Crash course in simulation

Target

Monte Carlo basics

MCMC

MCMC difficulties

Importance Sampling

Pros & cons

Population Monte Carlo Algorithm

Illustrations

Further advances

General purpose

Given a density $\pi$ known up to a normalizing constant, and a function $h$, compute

$$\Pi(h) = \int h(x)\pi(x)\mu(dx) = \frac{\int h(x)\tilde{\pi}(x)\mu(dx)}{\int \tilde{\pi}(x)\mu(dx)}$$

when $\int h(x)\tilde{\pi}(x)\mu(dx)$ is intractable.
Monte Carlo basics

Generate an iid sample \( x_1, \ldots, x_N \) from \( \pi \) and estimate \( \Pi(h) \) by

\[
\hat{\Pi}_N^{MC}(h) = N^{-1} \sum_{i=1}^{N} h(x_i).
\]

**LLN:** \( \hat{\Pi}_N^{MC}(h) \) as \( N \to \infty \)

If \( \Pi(h^2) = \int h^2(x) \pi(x) \mu(dx) < \infty \),

**CLT:** \( \sqrt{N} \left( \hat{\Pi}_N^{MC}(h) - \Pi(h) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Pi \{ [h - \Pi(h)]^2 \}) \).

**Caveat**

Often impossible or inefficient to simulate directly from \( \Pi \)

MCMC basics

Generate \( x^{(1)}, \ldots, x^{(T)} \) ergodic Markov chain \((x_t)_{t \in \mathbb{N}}\) with stationary distribution \( \pi \)

Estimate \( \Pi(h) \) by

\[
\hat{\Pi}_N^{MCMC}(h) = N^{-1} \sum_{i=T-N}^{T} h \left( x^{(i)} \right).
\]

[Robert & Casella, 2004]

A generic algorithm

**Metropolis–Hastings algorithm:**

Given \( x^{(t)} \) and a proposal \( q(\cdot|x^{(t)}) \),

1. Generate \( Y_t \sim q(y|x^{(t)}) \)
2. Take

\[
X^{(t+1)} = \begin{cases} 
Y_t & \text{with prob. } \rho(x^{(t)}, Y_t), \\
x^{(t)} & \text{with prob. } 1 - \rho(x^{(t)}, Y_t),
\end{cases}
\]

where

\[
\rho(x, y) = \min \left\{ \frac{\pi(y) q(x|y)}{\pi(x) q(y|x)}, 1 \right\}
\]

A generic MH: RWMH

Choose for proposal the **random walk**

\[
q(x|y) = g(y - x) = g(x - y)
\]

- local exploration of the space
- posterior-ratio acceptance probability
- only requires a scale but does require a scale!
- often targeted at **optimal acceptance rate**
Example (Mixture models)

\[
\pi(\theta | x) \propto \prod_{j=1}^{n} \left( \sum_{\ell=1}^{k} p_{\ell} f(x_j | \mu_\ell, \sigma_\ell) \right) \pi(\theta)
\]

Metropolis-Hastings proposal:

\[
\theta(t+1) = \begin{cases} 
\theta(t) + \omega \varepsilon(t) & \text{if } u(t) < \rho(t) \\
\theta(t) & \text{otherwise}
\end{cases}
\]

where

\[
\rho(t) = \frac{\pi(\theta(t) + \omega \varepsilon(t) | x)}{\pi(\theta(t) | x)} \wedge 1
\]

and \(\omega\) scaled for good acceptance rate.

MCMC difficulties

Trapping modes may remain undetected
Convergence to the stationary distribution can be very slow or intractable
Difficult adaptivity
A wee problem with Gibbs on mixtures

Gibbs stuck at the wrong mode

Gibbs started at random

Trapping modes may remain undetected
Convergence to the stationary distribution can be very slow or intractable
Difficult adaptivity

Example (Bimodal target)

Density

\[
f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \left( \frac{4(x - .3)^2 + .01}{4(1 + (.3)^2) + .01} \right).
\]

and use of random walk Metropolis–Hastings algorithm with variance .04
Evaluation of the missing mass by

\[
\sum_{t=1}^{T-1} [\theta_{(t+1)} - \theta_{(t)}] f(\theta_{(t)})
\]

[Philippe & Robert, 2001]
MCMC difficulties

Trapping modes may remain undetected
Convergence to the stationary distribution can be very slow or intractable
Difficult adaptivity

Simple adaptive MCMC is not possible

Example (Poly $t$ distribution)
Consider a $t$-distribution $T(3, \theta, 1)$ sample $(x_1, \ldots, x_n)$ with a flat prior $\pi(\theta) = 1$
If we try fit a normal proposal from empirical mean and variance of the chain so far,

$$
\mu_t = \frac{1}{t} \sum_{i=1}^{t} \theta^{(i)} \quad \text{and} \quad \sigma_t^2 = \frac{1}{t} \sum_{i=1}^{t} (\theta^{(i)} - \mu_t)^2,
$$

Metropolis–Hastings algorithm with acceptance probability

$$
\prod_{j=2}^{n} \left[ \frac{\nu + (x_j - \theta^{(i)})^2}{\nu + (x_j - \xi)^2} \right]^{-(\nu+1)/2} \exp \frac{- (\mu_t - \theta^{(i)})^2}{2\sigma_t^2} \exp \frac{- (\mu_t - \xi)^2}{2\sigma_t^2},
$$

where $\xi \sim \mathcal{N}(\mu_t, \sigma_t^2)$.

