\pi}_n(x_n|\,x_1)\,dx_n - E_{\pi_n}(\varphi)\right)^2 dx_1 \\
&+ \sum_{k=2}^{n-1} \int \frac{\left(\widetilde{\pi}_n(x_k) L_{k-1}(x_k, x_{k-1})\right)^2}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)}\left(\int \varphi(x_n)\,\widetilde{\pi}_n(x_n|\,x_k)\,dx_n - E_{\pi_n}(\varphi)\right)^2 dx_{k-1:k} \\
&+ \int \frac{\left(\pi_n(x_n) L_{n-1}(x_n, x_{n-1})\right)^2}{\pi_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n)}\left(\varphi(x_n) - E_{\pi_n}(\varphi)\right)^2 dx_{n-1:n}.
\end{aligned}$$

• The variance expressions obtained for the CLT demonstrate the role of the auxiliary kernels $\left\{L_n^{\text{opt}}\right\}_{n\in\mathbb{N}}$.

• When multinomial resampling is used at each stage, the variance expression demonstrates that the faster the L kernels are mixing, then the faster $\widetilde{\pi}_n\left(x_n|\, x_k\right) = \frac{\int \widetilde{\pi}_n(x_{1:n})dx_{1:k-1}dx_{k+1:n}}{\widetilde{\pi}_n(x_k)}$ converges to $\pi_n\left(x_n\right)$

• So ultimately when L kernels are mixing quickly, then the most significant term in the expression for the asymptotic variance is just the term at time n given by

$$\int \frac{\left(\pi_n\left(x_n\right) L_{n-1}\left(x_n, x_{n-1}\right)\right)^2}{\pi_{n-1}\left(x_{n-1}\right) K_n\left(x_{n-1}, x_n\right)} \left(\varphi\left(x_n\right) - E_{\pi_n}\left(\varphi\right)\right)^2 dx_{n-1:n}$$

Extra Comments : Example Bayesian Variable Selection

- Regression model

$$Y = \sum_{k=1}^{M} I_k \beta_k \Psi_k (X) + V;\ V \sim \mathcal{N}\left(0, \sigma^2\right).$$

- Indicator variable $I_k \in \{0, 1\}$ is such that $\beta_k = 0$ if $I_k = 0$ and $\beta_k \neq 0$ if $I_k = 1$
- Assuming T independent identically distributed data $(X_{1:T}, Y_{1:T})$ are available

$$Y_{1:T} = D(I_{1:M})\, \beta(I_{1:M}) + V_{1:T},$$

- $D(I_{1:M})$ is a $T \times l(I_{1:M})$ matrix where $l(I_{1:M}) = \sum_{k=1}^{M} I_k$ is the number of basis terms included in the model.
- The j^{th} column of $D(I_{1:M})$ corresponds to $\left(\Psi_{\alpha(I_{1:M},j)}(X_1), \ldots, \Psi_{\alpha(I_{1:M},j)}(X_T)\right)^{\mathsf{T}}$ where $\alpha(I_{1:M}, j)$ is the index of the j^{th} non-null coefficient of the sequence $I_{1:M}$.

- $\beta_{I_{1:M}}$ is the associated $l\left(I_{1:M}\right) \times 1$ vector of non-null regression coefficients.

$$\begin{aligned} \beta\left(I_{1:M}\right) | \left(\sigma^2, I_{1:M}\right) &\sim \mathcal{N}\left(0, \delta^2 \sigma^2 \left(D^{\mathsf{T}}\left(I_{1:M}\right) D\left(I_{1:M}\right)\right)^{-1}\right), \\ \sigma^2 &\sim \mathcal{IG}\left(\frac{\gamma_0}{2}, \frac{\nu_0}{2}\right), \end{aligned}$$

and $\mathsf{Pr}\left(I_k = 1 | \lambda\right) = \lambda$ where λ is uniformly distributed on $[0, 1]$.

- Finally γ_0, ν_0 and δ are fixed hyperparameters.

- The data are taken to be the sinc function, i.e. $\mathsf{sinc}(x) = \mathsf{sin}\left(x\right)/x$, corrupted by additive Gaussian noise with $\sigma = 0.1$ for $T = 50$ evenly spaced points in the interval $[-10, 10]$.

