Sequential Monte-Carlo Samplers for Bayesian Inference in Complex Systems

François Septier

Institut Mines-Télécom/Télécom Lille/LAGIS UMR CNRS 8219

STM2014/CSM2014

July 28-31 2014, ISM

Outline

1. Introduction
   - Context
   - Traditional Monte Carlo methods
   - SMC Samplers

2. Variance reduction schemes for SMC samplers
   - Adaptive sequence of target distributions
   - Recycling schemes
   - Conclusion

Outline

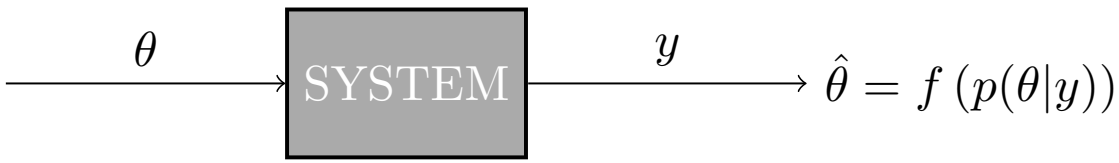
1. Introduction
   - Context
   - Traditional Monte Carlo methods
   - SMC Samplers

2. Variance reduction schemes for SMC samplers
   - Adaptive sequence of target distributions
   - Recycling schemes
   - Conclusion

Context

General Aim : make statements, *inferences*, about unknown features of the physical system based on observed data.

Bayesian Inference :



One important task : finding the estimation of unknown parameters $\boldsymbol{\theta}$ and their distribution.

$\Rightarrow$ Contain all the statistical information about phenomenon.

Context

Bayesian framework :

1. *Prior*, $p(\boldsymbol{\theta})$: expresses what is known about $\boldsymbol{\theta}$ prior to observing data.
2. *Likelihood* , $p(\mathbf{y}|\boldsymbol{\theta})$: probability of observing a data if we have a certain set of parameter values.
3. *Posterior*, $p(\boldsymbol{\theta}|\mathbf{y})$: expresses what is known about $\boldsymbol{\theta}$ after observing data.

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$$

$p(\mathbf{y}) = \int_E p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$: *normalizing constant/ marginal likelihood/ Bayesian Evidence*.

4. *Inference* : derive appropriate inference statements from the posterior distribution. e.g,

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[\varphi(\boldsymbol{\theta})] = \int \varphi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

of some function $\varphi(\boldsymbol{\theta})$.

Generally impossible to obtain a closed-form expression of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$!

Solution

Numerical integration techniques :, e.g, Gaussian Quadrature and Simpson rule, [Ruanaidh $\mathrm{et\ al.}$, 1996] : require a grid of points $\Rightarrow$ are fine in low dimensions, **BUT**

too costly for high dimensional integrals !

Monte Carlo methods : Generate a large number of samples distributed according to $p(\boldsymbol{\theta}|\mathbf{y})$ to obtain consistent simulation-based estimators.

Remarkably flexible and extremely powerful to adapt to many statistical models.

Solution

Numerical integration techniques :, e.g, Gaussian Quadrature and Simpson rule, [Ruanaidh $\mathrm{et\ al.}$, 1996] : require a grid of points $\Rightarrow$ are fine in low dimensions, **BUT**

too costly for high dimensional integrals !

Monte Carlo methods : Generate a large number of samples distributed according to $p(\boldsymbol{\theta}|\mathbf{y})$ to obtain consistent simulation-based estimators.

Remarkably flexible and extremely powerful to adapt to many statistical models.

Bayesian Formulation

Formulation : Target distribution $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$, only known up to a normalizing constant

$$\pi(\boldsymbol{\theta}) = \frac{\gamma(\boldsymbol{\theta})}{Z} \propto \gamma(\boldsymbol{\theta})$$

$$\gamma(\boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$$

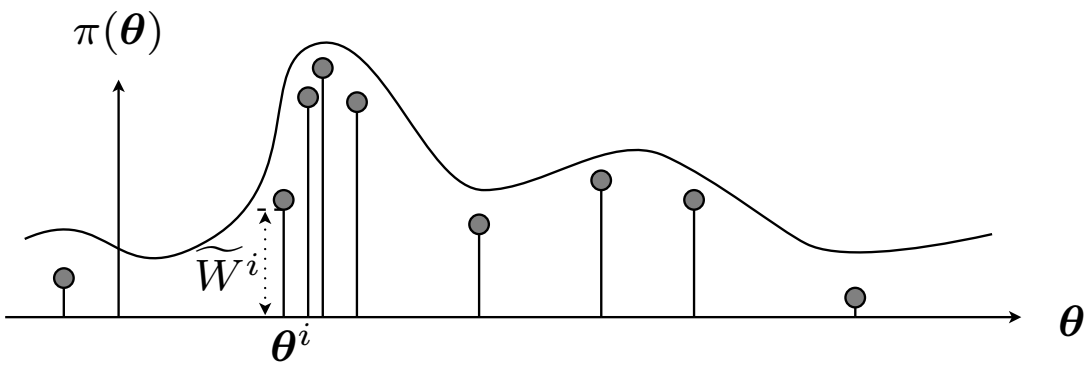
and

$$Z = \int_E \gamma(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_E p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} = p(\mathbf{y}) \quad \textit{Normalizing constant}$$

Importance sampling(IS)

Basic idea : Sample from a *proposal distribution* $\eta(\boldsymbol{\theta})$ instead of $\pi(\boldsymbol{\theta})$ and use weights as correction.

1. Sample $\boldsymbol{\theta}^i \sim \eta(\boldsymbol{\theta})$
2. Correction Step : $W^i = \frac{\gamma(\boldsymbol{\theta}^i)}{\eta(\boldsymbol{\theta})}$



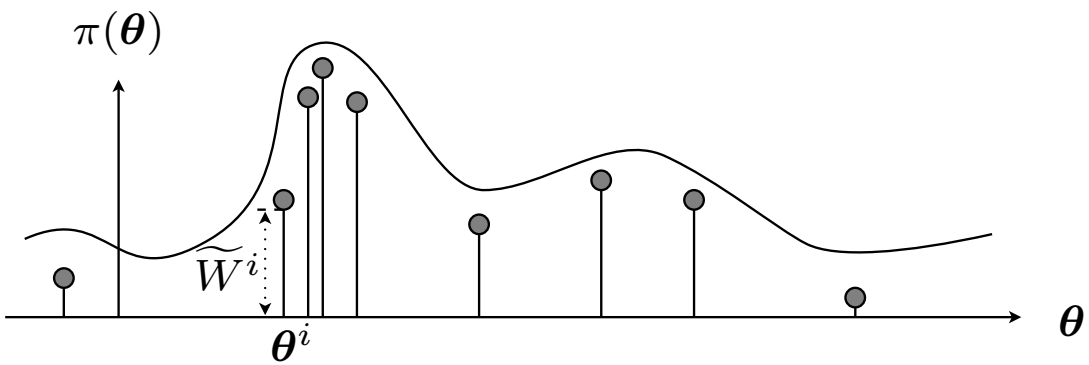
Pros : Good convergence properties, easy to implement.

Cons : Difficult and challenging to choose a good proposal distribution.

Importance sampling(IS)

Basic idea : Sample from a *proposal distribution* $\eta(\boldsymbol{\theta})$ instead of $\pi(\boldsymbol{\theta})$ and use weights as correction.

1. Sample $\boldsymbol{\theta}^i \sim \eta(\boldsymbol{\theta})$
2. Correction Step : $W^i = \frac{\gamma(\boldsymbol{\theta}^i)}{\eta(\boldsymbol{\theta})}$



Pros : Good convergence properties, easy to implement.

Cons : Difficult and challenging to choose a good proposal distribution.

Markov chain Monte Carlo (MCMC)

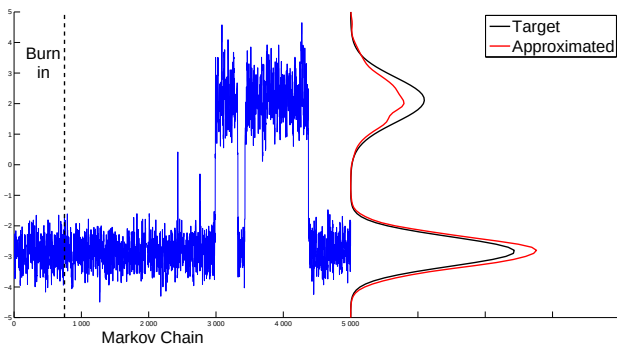
Basic idea : Construct a Markov Chain whose stationary limiting distribution is $p(\boldsymbol{\theta}|\mathbf{y})$.

