Sequential Monte Carlo Methods for Bayesian Computation

A. Doucet

Feb. 2015

Summary of Previous Lectures

- SMC allows us to sample approximately from $\{p(x_{1:t}|y_{1:t})\}_{t\geq 1}$ and compute $\{p(y_{1:t})\}_{t\geq 1}$ for state-space models.
- SMC are useful far beyond this class of models
  - More general time series models
  - General "static" Bayesian computation
  - Some control problems
  - Rare event simulation
- We show here how to extend trivially SMC techniques previously discussed to a general class of problems.

Generic Problem

- Consider a sequence of probability distributions $\{\pi_t\}_{t\geq 1}$ defined on a sequence of spaces $\{\mathcal{X}_t\}_{t\geq 1}$ where $\mathcal{X}_t = \mathcal{X}^t$.
- Each distribution $\pi_t(x_{1:t})$ is assumed known *up to a normalizing constant*, i.e.

$$\pi_t(x_{1:t}) = \frac{\gamma_t(x_{1:t})}{Z_t}$$

  where $\gamma_t : \mathcal{X}_t \to \mathbb{R}^+$ can be computed pointwise but Z_t cannot.
- SMC methods can be used to sample sequentially from $\{\pi_t\}_{t\geq 1}$ and compute Z_t.

Examples

- **State-space models**

$$\gamma_t(x_{1:t}) = p(x_{1:t}, y_{1:t}) = \mu(x_1)\prod_{k=2}^{t} f(x_k|x_{k-1})\prod_{k=1}^{t} g(y_k|x_k),$$

$$Z_t = p(y_{1:t}) = \int\cdots\int \mu(x_1)\prod_{k=2}^{t} f(x_k|x_{k-1})\prod_{k=1}^{t} g(y_k|x_k)\,dx_{1:t}$$

- **General time series models**

$$\begin{aligned}\gamma_t(x_{1:t}) &= p(x_{1:t}, y_{1:t})\\ &= \mu(x_1)\prod_{k=2}^{t} f(x_k|y_{1:k-1}, x_{1:k-1})\prod_{k=1}^{t} g(y_k|y_{1:k-1}, x_{1:k})\\ Z_t &= p(y_{1:t})\end{aligned}$$

- **Nonparametric Bayes**: Wood & Griffiths, 2006; Caron & D., 2007, Saeedi & Bouchard-Côté, 2011; Ahmed & Smola, 2012 etc.
- **Graphical models, coalescent models, phylogenetic trees**: Gorur & Teh, 2008, Ihler, Frank, Smyth, 2009 etc.

Examples

- **Risk sensitive control** (Kantas, 2009):

$$\gamma_t(x_{1:t}) = \mu(x_1)\prod_{k=2}^{t} f(x_k|x_{1:k-1})\prod_{k=1}^{t}\exp(C(x_k))$$

$$Z_t = \mathbb{E}_\mu\left[\exp\left(\sum_{k=1}^{t} C(x_k)\right)\right]$$

- **Eigenmeasure/Eigenvalue** (Heterington, 1984): $K(x,y) \geq 0$ with $\int K(x,y)\,dy \neq 1$, $\nu(y) = \lambda\int \nu(x)K(x,y)\,dx$

$$\gamma_t(x_{1:t}) = \mu(x_1)\prod_{k=2}^{t} K(x_{t-1},x_t),$$

$$\pi_t(x_t) \longrightarrow \nu(x_t) \text{ and } \frac{Z_{t+1}}{Z_t} \longrightarrow \lambda$$

SMC for General Target Distributions

- Assume at time $t-1$, we have $\left\{W_{t-1}^{(i)}, X_{1:t-1}^{(i)}\right\}$ approximating $\pi_{t-1}(x_{1:t-1})$.
- We introduce $q_t(x_t|x_{1:t-1})$ and use

$$\pi_t(x_{1:t}) = \frac{\gamma_t(x_{1:t})}{Z_t} = \frac{w_t(x_{1:t})\,\pi_{t-1}(x_{1:t-1})\,q_t(x_1|x_{1:t-1})}{Z_t}$$

where the incremental weight is

$$w_t(x_{1:t}) = \frac{\gamma_t(x_{1:t})}{\gamma_{t-1}(x_{1:t-1})\,q_t(x_1|x_{1:t-1})}.$$