- We select $M = T$ basis functions of the form

$$\Psi_k\left(x\right) = \frac{1}{\sqrt{2\pi}\phi} \mathsf{exp}\left(-\frac{\left(x - x_k\right)^2}{2\phi^2}\right)$$

where $\phi = 1.6$.

- To sample from $p\left(i_{1:M} | \, y_{1:T}, x_{1:T}\right)$, we consider the sequence of distributions

$$\pi_n\left(i_{1:M}\right) \propto \left[p\left(i_{1:M} | \, x_{1:T}, y_{1:T}\right)\right]^{\gamma_n}$$

- $n \in \{1, \ldots, p\}$, $p = 10000$ and $\gamma_1 = 0$ (i.e. $\pi_1 = \mu_1$ is the uniform distribution)
- γ_n increases according to $a \log(n) + b$ with $\gamma_p = 1$.
- To find the modes of $p\left(i_{1:M} | \, x_{1:T}, y_{1:T}\right)$, we consider $n \in \{1, \ldots, p\}$, $p = 50000$
- In this case, we use $a \log(n) + b$ such that $\gamma_1 = 0$ and $\gamma_p = 10$.
- Select K_n as a deterministic scan Gibbs sampler of invariant distribution π_n.
- Only one variable is updated at iteration n, hence $p \gg 1$.
- $N = 1000$ particles
- Resampling is performed when the ESS is below $N/2$.
- 50 simulations using the same dataset.

Extra Comments : SMC Estimates

- Particle approximation after the mutation stage $\pi_n K_n$ is $\left\{W_{n-1}^{(i)}, X_{1:n}^{(i)}\right\}$
- Ratio of normalizing constants

$$\frac{Z_n}{Z_{n-1}} = \frac{\int f_n\left(x_{1:n}\right) dx_{1:n}}{\int f_{n-1}\left(x_{1:n-1}\right) dx_{1:n-1}}$$

- note

$$\begin{aligned} \frac{\pi_n}{\pi_{n-1} K_n} &= \frac{Z_n^{-1} f_n}{Z_{n-1}^{-1} f_{n-1} K_n} \\ \therefore \quad \frac{Z_n \pi_n}{Z_{n-1} \pi_{n-1} K_n} &= \frac{f_n}{f_{n-1} K_n} = \widetilde{w}_n \end{aligned}$$

$$
\begin{aligned}
\therefore \quad \frac{Z_n}{Z_{n-1}} &= \frac{\pi_{n-1} K_n \widetilde{w}_n}{\pi_n} \\
&= \frac{\sum_{i=1}^{N} W_{n-1}^{(i)} \widetilde{w}_n \delta_{X_{1:n}^{(i)}}}{\sum_{i=1}^{N} W_n^{(i)} \delta_{X_{1:n}^{(i)}}} \\
&= \frac{\sum_{i=1}^{N} W_{n-1}^{(i)} \widetilde{w}_n \delta_{X_{1:n}^{(i)}}}{\frac{\sum_{i=1}^{N} W_{n-1}^{(i)} \widetilde{w}_n \delta_{X_{1:n}^{(i)}}}{\sum_{j=1}^{N} W_{n-1}^{(j)} \widetilde{w}_n \delta_{X_{1:n}^{(j)}}}} \\
&= \sum_{i=1}^{N} W_{n-1}^{(i)} \widetilde{w}_n \left(X_{1:n}^{(i)} \right)
\end{aligned}
$$

$$\widehat{\frac{Z_n}{Z_{n-1}}} = \sum_{i=1}^{N} W_{n-1}^{(i)} \widetilde{w}_n \left(X_{1:n}^{(i)} \right)$$

where

$$\widetilde{w}_n \left(x_{n-1:n} \right) = \frac{f_n \left(x_n \right) L_{n-1} \left(x_n, x_{n-1} \right)}{f_{n-1} \left(x_{n-1} \right) K_n \left(x_{n-1}, x_n \right)}.$$