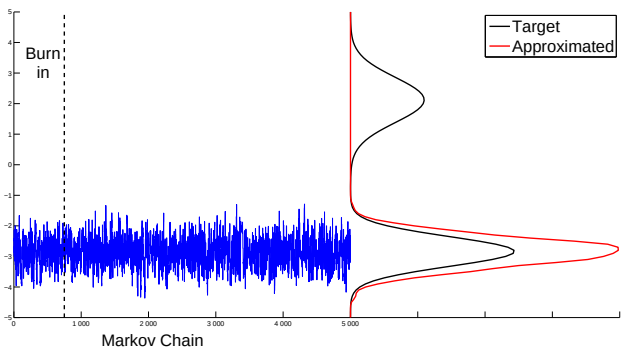
1. Sample $\boldsymbol{\theta}^* \sim \mathcal{K}_t(\boldsymbol{\theta}^{i-1}, \cdot)$
2. Accept $[\boldsymbol{\theta}^i = \boldsymbol{\theta}^*]$ or Reject $[\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}]$ with some probability.

Pros : A lot of available sampling strategies [e.g, local moves all elements or sub-blocks]

Cons :

- Difficult to assess when the Markov chain has reached its stationary regime of interest [Burn-in period].
- Can easily become trapped in local modes.
- Extra complexity cost for estimating normalizing constant.





Markov chain Monte Carlo (MCMC)

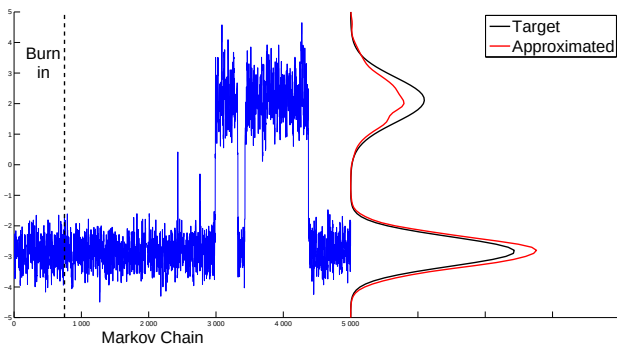
Basic idea : Construct a Markov Chain whose stationary limiting distribution is $p(\boldsymbol{\theta}|\mathbf{y})$.

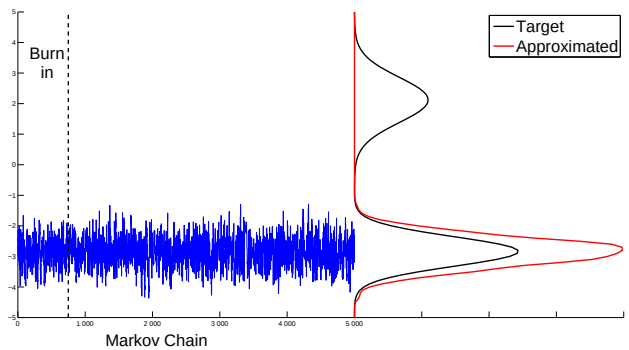
1. Sample $\boldsymbol{\theta}^* \sim \mathcal{K}_t(\boldsymbol{\theta}^{i-1}, \cdot)$
2. Accept $[\boldsymbol{\theta}^i = \boldsymbol{\theta}^*]$ or Reject $[\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}]$ with some probability.

Pros : A lot of available sampling strategies [e.g, local moves all elements or sub-blocks]

Cons :

- Difficult to assess when the Markov chain has reached its stationary regime of interest [Burn-in period].
- Can easily become trapped in local modes.
- Extra complexity cost for estimating normalizing constant.



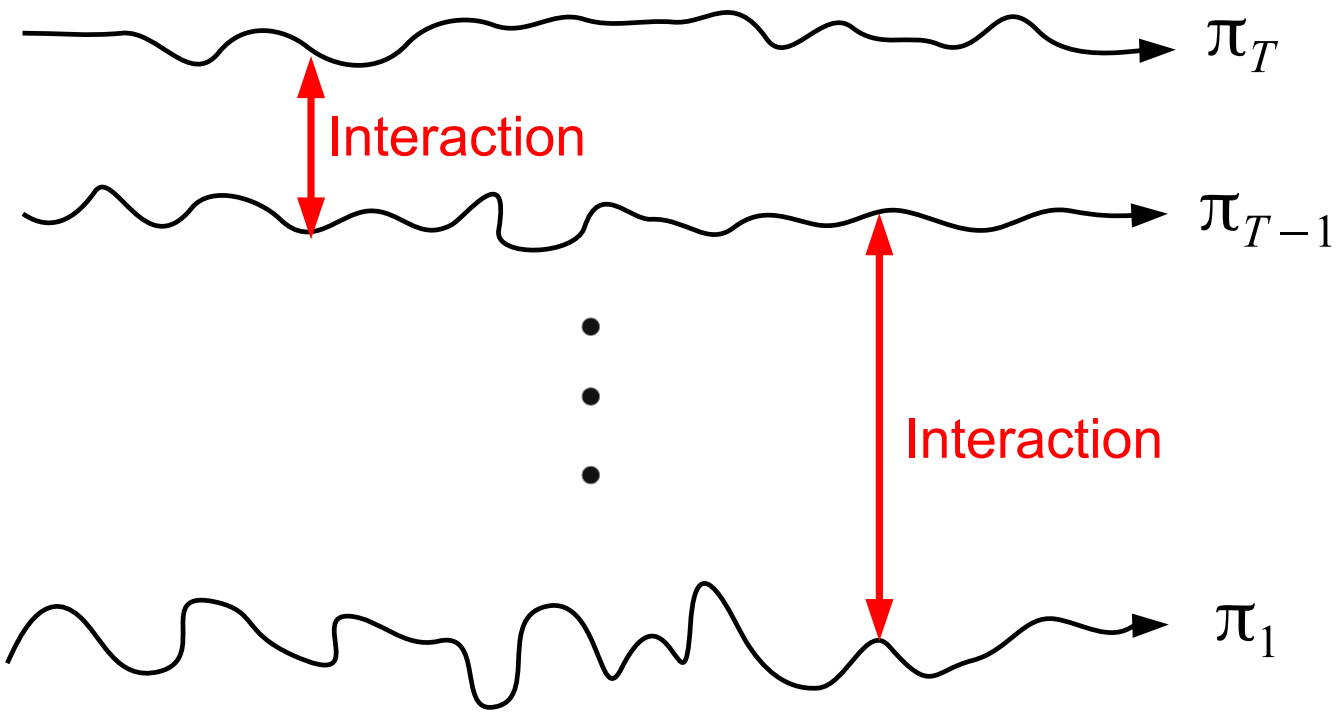


Population MCMC

Population-based MCMC was originally developed by Geyer [Geyer, 1991].

↪ Further advances came in [Liang and Wong, 2000, Liang and Wong, 2001]

Main Idea : Runs T MCMC chains in parallel, each one targeting $\neq$ versions (e.g. annealed) of the posterior distribution and include some interactions between the Markov chains.



Population MCMC

Population-based MCMC was originally developed by Geyer [Geyer, 1991].

$\hookrightarrow$ Further advances came in [Liang and Wong, 2000, Liang and Wong, 2001]

Main Idea : Runs T MCMC chains in parallel, each one targeting $\neq$ versions (e.g. annealed) of the posterior distribution and include some interactions between the Markov chains.

The new target distribution defined in the population-based MCMC is defined as :

$$\pi^*(\boldsymbol{\theta}_{1:T}) = \prod_{k=1}^{T} \pi_k(\boldsymbol{\theta}_k) \tag{1}$$

where it is assumed that the true target of interest (the posterior distribution in Bayesian inference) $\pi = \pi_k$ for at least one $k = 1, \ldots, T$.

Typical choice (for multimodal posterior distribution) :

$$\pi_k(\boldsymbol{\theta}) \propto p(\mathbf{y}|\boldsymbol{\theta})^{\phi_k} p(\boldsymbol{\theta}) \tag{2}$$

with $\forall k$, $\phi_k \in (0, 1]$ and for at least one $k = 1, \ldots, T$, $\phi_k = 1$.

Then 2 $\neq$ MCMC kernels are involved :

- Update each chain
- Interact 2 chains by crossover of exchange move.

Population MCMC

Population-based MCMC was originally developed by Geyer [Geyer, 1991].
$\hookrightarrow$ Further advances came in [Liang and Wong, 2000, Liang and Wong, 2001]
Main Idea : Runs T MCMC chains in parallel, each one targeting $\neq$ versions (e.g. annealed) of the posterior distribution and include some interactions between the Markov chains.