- By sampling $X_t^{(i)} \sim q_t(x_t|x_{1:t-1})$, we obtain

$$\widehat{\pi}_t(x_{1:t}) = \sum_{i=1}^{N} W_{1:t}^{(i)}\delta_{X_{1:t}^{(i)}}(x_{1:t}),\ W_t^{(i)} \propto W_{t-1}^{(i)} w_t\left(X_{1:t}^{(i)}\right).$$

SMC for General Target Distributions

- At time 1, you are interested in

$$\pi_1(x_1) = \frac{\gamma_1(x_1)}{Z_1}.$$

- Introduce a distribution $q_1(x_1)$ easy to sample from and use

$$\pi_1(x_1) = \frac{w_1(x_1)\,q_1(x_1)}{Z_1},\ w_1(x_1) = \frac{\gamma_1(x_1)}{q(x_1)}.$$

- Sample $X_1^{(i)} \sim q_1(x_1)$ thus $\widehat{q}_1(x_1) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_1^{(i)}}(x_1)$ and

$$\widehat{\pi}_1(x_1) = \sum_{i=1}^{N} W_1^{(i)}\delta_{X_1^{(i)}}(x_1),\ W_1^{(i)} \propto w_1\left(X_1^{(i)}\right).$$

A General SMC Algorithm

Assume we have N weighted particles $\left\{W_{t-1}^{(i)}, X_{1:t-1}^{(i)}\right\}$ approximating $\pi_{t-1}(x_{1:t-1})$ then at time t,

- Sample $\widetilde{X}_t^{(i)} \sim q_t\left(x_t|X_{1:t-1}^{(i)}\right)$, set $\widetilde{X}_{1:t}^{(i)} = \left(X_{1:t-1}^{(i)}, \widetilde{X}_t^{(i)}\right)$ and

$$\widetilde{\pi}_t(x_{1:t}) = \sum_{i=1}^{N} W_{1:t}^{(i)}\delta_{X_{1:t}^{(i)}}(x_{1:t}),$$

$$W_t^{(i)} \propto W_{t-1}^{(i)} w_t\left(X_{1:t-1}^{(i)}, \widetilde{X}_t^{(i)}\right),$$

$$\widehat{Z_t/Z_{t-1}} = \sum_{i=1}^{N} W_{t-1}^{(i)} w_t\left(X_{1:t-1}^{(i)}, \widetilde{X}_t^{(i)}\right).$$

- If ESS$< N/2$ resample $X_{1:t}^{(i)} \sim \widetilde{\pi}_t(x_{1:t})$ and set $W_t^{(i)} \leftarrow \frac{1}{N}$ to obtain $\widehat{\pi}_t(x_{1:t}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_{1:t}^{(i)}}(x_{1:t})$.

Selection of the Proposal

- The proposal minimizing the variance of incremental weight

$$w_t\left(x_{1:t}\right) = \frac{\gamma_t\left(x_{1:t}\right)}{\gamma_{t-1}\left(x_{1:t-1}\right) q_t\left(x_t \middle| x_{1:t-1}\right)} \propto \frac{\pi_t\left(x_{1:t}\right)}{\pi_{t-1}\left(x_{1:t-1}\right) q_t\left(x_t \middle| x_{1:t-1}\right)}$$

  is

$$q_t^{\text{opt}}\left(x_t \middle| x_{1:t-1}\right) = \pi_t\left(x_t \middle| x_{1:t-1}\right).$$