Extra Comments : $\widetilde{\pi}_n\left(x_{1:n-1}|x_n\right)$

• Choice of $\widetilde{\pi}_n\left(x_{1:n}\right) = \pi_n\left(x_n\right)\prod_{k=1}^{n-1} L_k\left(x_{k+1}, x_k\right)$ is useful as it makes evaluating the IS weights simpler :

$$
\begin{aligned}
w_n\left(x_{1:n}\right) &= \frac{\widetilde{\pi}_n\left(x_{1:n}\right)}{\widetilde{\pi}_{n-1}\left(x_{1:n}\right) K_n\left(x_{n-1}, x_n\right)} \\
&= \frac{\pi_n\left(x_n\right)\widetilde{\pi}_n\left(x_{1:n-1}|x_n\right)}{\pi_{n-1}\left(x_{n-1}\right)\widetilde{\pi}_{n-1}\left(x_{1:n-1}|x_n\right) K_n\left(x_{n-1}, x_n\right)} \\
&= \frac{\pi_n\left(x_n\right)\prod_{k=1}^{n-1} L_k\left(x_{k+1}, x_k\right)}{\pi_{n-1}\left(x_{n-1}\right)\prod_{k=1}^{n-2} L_k\left(x_{k+1}, x_k\right) K_n\left(x_{n-1}, x_n\right)} \\
&= \frac{\pi_n\left(x_n\right) L_{n-1}\left(x_{n+1}, x_n\right)}{\pi_{n-1}\left(x_{n-1}\right) K_n\left(x_{n-1}, x_n\right)}
\end{aligned}
$$

Extra Comments : Details of Possible $\{\pi_n\}$

• **(Neal, 2001)** Move smoothly from tractable distribution μ_1 via a sequence of intermediate distributions to π.

• eg. geometric path :

$$\pi_n(x) \propto [\mu_1(x)]^{\gamma_n} [\pi(x)]^{1-\gamma_n}$$

where μ_1 easy to sample and $\gamma_1 = 1$, $\gamma_n < \gamma_{n-1}$ and $\gamma_P = 0$.

• **Global Optimisation** : eg. Simulated Annealing

• Can select $\pi_n(x) \propto [\pi(x)]^{\gamma_n}$ with $\{\gamma_n\}$ increasing sequence such that $\gamma_n \to \infty$.

• Resulting algorithm is a genetic algorithm where the sampling step is a "mutation" and resampling step is a "selection" step.

• Significant difference to standard genetic algorithms is we control the asymptotic ($N \to \infty$) distribution of the particles.

• Standard selection/mutation algorithms are based on

$$X_n^{(i)} \sim K_n\left(X_{n-1}^{(i)}, .\right) \quad W_n^{(i)} \propto \pi_n\left(X_n^{(i)}\right)$$

• This corresponds to sampling from a joint distribution

$$\widetilde{\pi}_n\left(x_{1:n}\right) \propto \mu_1\left(x_1\right) \pi_1\left(x_1\right) \prod_{k=2}^{n} K_k\left(x_{k+1}, x_k\right) \pi_k\left(x_k\right)$$

• In general $\widetilde{\pi}_n\left(x_n\right) \neq \pi_n\left(x_n\right)$ and particles may not concentrate on a set of global maxima of π.

• **Rare Event Simuation**

• Want to estimate the probability of a very rare event, A, under the condition $\pi\left(A\right) << 1$.

$\Rightarrow$ Applications in finance and telecommunications.

• Most of these applications π is typically easy to sample and its normalizing constant is known.

• Sequence of distributions in this application will be as follows :

$$\pi_n(x) \propto \pi(x)\, \mathbb{I}_{A_n}(x)$$

with $N = \{1, ..., P\}$ and $A_1 = E \supseteq A_2 \supseteq ...A_{P-1} \supseteq A_P = A$.