Pros :

- A lot of available sampling strategies [e.g, local moves all elements or sub-blocks]
- Robust to multimodality of the posterior distribution

Cons :

- Difficult to assess when the Markov chain has reached its stationary regime of interest [Burn-in period].
- Extra complexity cost for estimating normalizing constant.

Goal of this work

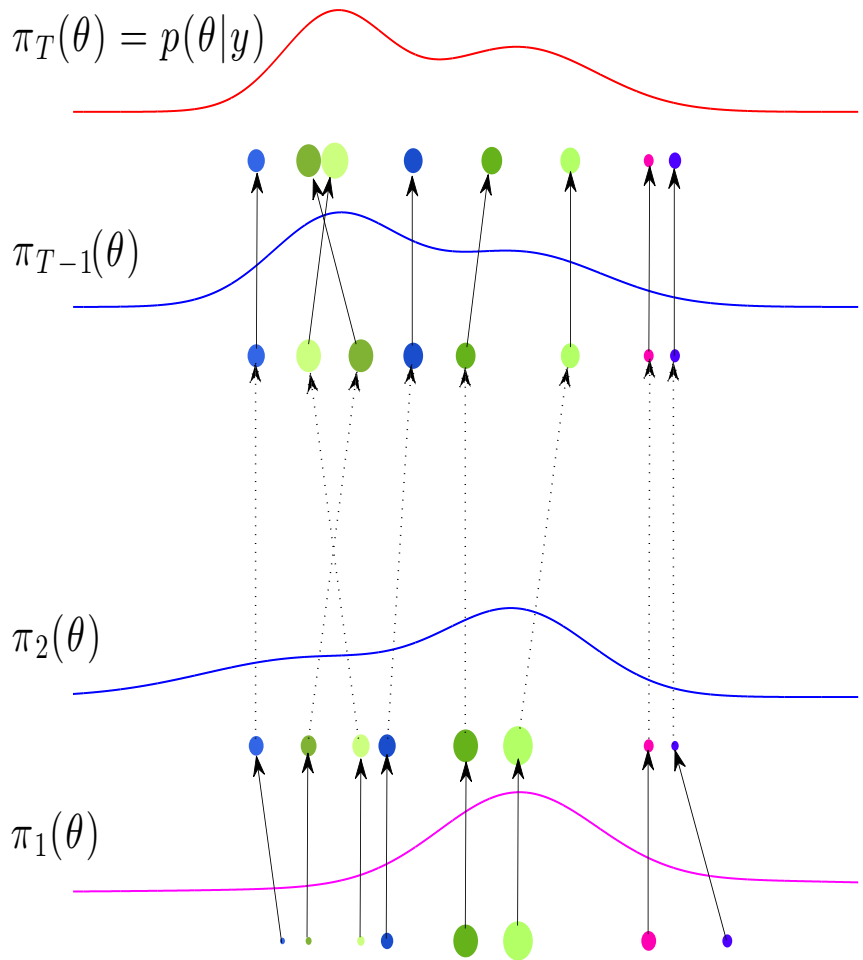
1. Study more robust and efficient Monte Carlo technique : SMC sampler.
2. Algorithm improvement by proposing new strategies to reduce the variance of the estimator
3. Applications to some challenging signal processing problems.

Joint work with Gareth Peters, T.L. Thu Nguyen

SMC Sampler : Main idea

Idea 1 : Design an **artificial** sequence of annealed distributions $\{\pi_t\}_{1 \leq t \leq T}$ from a distribution easy to sample from to the posterior of interest.

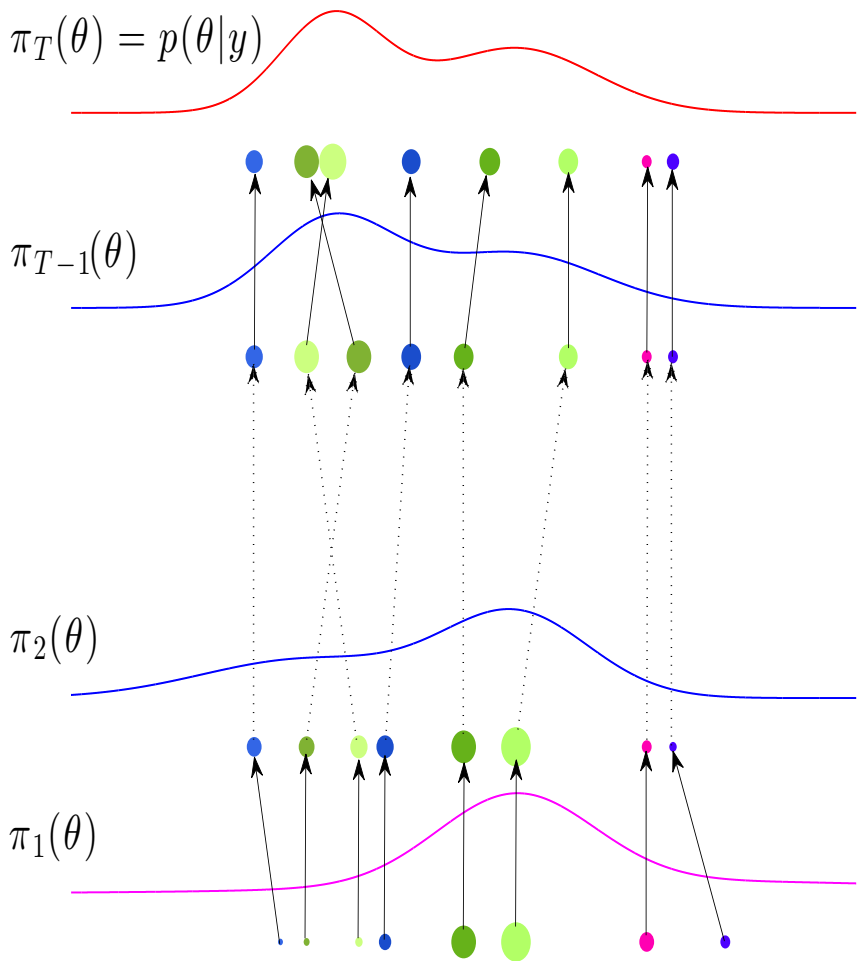
Idea 2 : Propagate a cloud of weighted random samples to approximate each distribution by combining IS and MCMC advantages.



SMC Sampler : Main idea

Idea 1 : Design an **artificial** sequence of annealed distributions $\{\pi_t\}_{1 \leq t \leq T}$ from a distribution easy to sample from to the posterior of interest.

Idea 2 : Propagate a cloud of weighted random samples to approximate each distribution by combining IS and MCMC advantages.



Methodology

At time 1 : Select π_1 which is easy to approximate by importance distribution η_1.

$$\vdots$$

At time t :

1. Propagate $\left\{\boldsymbol{\theta}_{t-1}^{(m)}\right\}_{m=1}^{N}$ by *mutation (MCMC) kernel* $\mathcal{K}_t(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)$ to obtain $\left\{\boldsymbol{\theta}_t^{(m)}\right\}_{m=1}^{N}$.
2. Correct using importance weights :

$$W_t(\boldsymbol{\theta}_t^{(m)}) = \frac{\gamma_t(\boldsymbol{\theta}_t^{(m)})}{\eta_t(\boldsymbol{\theta}_t^{(m)})}$$

However

$$\eta_t(\boldsymbol{\theta}_t) = \int_E \eta_{t-1}(\boldsymbol{\theta}_{t-1})\mathcal{K}_t(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)d\boldsymbol{\theta}_{t-1}$$

is typically not available.

$\Rightarrow$ Cannot directly use Importance Sampling.

Methodology

Solution : Perform importance sampling on extended space by introducing a sequence of extended probability distributions $\{\tilde{\pi}_t\}_{t=1}^T$ on E^t admitting $\{\pi_t\}_{t=1}^T$ as marginals

$$\tilde{\pi}_t(\boldsymbol{\theta}_{1:t}) = \frac{\tilde{\gamma}_t(\boldsymbol{\theta}_{1:t})}{Z_t}$$

$$\tilde{\gamma}_t(\boldsymbol{\theta}_{1:t}) = \gamma_t(\boldsymbol{\theta}_t) \prod_{k=1}^{t-1} \mathcal{L}_k(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_k)$$

in which $\mathcal{L}_t(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t)$ termed *backward* Markov kernels.

$\Rightarrow$ Allow the use of IS without computing $\eta_t(\boldsymbol{\theta}_t)$.