- In this case, we have

$$\begin{aligned} w_t\left(x_{1:t}\right) &= w_{t-1}\left(x_{1:t-1}\right) \frac{\gamma_t\left(x_{1:t}\right)}{\gamma_{t-1}\left(x_{1:t-1}\right) q_t^{\text{opt}}\left(x_t \middle| x_{1:t-1}\right)} \\ &= w_{t-1}\left(x_{1:t-1}\right) \frac{\gamma_t\left(x_{1:t-1}\right)}{\gamma_{t-1}\left(x_{1:t-1}\right)}. \end{aligned}$$

Application: Rao-Blackwellized SMC Methods

- "A Good Monte Carlo is a Dead Monte Carlo": Whenever you can integrate out analytically variables, do it.
- **Example**. Consider $X_t = \left(U_t, Z_t\right)$ where

$$\begin{aligned} U_t \middle| U_{t-1} &\sim f\left(\cdot \middle| U_{t-1}\right), \\ Z_t &= A_{U_t} Z_{t-1} + B_{U_t} V_t, \ V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, I_{t_v}\right), \\ Y_t &= C_{U_t} Z_t + D_{U_t} W_t, \ W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, I_{t_w}\right). \end{aligned}$$

- SIS can be used to approximate $p\left(x_{1:t} \middle| y_{1:t}\right)$ but this is inefficient as

$$p\left(x_{1:t} \middle| y_{1:t}\right) = p\left(u_{1:t} \middle| y_{1:t}\right) \underbrace{p\left(z_{1:t} \middle| y_{1:t}, u_{1:t}\right)}_{\text{Gaussian}}.$$

- Rao-Blackwellized SIS methods target $p\left(u_{1:t} \middle| y_{1:t}\right) \propto p\left(y_{1:t} \middle| u_{1:t}\right) p\left(u_{1:t}\right)$ instead of $p\left(x_{1:t} \middle| y_{1:t}\right)$ (D, Godsill & Andrieu, 2000; Liu & Chen, 2000). Also apply to dynamic Bayesian nets (Murphy, D., De Freitas, Russell, 2000).

Another Generic Sampling Problem

- Let $\left\{\pi_t\right\}_{t \geq 1}$ be a *sequence of probability distributions* defined on $\mathcal{X}$ such that $\pi_t\left(x\right) = \pi_t\left(x\right)$ and each $\pi_t\left(x\right)$ is known *up to a normalizing constant*, i.e.

$$\pi_t\left(x\right) = \underbrace{Z_t^{-1}}_{\text{unknown}} . \underbrace{\gamma_t\left(x\right)}_{\text{known}}.$$

- Standard SMC do not apply as all the distributions are defined on the same space.
- It is possible to adapt SMC to this problem.

Applications

- *Sequential Bayesian Inference:* $\pi_t\left(x\right) = p\left(x \middle| y_{1:t}\right)$.
- *Global optimization:* $\pi_t\left(x\right) \propto \left[\pi\left(x\right)\right]^{\eta_t}$ with $\left\{\eta_t\right\}$ increasing sequence such that $\eta_t \rightarrow \infty$.
- *Sampling from a fixed target* π : $\pi_t\left(x\right) \propto \mu\left(x\right)\left[g\left(y \middle| x\right)\right]^{\phi_t}$ where μ easy to sample and $\phi_t = 0$, $\phi_t > \phi_{t-1}$ and $\phi_P = 1$.
- *Rare event simulation* $\pi\left(A\right) \ll 1$: $\gamma_t\left(x\right) = \pi\left(x\right) 1_{\mathcal{X}_t}\left(x\right)$ with Z_1 known, $\mathcal{X}_1 = \mathcal{X}$, $\mathcal{X}_t \subset \mathcal{X}_{t-1}$ and $\mathcal{X}_P = A$ then $Z_T = \pi\left(A\right)$.

Standard Approaches

- Run a MCMC (e.g. Metropolis-Hastings) algorithm to sample from each target distribution π_t; i.e. build a Markov kernel $K_t\left(x' \middle| x\right)$ such that

$$\pi_t\left(x'\right) = \int_{\mathcal{X}} \pi_t\left(x\right) K_t\left(x' \middle| x\right) dx$$

and simulate a Markov chain $\left\{X_t^{(i)}\right\}$: $X_t^{(1)} \sim \mu_t$ and $X_t^{(i)} \sim K_t\left(x' \middle| X_t^{(i-1)}\right)$.