• Can use the estimate for the ratio of normalizing constants

$$\widehat{\frac{Z_n}{Z_{n-1}}} = \sum_{i=1}^{N} W_{n-1}^{(i)} \widetilde{w}_n \left(X_{1:n}^{(i)} \right)$$

to estimate the rare event probability as follows :

$$\widehat{\pi}(A) = Z_1 \prod_{k=1}^{p-1} \widehat{\frac{Z_{k+1}}{Z_k}} \approx Z_p = \int_A \pi(x)\, dx$$

Extra Comments : Details of Possible $\{L_n\}$

- In standard applications of SMC only the proposal kernels $\{K_n\}$ need to be selected as the joint distributions $\{\widetilde{\pi}_n\}$ are given by the problem at hand.

- In the framework considered the$\{L_n\}$ kernels are completely arbitrary.

- In practice $\{L_n\}$ should be optimised with respect to $\{K_n\}$.

- To understand how to select $\{L_n\}$, the key is to remember

$$X_{1:n}^{(i)} \sim \mu_n(\cdot) \text{ where } \mu_n(x_{1:n}) = \mu_1(x_1) \prod_{k=2}^{n} K_k(x_{k-1}, x_k)$$

- Discrepancy between $\pi_n(x_n)$ and $\mu_n(x_n) \Rightarrow$ IS correction

$$w_n(x_n) = \frac{\pi_n(x_n)}{\mu_n(x_n)}$$

$\Rightarrow$ Impossible to implement as $\mu_n(x_n)$ admits no closed-form expression.

$$\mu_n(x_n) = \begin{cases} \mu_1 K_{2:n}(x_n) & \text{if no resampling before } n \\ \pi_s K_{s:n}(x_n) & \text{last resampled at } s \end{cases}$$

• An obvious Monte Carlo approximation of $\mu_n(x_n)$ is based on the current particles :

$$\widehat{\mu}_n(x_n) = \frac{1}{N} \sum_{i=1}^{N} K_n\left(X_{n-1}^{(i)}, x_n\right)$$

however then the complexity of the algorithm would be $O\left(N^2\right)$.

• Introduction of kernels $\{L_n\}$ allows us to perform Importance Sampling with

out having to compute the marginal distribution

$$w_n(x_{1:n}) = \frac{\pi_n(x_n) \prod_{k=1}^{n-1} L_k(x_{k+1}, x_k)}{\mu_1(x_1) \prod_{k=2}^{n} K_k(x_{k-1}, x_k)} = w_{n-1}(x_{1:n-1}) \frac{\pi_n(x_n)}{\pi_{n-1}(x_n)} \frac{L_{n-1}(x_n, x_{n-1})}{K_n(x_{n-1}, x_n)}.$$

$\Rightarrow$ Possible to compute consistent IS weight without having to compute $\mu_n(x_n)$.

$\Rightarrow$ No free lunch theorem: price to pay is variance increase as now weights are evaluated on E^n as opposed to E.

- Optimal sequence of kernels $\left\{L_n^{\text{opt}}\right\}$ brings you back to IS on E

$$L_{k-1}^{\text{opt}}(x_k, x_{k-1}) = \frac{\mu_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)}{\mu_k(x_k)}.$$

Extra Comments : Selection of Artificial Kernels

• Hard to use $\left\{L_{k-1}^{\text{opt}}\right\}$ as do not know $\mu_n(x_n)$ but can be approximated.

$\Rightarrow$ Suboptimal but asymptoticaly consistent strategy consists of using the approximation of L_{n-1}^{opt} given below

$$L_{n-1}(x_n, x_{n-1}) = \frac{\pi_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n)}{\pi_{n-1} K_n(x_n)} \Rightarrow w_n(x_{n-1:n}) = \frac{\pi_n(x_n)}{\pi_{n-1} K_n(x_n)}$$

• If K_n is an MCMC kernel of invariant distribution π_n and $\pi_{n-1} \simeq \pi_n$ then a convenient choice is

$$L_{n-1}(x_n, x_{n-1}) = \frac{\pi_n(x_{n-1}) K_n(x_{n-1}, x_n)}{\pi_n(x_n)} \Rightarrow w_n(x_{n-1:n}) = \frac{\pi_n(x_{n-1})}{\pi_{n-1}(x_{n-1})}.$$

$\Rightarrow$ this choice of L_n kernel is a good approximation of L_{n-1}^{opt} when $\pi_{n-1} \approx \pi_n$

• Remark. Aestethic but generally bad idea: $L_{n-1} = K_n$.