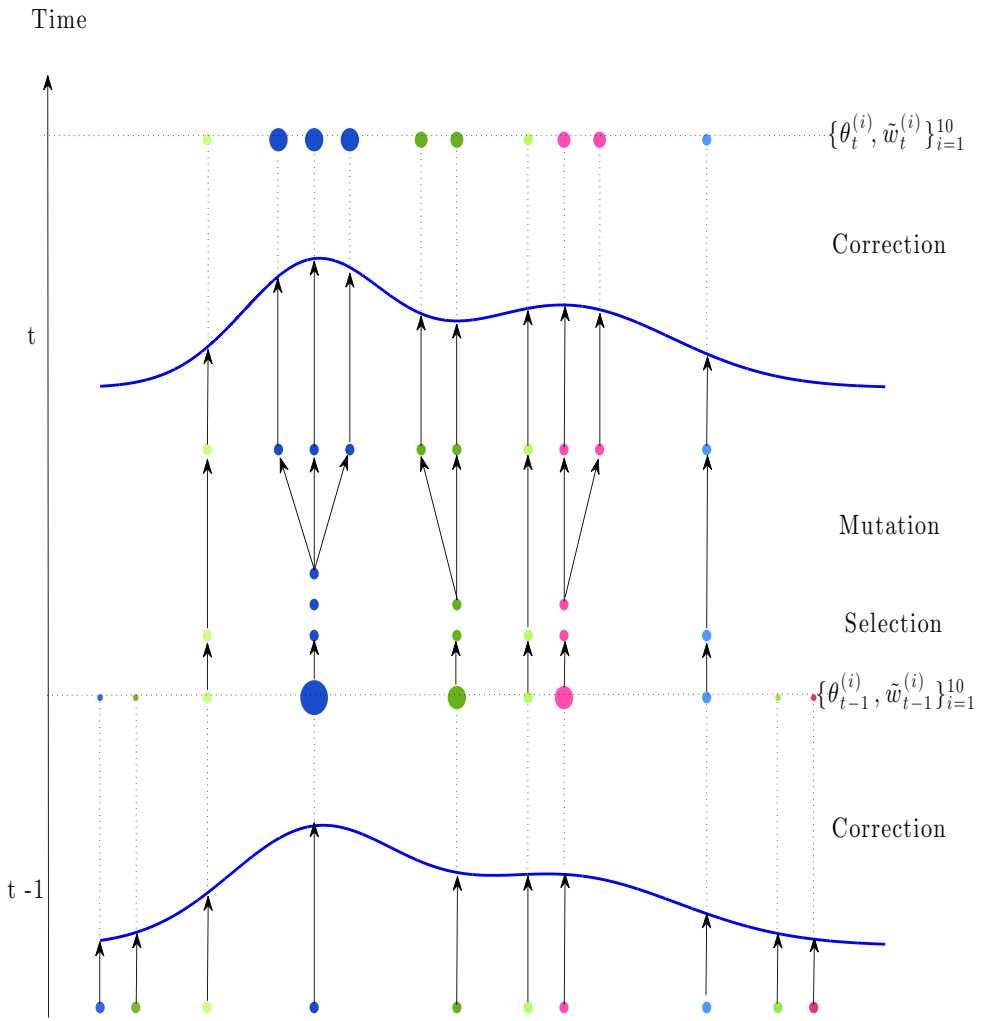
$$W_t^{(m)} \propto \frac{\tilde{\pi}_t(\boldsymbol{\theta}_{1:t}^{(m)})}{\eta_t(\boldsymbol{\theta}_{1:t}^{(m)})} \propto w_t(\boldsymbol{\theta}_{t-1}^{(m)}, \boldsymbol{\theta}_t^{(m)}) W_{t-1}^{(m)}$$

where incremental weights

$$w_t(\boldsymbol{\theta}_{t-1}^{(m)}, \boldsymbol{\theta}_t^{(m)}) = \frac{\gamma_t(\boldsymbol{\theta}_t^{(m)}) \mathcal{L}_{t-1}(\boldsymbol{\theta}_t^{(m)}, \boldsymbol{\theta}_{t-1}^{(m)})}{\gamma_{t-1}(\boldsymbol{\theta}_{t-1}^{(m)}) \mathcal{K}_t(\boldsymbol{\theta}_{t-1}^{(m)}, \boldsymbol{\theta}_t^{(m)})}$$

Algorithm summary

1. *Mutation*, particles moved from $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_t$ via a MCMC mutation kernel ;
2. *Correction*, particles are reweighted with respect to π_t ;
3. *Selection*, resampling weighted particles reduce the variability of the importance weights.



SMC Sampler

1: Initialize particle system

2: Sample $\left\{\boldsymbol{\theta}_1^{(m)}\right\}_{m=1}^{N} \sim \eta_1(\cdot)$ and compute $\widetilde{W}_1^{(m)} = \left(\frac{\gamma_1(\boldsymbol{\theta}_1^{(m)})}{\eta_1(\boldsymbol{\theta}_1^{(m)})}\right)\left[\sum_{j=1}^{N}\frac{\gamma_1(\boldsymbol{\theta}_1^{(j)})}{\eta_1(\boldsymbol{\theta}_1^{(j)})}\right]^{-1}$ and do resampling if $\mathbb{ESS} < \overline{\mathbb{ESS}}$

3: **for** $t = 2, \ldots, T$ **do**

4: Mutation : for each $m = 1, \ldots, N$: Sample $\boldsymbol{\theta}_t^m \sim \mathcal{K}_t(\boldsymbol{\theta}_{t-1}^{(m)}; \cdot)$ where $\mathcal{K}_t(\cdot;\cdot)$ is a $\pi_t(\cdot)$ invariant Markov kernel.

5: Computation of the weights : for each $m = 1, \ldots, N$

$$W_t^{(m)} = \widetilde{W}_{t-1}^{(m)} \frac{\gamma_t(\boldsymbol{\theta}_t^{(m)})\mathcal{L}_{t-1}(\boldsymbol{\theta}_t^{(m)}, \boldsymbol{\theta}_{t-1}^{(m)})}{\gamma_{t-1}(\boldsymbol{\theta}_{t-1}^{(m)})\mathcal{K}_t(\boldsymbol{\theta}_{t-1}^{(m)}, \boldsymbol{\theta}_t^{(m)})}$$

Normalization of the weights : $\widetilde{W}_t^{(m)} = W_t^{(m)}\left[\sum_{j=1}^{N} W_t^{(j)}\right]^{-1}$

6: Selection : if $ESS < \overline{\mathbb{ESS}}$ then Resample

7: **end for**

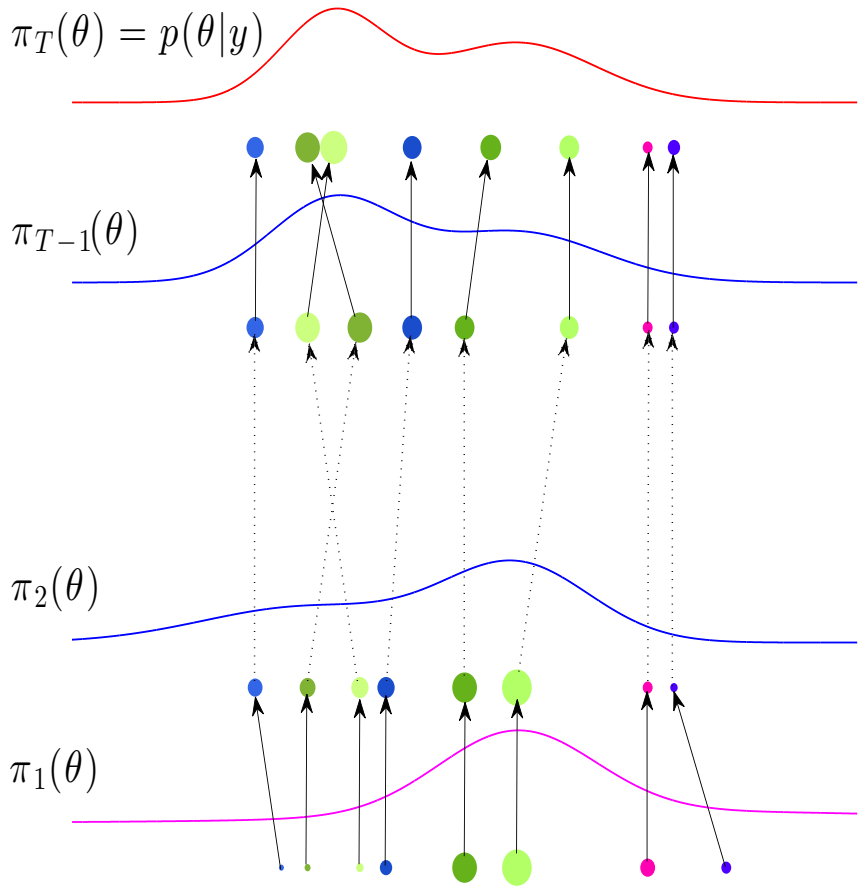
Advantages

1. No Burn-in period.
2. Framework that allowed to use interacting parallel MCMC.
3. Flexible choice of forward kernel $\mathcal{K}_t$.
4. Well suited for the computation of Bayesian evidence : unbiased estimate of normalizing constant.