- Under weak assumptions, we have $\lim_{i\to\infty} \mathcal{L}\left(X_t^{(i)}\right) = \pi_t$ i.e. $X_t^{(i)}$ is asymptotically distributed according to π_t and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(X_t^{(i)}\right) = \int \varphi\left(x\right) \pi_t\left(x\right) dx.$$

Some Potential Limitations

- Convergence to π_t can be slow and is difficult to diagnose.
- It does not give an estimate of Z_t with 'good' properties.
- If π_{t-1} and π_t are 'close', then it should be possible to devise a cleverer strategy.
- We can use instead importance Sampling.

Importance Sampling

- Let the *target distribution* be $\pi_t\left(x\right) = Z_t^{-1}\gamma_t\left(x\right)$ and μ_t be an *importance distribution* then

$$\pi_t\left(x\right) = \frac{w_t\left(x\right) q_t\left(x\right)}{\int w_t\left(x\right) q_t\left(x\right) dx} \text{ where } w_t\left(x\right) = \frac{\gamma_t\left(x\right)}{q_t\left(x\right)},$$
$$Z_t = \int w_t\left(x\right) q_t\left(x\right) dx$$

- By sampling N i.i.d. particles $X_t^{(i)} \sim q_t$ then $\widehat{q}_t\left(x\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^{(i)}}\left(x\right)$ and

$$\widehat{\pi}_t\left(x\right) = \sum_{i=1}^{N} W_t^{(i)} \delta_{X_t^{(i)}}\left(x\right) \text{ where } W_t^{(i)} \propto w_t\left(X_t^{(i)}\right),$$
$$\widehat{Z}_t = \frac{1}{N}\sum_{i=1}^{N} w_t\left(X_t^{(i)}\right).$$

Limitations of Importance Sampling

- Importance Sampling (IS) is a straightforward method to use if q_t is easy to sample.
- For the estimates to have reasonable variances, we need to select very carefully the importance distribution.
- Naive strategies provide typically estimates with exponential variance in the dimension.
- For state-space models discussed previously, $\dim\left(\mathcal{X}\right) < 10$ in most cases. For static problems, we often have $\dim\left(\mathcal{X}\right) > 1000$.

General Principle

- "Philosophy": Start by doing simple things before trying to do complex things.
- Develop a sequential/iterative IS strategy where we start by approximating a simple target distribution π_1. Then targets evolve over time and *we build the importance distribution sequentially*; i.e. at time t, we use q_{t-1} to build q_t.
- This approach only makes sense if the sequence $\{\pi_t\}$ is not arbitrary; i.e. π_{t-1} somewhat close to π_t.

Sequential Importance Sampling

- At time 1, sample N ($N \gg 1$) particles $X_1^{(i)} \sim q_1$ to obtain the following IS estimates

$$\widehat{\pi}_1(x) = \sum_{i=1}^{N} W_1^{(i)} \delta_{X_1^{(i)}}(x) \text{ where } W_1^{(i)} \propto w_1\left(X_1^{(i)}\right),$$

$$\widehat{Z}_1 = \frac{1}{N}\sum_{i=1}^{N} w_1\left(X_1^{(i)}\right)$$

- *Remark*: Estimates have reasonable variance only if discrepancy between π_1 and μ_1 small; hence the need to start with easy to sample or approximate π_1.

Sequential Importance Sampling

- At time $t-1$, one has N particles $\left\{X_{t-1}^{(i)}, W_{t-1}^{(i)}\right\}$

$$X_{t-1}^{(i)} \sim q_{t-1}(x) \cdot \text{ and } W_{t-1}^{(i)} \propto \frac{\pi_{t-1}\left(X_{t-1}^{(i)}\right)}{q_{t-1}\left(X_{t-1}^{(i)}\right)}.$$

- Move the particles according to transition kernel

$$X_t^{(i)} \sim K_t\left(x \middle| X_{t-1}^{(i)}\right) \Rightarrow q_t(x') = \int q_{t-1}(x) K_t\left(x' \middle| x\right) dx$$

- Optimal transition kernel $K_t(x'|x) = \pi_t(x')$ cannot be used so we need alternatives.