Extra Comments : $L_{n-1} = K_n$

- By selecting $L_{n-1}{=}K_n$, the incremental weight looks like a MH ratio.

- This choice might be "aesthetic" but is inefficient in most cases.

- Toy example where $E = R$, $\pi_1(x) = N\left(x; 0, \sigma_1^2\right)$, $\pi_2(x) = N\left(x; 0, \sigma_2^2\right)$ and $K_2(x, x') = N\left(x'; x, \sigma_K^2\right)$ with $\sigma_1 \geq \sigma_2$.

- Assume particles at time 1 have been resampled so that $\mu_1 = \pi_1$. Whatever σ_K , the IS weight on the marginal space is upper bounded

$$\frac{\pi_2(x)}{\mu_2(x)} < \infty.$$

and has finite variance.

• Now assume one does not perform importance sampling on E but on the extended space $E \times E$ by introducing an auxiliary kernel $L_1 = K_2$. In this case the importance weight is given by

$$\frac{\pi_2(x') K_2(x', x)}{\pi_1(x) K_2(x, x')} = \frac{\pi_2(x')}{\pi_1(x)}.$$

• It is not upper bounded over $E \times E$ and does not admit a finite variance.

• So if E is not a compact set, one cannot expect this choice to perform well.

Extra Comments : $L_{n-1} = \frac{\pi_n(x_{n-1})K_n(x_{n-1},x_n)}{\pi_n(x_n)}$

- Consider the case where K_n is an MCMC kernel of invariant distribution π_n.
- Convenient choice for L_{n-1} is

$$L_{n-1}\left(x_n, x_{n-1}\right) = \frac{\pi_n\left(x_{n-1}\right) K_n\left(x_{n-1}, x_n\right)}{\pi_n\left(x_n\right)}$$

- Get incremental weight which is independent of x_n

$$\frac{\pi_n\left(x_{n-1}\right)}{\pi_{n-1}\left(x_{n-1}\right)}.$$

- This kernel is a good approximation of L_{n-1}^{opt} if $\pi_{n-1} \approx \pi_n$.
- The weights are independent of x_n.

• Therefore even if K_n is fast mixing (in the limiting case $K_n(x, x') = \pi_n(x')$) then the particles are still weighted according to $\frac{\pi_n(x_{n-1})}{\pi_{n-1}(x_{n-1})}$.

• So if resampling is used, it is more efficient to first resample particles with respect to their weights $\frac{\pi_n(x_{n-1})}{\pi_{n-1}(x_{n-1})}$ before sampling them according to K_n, instead of using sampling then resampling.

• This increases the diversity among particles at time n and it is possible since the weight is independent of x'.

• This simple approach is attractive but could perform poorly in some applications.

• Consider a case where the regions of high probability mass for π_{n-1} and π_n are located in disjoint parts of the state space; e.g. say a sequential Bayesian estimation problem and y_n is an informative observation.

• In this case when the particles are resampled, only very few will survive and these surviving particles might not even be located in the regions of high probability mass of π_n.

• Hence, except when K_n is mixing very quickly, performance will not be satisfactory.

Extra Comments : homogeneous case

• Homogeneous case: $\pi_n = \pi$, $K_n = K$, $L_n = L$ where K is an MCMC kernel of invariant distribution π and

$$L\left(x, x'\right) = \frac{\pi\left(x'\right) K\left(x', x\right)}{\pi\left(x\right)}.$$

• After having resampled the particles once, they are approximately distributed according to π.

• In this case, $L\left(x, x'\right)$ is equal to the optimal kernel and the associated importance weights are now equal to 1.

• Each particle evolves independently according to K and it is not necessary to make them interact anymore.

• Tempting to consider a different L kernel, in this case the particles would have to be resampled periodically.

• Such a strategy does not appear justified as resampling is not performed to modify the marginal distribution of the particles but just the correlation between surviving particles at two successive time instants.

• This only limits the diversity in the set of particles and one cannot expect the variance of the resulting estimate to be lower than the one obtained using non-interacting MCMC chains.