$\Rightarrow$ **Promising alternative to standard MCMC methods.**

Challenging problems

1. **How to choose the sequence of target distributions ?**
2. **How to optimize and reuse all the particles generated through all SMC iterations ?**



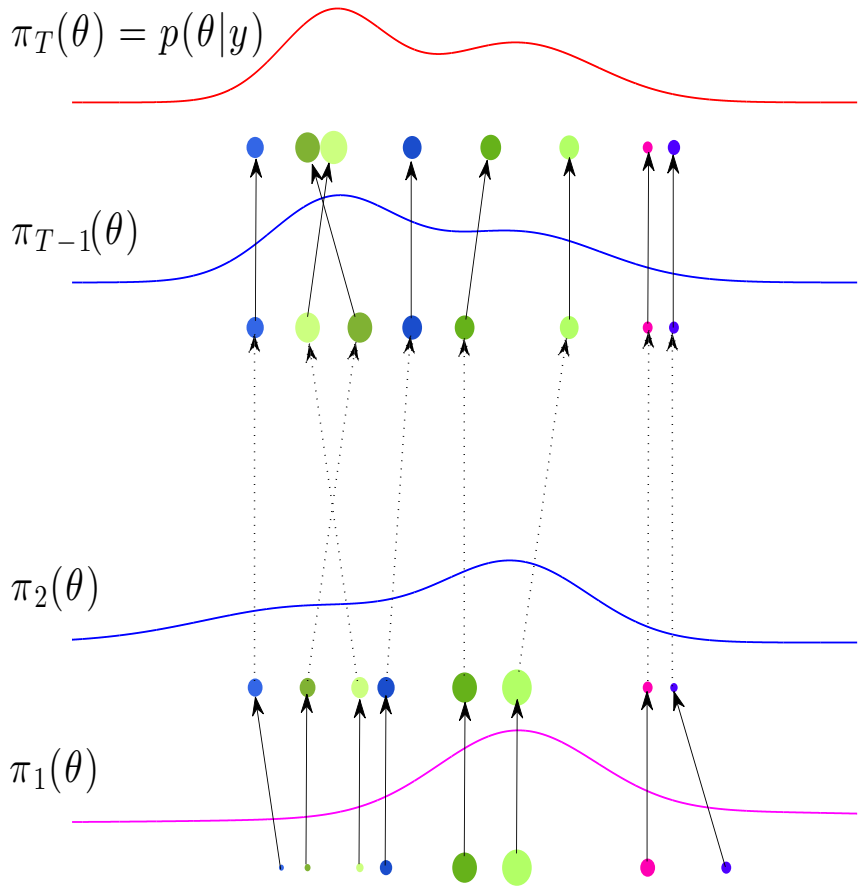
Outline

1 Introduction
- Context
- Traditional Monte Carlo methods
- SMC Samplers

2 Variance reduction schemes for SMC samplers
- Adaptive sequence of target distributions
- Recycling schemes
- Conclusion

Challenging problems

1. **How to choose the sequence of target distributions ?**
2. How to optimize and reuse all the particles generated through all SMC iterations.



Typical choice of the target sequence

Utilize *likelihood tempered* target sequence ([Neal, 2001])

$$\pi_t(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta})^{\phi_t}$$

$\{\phi_t\}$: non-decreasing temperature schedule satisfies $\phi_0 = 0$ and $\phi_T = 1$.

$\Rightarrow$ sample initially from the prior distribution $\pi_0 = p(\boldsymbol{\theta})$ and gradually increase the effect of likelihood in order to obtain the approximation of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$.

Related Works

Idea : Automatically approximate discrepancy, $\varrho_t = \phi_t - \phi_{t-1}$, between π_t and π_{t-1}.

1. [Jasra et al., 2011] : Based on controlling the rate of $\mathbb{ESS}_t$:

   $\mathbb{ESS}_t$: empirical measure of the discrepancy η_t and π_t.

2. [Zhou et al., 2013] : Based on controlling the rate of $\mathbb{CESS}_t$:

   $\mathbb{CESS}_t$ - variant of $\mathbb{ESS}_t$: empirical measure of the discrepancy between π_t and π_{t-1}.

Pros : Easy to implement.

Cons : On-line scheme : one step ahead (not global) optimization

$\Rightarrow$ Impossible to control the complexity of the algorithm.

Our Idea : Propose an adaptive scheme based on global optimization of $\{\phi_t\}$ before running SMC samplers.

Related Works

Idea : Automatically approximate discrepancy, $\varrho_t = \phi_t - \phi_{t-1}$, between π_t and π_{t-1}.

1. [Jasra et al., 2011] : Based on controlling the rate of $\mathbb{ESS}_t$:

   $\mathbb{ESS}_t$: empirical measure of the discrepancy η_t and π_t.

2. [Zhou et al., 2013] : Based on controlling the rate of $\mathbb{CESS}_t$:

   $\mathbb{CESS}_t$ - variant of $\mathbb{ESS}_t$: empirical measure of the discrepancy between π_t and π_{t-1}.

Pros : Easy to implement.

Cons : On-line scheme : one step ahead (not global) optimization

$\Rightarrow$ Impossible to control the complexity of the algorithm.

Our Idea : Propose an adaptive scheme based on global optimization of $\{\phi_t\}$ before running SMC samplers.

Related Works

Idea : Automatically approximate discrepancy, $\varrho_t = \phi_t - \phi_{t-1}$, between π_t and π_{t-1}.

1. [Jasra et al., 2011] : Based on controlling the rate of $\mathbb{ESS}_t$:

   $\mathbb{ESS}_t$: empirical measure of the discrepancy η_t and π_t.

2. [Zhou et al., 2013] : Based on controlling the rate of $\mathbb{CESS}_t$:

   $\mathbb{CESS}_t$ - variant of $\mathbb{ESS}_t$: empirical measure of the discrepancy between π_t and π_{t-1}.

Pros : Easy to implement.

Cons : On-line scheme : one step ahead (not global) optimization

$\Rightarrow$ Impossible to control the complexity of the algorithm.

Our Idea : Propose an adaptive scheme based on global optimization of $\{\phi_t\}$ before running SMC samplers.

Asymptotic Convergence results

Goal : Derive from [Del Moral et al., 2006] some specified expression of asymptotic variance to easily understand the impact of sequence of target distributions on the accuracy of the SMC sampler estimate.

Assumptions for these derivations :

- Forward kernel which mixes perfectly, i.e. :

$$\mathcal{K}_t(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t) = \pi_t(\boldsymbol{\theta}_t)$$

- Backward Kernel typically used when MCMC kernel is used as forward kernel :

$$\mathcal{L}_{t-1}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) = \frac{\pi_t(\boldsymbol{\theta}_{t-1})\mathcal{K}_t(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)}{\pi_t(\boldsymbol{\theta}_t)}$$

Conclusion :

- The asymptotic variance is reduced by conducting resampling before sampling $\Rightarrow$ Preferable to do resampling before the sampling stage.
- The asymptotic variance is a function of the dissimilarity between two successive distribution in the sequence.