Some Potential Choices

- $K_t(x'|x) = K_t(x')$ with
  - simple parametric form (e.g. Gaussian, multinomial etc.) (e.g. Cappé et al, 2005);
  - semi-parametric based on $\widehat{\mu}_{t-1}(x)$ (e.g. West, 1993; Titterington, 2001)
- $K_t(x'|x)$ MCMC kernel of invariant distribution π_t.
  - burn-in correction by importance sampling (Gilks & Berzuini, 2001; Neal, 2001; Crisan & D., 2000).
- $K_t(x'|x)$ approximation of a Gibbs sampler of invariant distribution π_t.

Sequential Importance Sampling

- At time $t = 1$, sample $X_1^{(i)} \sim q_1(x)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- At time $t \geq 2$, sample $X_t^{(i)} \sim K_t\left(x \middle| X_{t-1}^{(i)}\right)$ and set $w_t\left(X_t^{(i)}\right) = \frac{\gamma_t\left(X_t^{(i)}\right)}{q_t\left(X_t^{(i)}\right)}$ where

$$q_t(x') = \int q_{t-1}(x) K_t(x' \mid x)\, dx.$$

- We have

$$\widehat{\pi}_t(x) = \sum_{i=1}^{N} W_t^{(i)} \delta_{X_t^{(i)}}(x_t), \quad \widehat{Z}_t = \frac{1}{N} \sum_{i=1}^{N} w_t\left(X_t^{(i)}\right).$$

Implementation Issues

- In most cases, we *cannot* compute the marginal importance distribution

$$\begin{aligned} q_t(x_t) &= \int q_{t-1}(x_{t-1}) K_t(x_t \mid x_{t-1})\, dx_{t-1} \\ &= \int q_1(x_1) \prod_{k=2}^{t} K_k(x_k \mid x_{k-1})\, dx_{1:t-1}. \end{aligned}$$

- Monte Carlo approximation is possible

$$\widetilde{q}_t(x_t) = \int \widehat{q}_{t-1}(x_{t-1}) K_t(x_t \mid x_{t-1})\, dx_{t-1} = \frac{1}{N} \sum_{i=1}^{N} K_t\left(x \middle| X_{t-1}^{(i)}\right)$$

but is computationally intensive $\mathcal{O}\left(N^2\right)$.

An Artificial Target Distribution

- *Problem summary*: It is impossible to compute pointwise $q_t(x_t)$ hence $\gamma_t(x_t)/q_t(x_t)$ except when $t = 1$.
- *Solution*: Perform importance sampling on extended space.
- At time 2,

$$\frac{\pi_2(x_2)}{q_2(x_2)} = \frac{\pi_2(x_2)}{\int q_1(x_1) K_2(x_2 \mid x_1)\, dx_1} \text{ cannot be evaluated}$$

but alternative weights can be defined

$$\frac{\text{new joint target distribution}}{\text{joint importance distribution}} = \frac{\pi_2(x_2) L_1(x_1 \mid x_2)}{q_1(x_1) K_2(x_2 \mid x_1)}$$

where $L_1(x_1 \mid x_2)$ is an *arbitrary* (backward) Markov kernel.