Impact of the cooling schedule

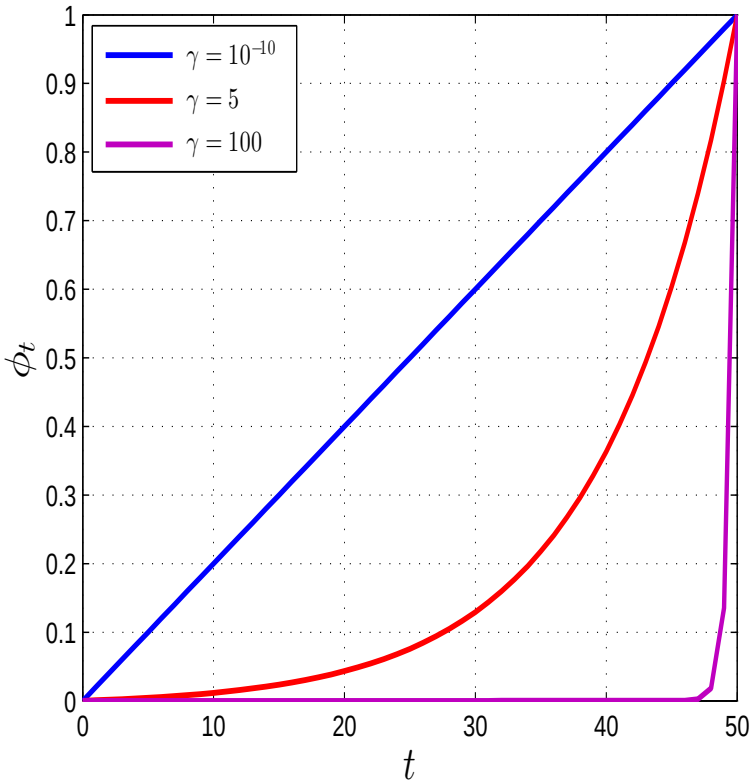
Model : Linear and Gaussian Model

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\boldsymbol{H\theta}, \boldsymbol{\Sigma}_y)$$

$$\Rightarrow \; p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

Parametric cooling temperature :

$$\phi_t = \frac{\exp(\gamma t/T) - 1}{\exp(\gamma) - 1}$$



$$\gamma = 10^{-10} \quad \rightarrow \mathbb{V}\text{ar}(\hat{Z}) = 2.97$$
$$\gamma = 5 \quad \rightarrow \mathbb{V}\text{ar}(\hat{Z}) = 0.3539$$
$$\gamma = 100 \quad \rightarrow \mathbb{V}\text{ar}(\hat{Z}) = 23.3$$

$\Rightarrow$ Choice of the cooling schedule is crucial !

Proposed approach

Proposed criterion : Take the sequence of distributions which minimizes the variance of normalizing constant.

Goal : Find optimal $\left\{\widehat{\phi}_t\right\}_{1\leq t\leq T}$ satisfies

$$\left\{\widehat{\phi}_1,\ldots,\widehat{\phi}_T\right\} = \underset{\phi_1,\ldots,\phi_T}{\arg\min} \underbrace{\sum_{t=1}^{T-1}\int \frac{\pi_{t+1}^2(\boldsymbol{\theta}_t)}{\pi_t(\boldsymbol{\theta}_t)} d\boldsymbol{\theta}_t - (T-1)}_{N\mathbb{V}\mathsf{ar}\{\hat{p}(\mathbf{y})\}} \tag{1}$$

which is related to the Rényi divergence

$$D_\alpha(f_1||f_2) = \frac{1}{\alpha-1}\log\int f_1^\alpha(x) f_2^{1-\alpha}(x)dx \geq 0 \tag{2}$$

Generally impossible to solve analytically !

Proposed approach

Proposed Solution : Avoid integral approximation by approximate the T artificial target distributions, π_t for $t = 1, \cdots, T$ by a multivariate normal distribution, i.e. :

$$\begin{aligned}\pi_t(\boldsymbol{\theta}) &\propto p(\mathbf{y}|\boldsymbol{\theta})^{\phi_t} p(\boldsymbol{\theta}) \\ &\approx \mathcal{N}(\boldsymbol{\theta}|\mu_t, \Sigma_t)\end{aligned}$$

$\Rightarrow$ enable to obtain the analytic expression for the asymptotic variance of normalizing constant.

Proposition : to obtain efficient algorithm by using :

1. Laplace's method or moment matching method.
2. Tempered multivariate normal distribution is proportional to multivariate normal distribution.
3. Product of 2 multivariate normal distributions is a multivariate normal distribution.

Proposed approach

Proposed Solution : Avoid integral approximation by approximate the T artificial target distributions, π_t for $t = 1, \cdots, T$ by a multivariate normal distribution, i.e. :

$$\begin{aligned}\pi_t(\boldsymbol{\theta}) &\propto p(\mathbf{y}|\boldsymbol{\theta})^{\phi_t} p(\boldsymbol{\theta}) \\ &\approx \mathcal{N}(\boldsymbol{\theta}|\mu_t, \Sigma_t)\end{aligned}$$

$\Rightarrow$ enable to obtain the analytic expression for the asymptotic variance of normalizing constant.

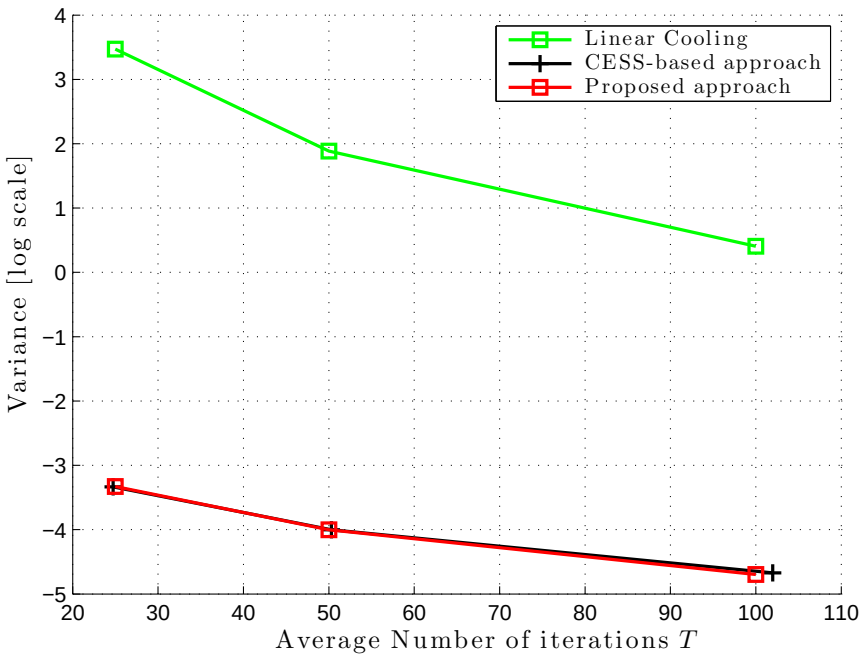
Pros :

- Global optimization
  $\hookrightarrow$ Obtain the complete view of cooling schedule performance before starting the SMC sampler.
- Empirically reduce the variance of Bayes evidence
  $\hookrightarrow$ Apply to model selection problem.

Cons : Based on Gaussian approximation of sequence of target distributions.

Illustration

Normal and linear model



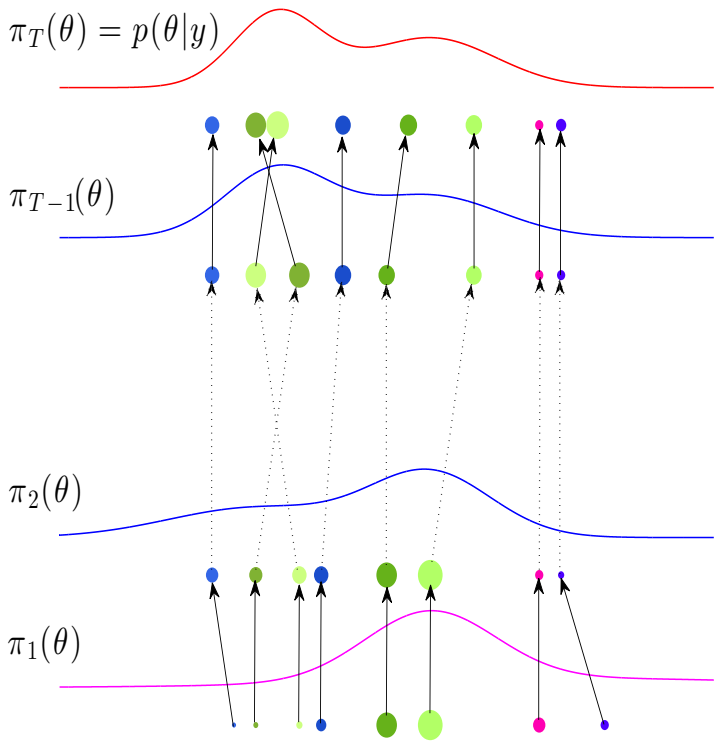
- Significant gain vs Linear cooling.
- Similar performance compared to $\mathbb{CESS}$-based approach

BUT

can totally control the complexity of the algorithm.

Challenging problems

1. How to choose the sequence of target distributions ?
2. **How to optimize and reuse all the particles generated through all SMC iterations ?**



Recycling Idea

Generally, only the weighted random samples from the last iteration are used :

$$\mathbb{E}_\pi[h(\boldsymbol{\theta})] \approx \sum_{j=1}^{N} \tilde{W}_T^{(j)} h(\boldsymbol{\theta}_T^{(j)})$$

BUT, we have generated T collections $\{\tilde{W}_t^{(j)}; \boldsymbol{\theta}_t^{(j)}\}_{j=1}^N$ that approximates :

$$\pi_t(\boldsymbol{\theta}) \approx \sum_{j=1}^{N} \tilde{W}_t^{(j)} \delta_{\boldsymbol{\theta}_t^{(j)}}(d\boldsymbol{\theta})$$

How can we combine all these collections to improve the estimator's property ?

$$\hookrightarrow \mathbb{E}_\pi[h(\boldsymbol{\theta})] \approx \sum_{t=1}^{T} \sum_{j=1}^{N} \tilde{W}_{COMBI,t}^{(j)} h(\boldsymbol{\theta}_t^{(j)})$$

Recycling Idea

Generally, only the weighted random samples from the last iteration are used :

$$\mathbb{E}_\pi[h(\boldsymbol{\theta})] \approx \sum_{j=1}^{N} \tilde{W}_T^{(j)} h(\boldsymbol{\theta}_T^{(j)})$$

BUT, we have generated T collections $\{\tilde{W}_t^{(j)}; \boldsymbol{\theta}_t^{(j)}\}_{j=1}^N$ that approximates :

$$\pi_t(\boldsymbol{\theta}) \approx \sum_{j=1}^{N} \tilde{W}_t^{(j)} \delta_{\boldsymbol{\theta}_t^{(j)}}(d\boldsymbol{\theta})$$

How can we combine all these collections
to improve the estimator's property ?

$$\hookrightarrow \mathbb{E}_\pi[h(\boldsymbol{\theta})] \approx \sum_{t=1}^{T} \sum_{j=1}^{N} \tilde{W}_{COMBI,t}^{(j)} h(\boldsymbol{\theta}_t^{(j)})$$

Recycling Idea

Generally, only the weighted random samples from the last iteration are used :

$$\mathbb{E}_\pi[h(\boldsymbol{\theta})] \approx \sum_{j=1}^{N} \tilde{W}_T^{(j)} h(\boldsymbol{\theta}_T^{(j)})$$

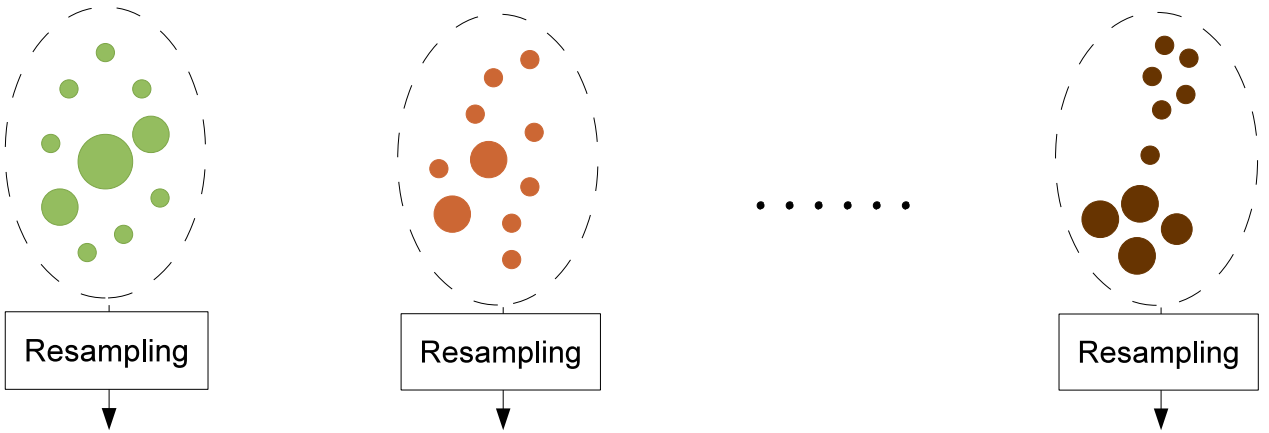
BUT, we have generated T collections $\{\tilde{W}_t^{(j)}; \boldsymbol{\theta}_t^{(j)}\}_{j=1}^N$ that approximates :

$$\pi_t(\boldsymbol{\theta}) \approx \sum_{j=1}^{N} \tilde{W}_t^{(j)} \delta_{\boldsymbol{\theta}_t^{(j)}}(d\boldsymbol{\theta})$$

How can we combine all these collections
to improve the estimator's property ?

$$\hookrightarrow \mathbb{E}_\pi[h(\boldsymbol{\theta})] \approx \sum_{t=1}^{T} \sum_{j=1}^{N} \tilde{W}_{COMBI,t}^{(j)} h(\boldsymbol{\theta}_t^{(j)})$$

Proposed Recycling Schemes

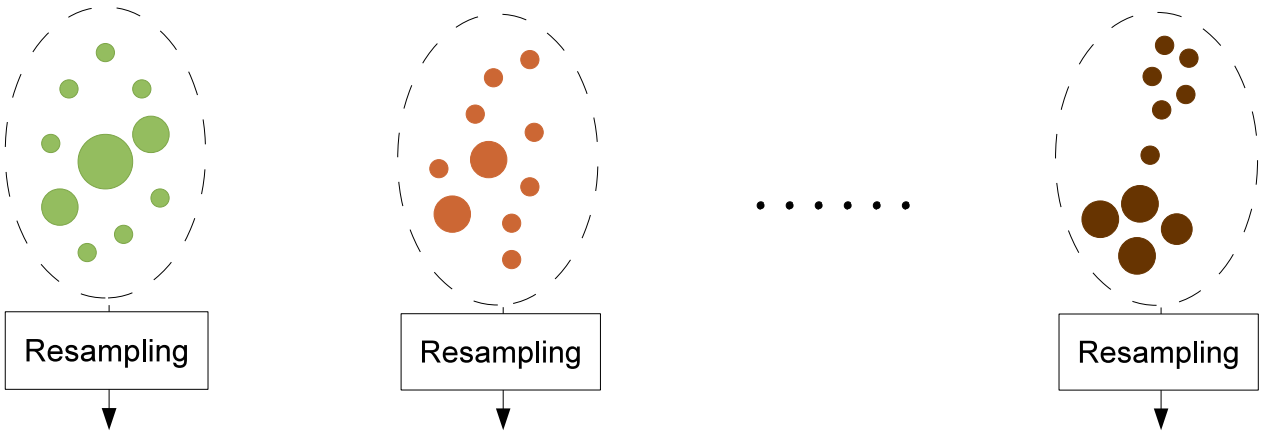


Idea : Correction of the random samples by an importance weighting step.

Prop 1 : Adapt to SMC sampler an idea proposed by [Gramacy et al., 2010] for MCMC.

1. Correction step : $W_{ESS,t}^{(j)} = \frac{\gamma(\tilde{\boldsymbol{\theta}}_t^{(j)})}{\gamma_t(\tilde{\boldsymbol{\theta}}_t^{(j)})}$.
2. Compute (local) estimator : $\hat{h}_t = \sum_{j=1}^N \tilde{W}_{ESS,t}^{(j)} h(\tilde{\boldsymbol{\theta}}_t^{(j)})$.
3. Combine these estimator : $\hat{h} = \sum_{t=1}^T \lambda_t \hat{h}_t$ such that λ_t optimize the ESS of the global population.

Proposed Recycling Schemes



$\{\tilde{\boldsymbol{\theta}}_1^{(j)}\}_{j=1}^N \sim \pi_1$ $\{\tilde{\boldsymbol{\theta}}_2^{(j)}\}_{j=1}^N \sim \pi_2$ *[Asympt.]* $\{\tilde{\boldsymbol{\theta}}_T^{(j)}\}_{j=1}^N \sim \pi$

Idea : Correction of the random samples by an importance weighting step.

Prop 2 : Use the deterministic mixture idea developed in [Veach and Guibas, 1995].
↪ Consider the entire available population coming from a "mixture".