- "Proof" of validity:

$$\int \pi_2(x_2) L_1(x_1 \mid x_2)\, dx_1 = \pi_2(x_2) \underbrace{\int L_1(x_1 \mid x_2)\, dx_1}_{=1 \text{ whatever being } L_1} = \pi_2(x_2)$$

An Extended Target Distribution Trick.... Again

- Similarly at time t, $q_t(x_t)$ cannot be computed so perform importance sampling on an extended space between

$$\widetilde{q}_t(x_{1:t}) = q_1(x_1) \prod_{k=2}^{t} K_k(x_k \mid x_{k-1})$$

and an extended artificial joint target distribution

$$\widetilde{\pi}_t(x_{1:t}) = \pi_t(x_t) \prod_{k=1}^{t-1} L_k(x_k \mid x_{k+1})$$

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

- "Proof" of validity

$$\int \widetilde{\pi}_t(x_{1:t})\, dx_{1:t-1} = \pi_t(x_t) \underbrace{\int \prod_{k=1}^{t-1} L_k(x_k \mid x_{k+1})\, dx_{1:t-1}}_{=1 \text{ whatever being } \{L_k\}} = \pi_t(x_t).$$

No Free Lunch

- By extending the integration space, the variance of the importance weights can only increase.
- The optimal kernels $\left\{L_k^{\text{opt}}\right\}$ are the ones bringing us back to the case where there is no space extension; i.e.

$$L_{t-1}^{\text{opt}}\left(x_{t-1}\middle|x_t\right)=\frac{q_{t-1}\left(x_{t-1}\right)K_t\left(x_t\middle|x_{t-1}\right)}{q_t\left(x_t\right)}$$

- The result follows straightforwardly from the forward-backward formula for Markov processes

$$\widetilde{q}_t\left(x_{1:t}\right)=q_1\left(x_1\right)\prod_{k=2}^{t}K_k\left(x_k\middle|x_{k-1}\right)=q_t\left(x_t\right)\prod_{k=2}^{t}L_{t-1}^{\text{opt}}\left(x_{t-1}\middle|x_t\right)$$

- L_{t-1}^{opt} cannot typically be computed (though there are important exceptions) but can be properly approximated in numerous cases.
- *Even if an approximation is used, the estimates are still asymptotically consistent.*

Back to Standard SMC Methods

- We need to sample from a sequence of (artificial) target distributions $\left\{\widetilde{\pi}_t\right\}$ of increasing dimension.
- **Conceptual difference**: Given $\left\{K_t\right\}$, $\left\{\widetilde{\pi}_t\right\}$ has been constructed in a "clever" way such that

$$\int\widetilde{\pi}_t\left(x_{1:t}\right)dx_{1:t-1}=\pi_t\left(x_t\right)$$

whereas usually the sequence of targets $\left\{\widetilde{\pi}_t\right\}$ is fixed and $\left\{K_t\right\}$ is designed accordingly.

- Because we typically cannot use $\left\{L_k^{\text{opt}}\right\}$, the variance of the weights typically increases over time and it is necessary to resample.

SMC Sampler

- At time $t=1$, sample $X_1^{(i)}\sim\mu_1\left(\cdot\right)$ and set $w_1\left(X_1^{(i)}\right)=\frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- Resample $\left\{X_1^{(i)},W_1^{(i)}\right\}$ to obtain new particles also denoted $\left\{X_1^{(i)}\right\}$
- At time $t\geq 2$
  - sample $X_t^{(i)}\sim K_t\left(X_{t-1}^{(i)},x_t\right)$.
  - compute $w_t\left(X_{t-1}^{(i)},X_t^{(i)}\right)=\frac{\gamma_t\left(X_t^{(i)}\right)L_{t-1}\left(X_{t-1}^{(i)}\middle|X_t^{(i)}\right)}{\gamma_{t-1}\left(X_{t-1}^{(i)}\right)K_t\left(X_t^{(i)}\middle|X_{t-1}^{(i)}\right)}$.
- Resample $\left\{X_t^{(i)},W_t^{(i)}\right\}$ to obtain new particles also denoted $\left\{X_t^{(i)}\right\}$.

Monte Carlo Estimates

- We obtain

$$\widehat{\pi}_t\left(x\right)=\sum_{i=1}^{N}W_t^{(i)}\delta_{X_t^{(i)}}\left(x\right).$$

- Ratio of Normalizing Constants

$$\frac{Z_t}{Z_{t-1}}=\int\frac{\gamma_t\left(x_t\right)L_{t-1}\left(x_{t-1}\middle|x_t\right)}{\gamma_{t-1}\left(x_{t-1}\right)K_t\left(x_t\middle|x_{t-1}\right)}\pi_{t-1}\left(x_{t-1}\right)K_t\left(x_t\middle|x_{t-1}\right)dx_{t-1:t}$$

so

$$\frac{\widehat{Z_t}}{Z_{t-1}}=\sum_{i=1}^{N}W_{t-1}^{(i)}\frac{\gamma_t\left(X_t^{(i)}\right)L_{t-1}\left(X_{t-1}^{(i)}\middle|X_t^{(i)}\right)}{\gamma_{t-1}\left(X_{t-1}^{(i)}\right)K_t\left(X_t^{(i)}\middle|X_{t-1}^{(i)}\right)}.$$

- This is a generalization of the celebrated Jarzynski-Crooks identity (1997); see Neal (2001).

Guidelines...

- **First step**: Build a sequence of distributions $\{\pi_t\}$ going from π_1 easy to sample/approximate to $\pi_P = \pi$; e.g. $\pi_t(x) \propto \mu(x)\left[g(y|x)\right]^{\phi_t}$ where μ easy to sample and $\phi_t = 0$, $\phi_t > \phi_{t-1}$ and $\phi_P = 1$.
- **Second step**: Introduce a sequence of transition kernels $\{K_t\}$; e.g. K_t MCMC sampler of invariant distribution π_t.
- **Third step**: Introduce a sequence of backward kernels $\{L_t\}$ equal/approximating L_t^{opt}; e.g.

$$L_{t-1}(x_{t-1}|x_t) = \frac{\pi_{t-1}(x_{t-1})K_t(x_t|x_{t-1})}{\int \pi_{t-1}(x)K_t(x_t|x)dx}$$
$$\Rightarrow \frac{\gamma_t(x_t)L_{t-1}(x_{t-1}|x_t)}{\gamma_{t-1}(x_{t-1})K_t(x_t|x_{t-1})} = \frac{\gamma_t(x_t)}{\int \gamma_{t-1}(x)K_t(x_t|x)dx}$$

$$L_{t-1}(x_{t-1}|x_t) = \frac{\pi_t(x_{t-1})K_t(x_t|x_{t-1})}{\pi_t(x_t)}$$
$$\Rightarrow \frac{\gamma_t(x_t)L_{t-1}(x_{t-1}|x_t)}{\gamma_{t-1}(x_{t-1})K_t(x_t|x_{t-1})} = \frac{\gamma_t(x_{t-1})}{\gamma_{t-1}(x_{t-1})}.$$

Bayesian Inference for Mixture Models

- We have observations $y_{1:m}$ with

$$p(y|\mu_{1:k}, \sigma_{1:k}, w_{1:k}) = \sum_{i=1}^{k} w_i \mathcal{N}(y; \mu_i, \sigma_i)$$

- Assume that $k = 4$, $w_i = 1/k$ and $\sigma_i = 0.55$ are known, $\mu_{1:k}$ is uniform on the k-dimensional hypercube $[-10, 10]^k$.
- We simulate $m = 100$ observations for $\mu = \mu_{1:4} = (-3, 0, 3, 6)$ and want to sample from

$$p(\mu_{1:k}|y_{1:m}) \propto p(\mu_{1:k})\, p(y_{1:m}|\mu_{1:k}, \sigma_{1:k}, w_{1:k}).$$

- Invariance of the posterior to permutation of the labels of the parameters gives it $k! = 24$ symmetric modes
- Basic random-walk MCMC and importance sampling methods fail.