1. Correction : $W_{DeMix,t}^{(j)} = \frac{\gamma(\tilde{\boldsymbol{\theta}}_t^{(j)})}{\sum_{t=1}^T c_t \pi_t(\tilde{\boldsymbol{\theta}}_t^{(j)})}$ with $c_t = \frac{1}{T}$ and $\pi_t(\cdot) = \frac{\gamma_t(\cdot)}{Z_t}$

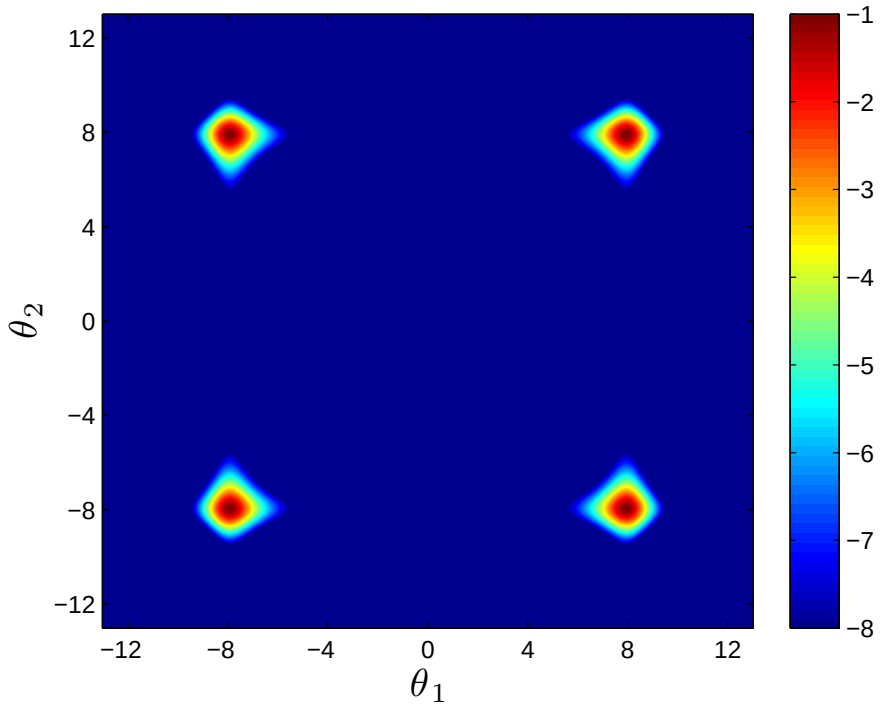
Prop Z_t is replaced by the unbiased estimate given by the SMC sampler.

2. Combine $\hat{h} = \sum_{t=1}^T \sum_{j=1}^N \tilde{W}_{DeMix,t}^{(j)} h(\tilde{\boldsymbol{\theta}}_t^{(j)})$

Numerical Simulations

Multimodal posterior distribution

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
p(\mathbf{y}|\boldsymbol{\theta}) &= \frac{\Gamma\left(\frac{\nu+n_{\mathbf{y}}}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{n_{\mathbf{y}}/2}}|\boldsymbol{\Sigma}_l|^{-\frac{1}{2}}\left[1+\frac{[\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\theta}]^T\boldsymbol{\Sigma}_l^{-1}[\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\theta}]}{\nu}\right]^{-\frac{(\nu+n_{\mathbf{y}})}{2}}
\end{aligned}
$$



$\log p(\boldsymbol{\theta}|\mathbf{y})$

Prior

$\boldsymbol{\mu} = \mathbf{0}_{2\times 1}$, $\boldsymbol{\Sigma} = 20\boldsymbol{I}_2$

Likelihood

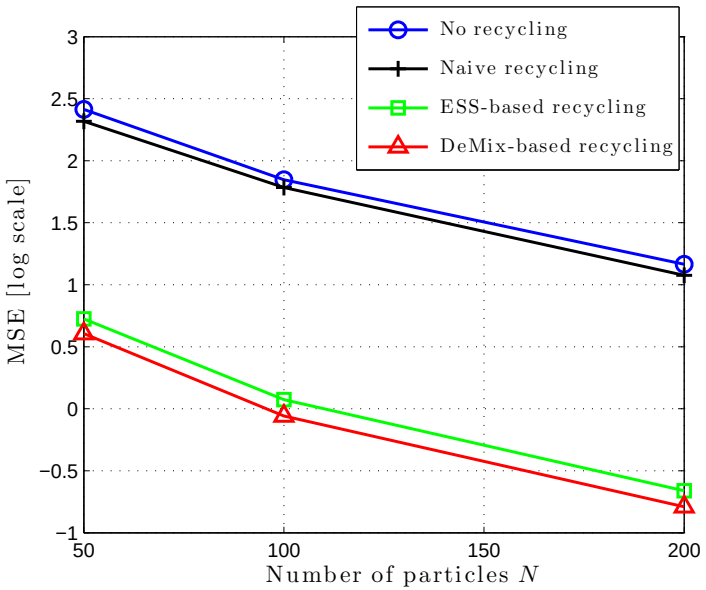
$\nu = 7$, $\boldsymbol{\Sigma}_l = 0.1\boldsymbol{I}_4$

$$
\boldsymbol{H} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}^T
$$

Observations

$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix}^T = \begin{bmatrix} 8 & -8 & 8 & -8 \end{bmatrix}^T$

Numerical Simulations



$T = 100.$

Proposed Recycling schemes
- Gain :

96% reduction
compared to classical estimator,

94% reduction
compared to naïve scheme.

$\Downarrow$

Significant improvement

Numerical Simulations

		No Recycling	Naive Recycling	ESS-based Recycling	DeMix Recycling
$T =$ 25 Iter.	$N = 50$	0.1607 (0.0615)	0.1571 (0.0612)	0.0861 (0.0417)	0.0839 (0.0390)
	$N = 100$	0.1048 (0.0331)	0.1026 (0.0325)	0.0596 (0.0216)	0.0578 (0.0203)
	$N = 200$	0.0825 (0.0299)	0.0809 (0.0296)	0.0494 (0.0201)	0.0476 (0.0188)
$T =$ 50 Iter.	$N = 50$	0.1641 (0.0651)	0.1499 (0.0649)	0.0678 (0.0289)	0.0655 (0.0274)
	$N = 100$	0.1126 (0.0392)	0.1020 (0.0385)	0.0517 (0.0215)	0.0500 (0.0204)
	$N = 200$	0.0878 (0.0378)	0.0803 (0.0369)	0.0404 (0.0147)	0.0396 (0.0139)
$T =$ 100 Iter.	$N = 50$	0.1795 (0.0883)	0.1528 (0.0845)	0.0623 (0.0420)	0.0604 (0.0393)
	$N = 100$	0.1261 (0.0580)	0.1092 (0.0570)	0.0475 (0.0229)	0.0459 (0.0214)
	$N = 200$	0.0901 (0.0329)	0.0761 (0.0326)	0.0352 (0.0141)	0.0342 (0.0135)

Table: Comparison of recycling schemes for the accuracy to approximate the posterior distribution $p(\theta_1|\mathbf{y})$ in terms of the Kolmogorov-Smirnov distance (mean and standard deviation in parentheses).

Conclusion and Future works

Conclusion

- Derive simple form for the asymptotic variances for SMC samplers estimate under some assumptions.
- Propose novel strategy to automatically and adaptively choose the sequence of target distribution.
- Propose two different approaches to recycle all past simulated particles for the approximation of posterior distribution.
- Obtain significant improvement by using both proposed strategies.

Future Work

- Theoretical Analysis of the proposed schemes

Del Moral, P., Doucet, A. and Jasra, A. (2006).

Sequential Monte Carlo samplers.
Journal of the Royal Statistical Society : Series B (Statistical Methodology) ***68***, 411–436.

Geyer, C. J. (1991).

Markov chain Monte Carlo maximum likelihood.
In Computing Science and Statistics : Proceedings of the 23rd Symposium on the Interface vol. 1, pp. 156–163,.

Gramacy, R., Samworth, R. and King, R. (2010).

Importance tempering.
Statistics and Computing ***20***, 1–7.

Jasra, A., Stephens, D. A., Doucet, A. and Tsagaris, T. (2011).

Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo.
Scandinavian Journal of Statistics ***38***, 1–22.

Liang, F. and Wong, W. (2001).

Real parameter evolutionary Monte Carlo with applications to Bayesian mixture models.
Journal of the American Statistical Association ***96***.

Liang, F. and Wong, W. H. (2000).

Evolutionary Monte Carlo : Applications to C_p Model Sampling and Change Point Problem.
Statistica Sinica ***10***, 317–342.

Neal, R. M. (2001).

Annealed importance sampling.
Statistics and Computing ***11***, 125–139.

Ruanaidh, Ó., Joseph, J. and Fitzgerald, W. J. (1996).

Numerical Bayesian methods applied to signal processing.
Springer.

Veach, E. and Guibas, L. (1995).

Optimally combining sampling techniques for Monte-Carlo rendering.
In Proc. SIGGRAPH'95 pp. 419–428,.

Zhou, Y., Johansen, A. M. and Aston, J. A. (2013).

Towards Automatic Model Comparison : An Adaptive Sequential Monte Carlo Approach.

arXiv preprint arXiv :1303.3123 .