Experimental Setups

- We build the sequence of P distributions

$$\pi_t(\mu_{1:k}) \propto p(\mu_{1:k})\left[p(y_{1:m}|\mu_{1:k}, \sigma_{1:k}, w_{1:k})\right]^{\phi_t}$$

  where $\phi_1 = 0 < \phi_2 < ... < \phi_P = 1$.
- MCMC sampler to sample from π_t : update $\mu_{1:4}$ via a MH kernel with additive normal random walk.
- We use

$$L_{t-1}(x_{t-1}|x_t) = \frac{\pi_t(x_{t-1})K_t(x_t|x_{t-1})}{\pi_t(x_t)}.$$

Running Times SMC Samplers: CPU vs GPU

Table: Running times for the Sequential Monte Carlo Sampler for various values of N.

N	CPU (mins)	8800GT (secs)	Speedup	GTX280 (secs)	Speedup
8192	4.44	1.192	223.5	0.597	446
16384	8.82	2.127	249	1.114	475
32768	17.7	3.995	266	2.114	502
65536	35.3	7.889	268	4.270	496
131072	70.6	15.671	270	8.075	525
262144	141	31.218	271	16.219	522

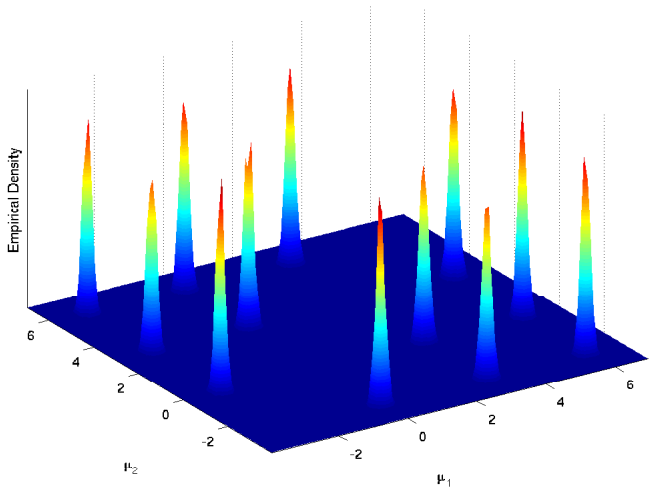

Figure: Estimated marginal posterior density $p(\mu_{1:2}|y_{1:m})$ from SMC samples, $N = 8192$.

Adaptive SMC Samplers

- A powerful advantage of SMC samplers over MCMC is that they offer more flexibility in terms of adaptation compared to adaptive MCMC as no reliance on ergodicity.
- Adaptive schedule $\{\phi_t\}_{t\geq 1}$ can be built to ensure $ESS_t = \alpha ESS_{t-1}$.
- Adaptive proposals can be build; e.g. scaling of random walk can depend on $\left\{X_t^{(i)}\right\}$.
- Adaptive numbers of particles can be used.

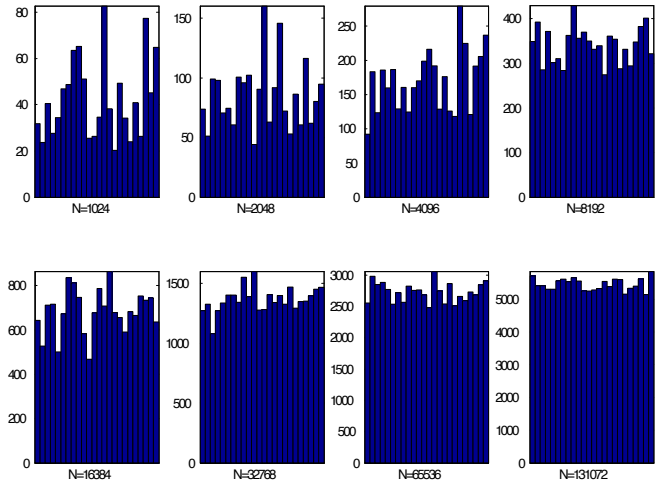

Figure: Effective number of SMC samples from each mode

Discussion

- SMC methods can be used to sample from non-standard high-dimensional distributions.
- This is a powerful alternative/complement to MCMC useful in complex scenarios.
- Very easily parallelizable and GPU implementations already available.
- Adaptive strategies can easily be implemented without affecting convergence.