Example (Poly $t$ distribution (2))

Invalid scheme:
- when range of initial values too small, the $\theta^{(i)}$’s cannot converge to the target distribution and concentrates on too small a support.
- long-range dependence on past values modifies the distribution of the sequence.
- using past simulations to create a non-parametric approximation to the target distribution does not work either
Adaptive scheme for a sample of $10 \ x_j \sim T_3$ and initial variances of (top) 0.1, (middle) 0.5, and (bottom) 2.5.

Sample produced by 50,000 iterations of a nonparametric adaptive MCMC scheme and comparison of its distribution with the target distribution.

Simply forget about it!

**Warning:**

One should not constantly adapt the proposal on past performances

Either adaptation ceases after a period of burnin or the adaptive scheme must be theoretically assessed on its own right...

Controlled MCMC

Optimal choice of a parameterised proposal $r(x, dy; \theta)$ against a proposal minimisation problem

$$\theta^* = \arg \min \Psi(\eta(\theta))$$

where

$$\eta(\theta) = \int_X r(\theta, x) \mu(\theta)(dx)$$

[Andrieu & Robert, 2001]

- Coerced acceptance
- Autocorrelations
- Moment matching
Two-time scale stochastic approximation

Set $\xi_i = (\eta_i, \dot{\eta}_i)$

Corresponding recursive system

$x_{i+1} \sim R(x_i, dx_{i+1}; \theta_i)$

$\xi_{i+1} = (1 - \gamma_{i+1})\xi_i + \gamma_{i+1}x_{i+1}$

$\theta_{i+1} = \theta_i - \gamma_{i+1}\dot{\eta}_i \Psi'(\eta_i)$

where $\{\gamma_i\}$ and $\{\varepsilon_i\}$ go to 0 at infinity

**Warning**

Hidden difficulties...

Importance Sampling

For $Q$ proposal distribution such that $Q(dx) = q(x)\mu(dx)$,

alternative representation

$\Pi(h) = \int h(x)q(x)q(x)\mu(dx).$

**Principle**

Generate an iid sample $x_1, \ldots, x_N \sim Q$ and estimate $\Pi(h)$ by

$\hat{\Pi}^{IS}_{Q,N}(h) = N^{-1} \sum_{i=1}^{N} h(x_i)q(x_i).$

**Then**

$\text{LLN : } \hat{\Pi}^{IS}_{Q,N}(h) \xrightarrow{as} \Pi(h)$

and if $Q((h\pi/q)^2) < \infty$,

$\text{CLT : } \sqrt{N}(\hat{\Pi}^{IS}_{Q,N}(h) - \Pi(h)) \xrightarrow{d} N(0, Q\{(h\pi/q - \Pi(h))^2\}).$

**Caveat**

If normalizing constant unknown, impossible to use $\hat{\Pi}^{IS}_{Q,N}$

Generic problem in Bayesian Statistics: $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$. 
Self-Normalised Importance Sampling

Self normalized version

\[ \hat{\Pi}_{Q,N}^{\text{SNIS}}(h) = \left( \sum_{i=1}^{N} \frac{\pi}{q}(x_i) \right)^{-1} \sum_{i=1}^{N} h(x_i) \frac{\pi}{q}(x_i). \]

**LLN**: \( \hat{\Pi}_{Q,N}^{\text{SNIS}}(h) \xrightarrow{\text{as}} \Pi(h) \)

and if \( \Pi((1 + h^2)(\pi/q)^2) < \infty \),

**CLT**: \( \sqrt{N}(\hat{\Pi}_{Q,N}^{\text{SNIS}}(h) - \Pi(h)) \xrightarrow{\text{d}} \mathcal{N}(0, \pi((\pi/q)(h - \Pi(h))^2)) \).

The quality of the SNIS approximation depends on the choice of \( Q \).

Sampling importance resampling

Importance sampling from \( g \) can also produce samples from the target \( \pi \)

[Rubin, 1987]

Theorem (Bootstraped importance sampling)

*If a sample \((x_i^*)_{1 \leq i \leq m}\) is derived from the weighted sample \((x_i, \omega_i)_{1 \leq i \leq n}\) by multinomial sampling with weights \(\bar{\omega}_i\), then*

\[ x_i^* \sim \pi(x) \]

where \(\bar{\omega}_{i,t} = \omega_{i,t} / \sum_{j=1}^{N} \omega_{j,t}\).

**Note**

Obviously, the \(x_i^*\)'s are not iid

Pros and cons of importance sampling vs. MCMC

- Production of a sample (IS) vs. a Markov chain (MCMC)
- Dependence on importance function (IS) vs. on previous value (MCMC)
- Unbiasedness (IS) vs. convergence to the true distribution (MCMC)
- Variance control (IS) vs. learning costs (MCMC)
- Recycling of past simulations (IS) vs. progressive adaptability (MCMC)
- Processing of moving targets (IS) vs. handling large dimensional problems (MCMC)
- Non-asymptotic validity (IS) vs. difficult asymptotia for adaptive algorithms (MCMC)

Population Monte Carlo Algorithm

1. Crash course in simulation
2. Population Monte Carlo Algorithm
   - Sequential importance sampling
   - Population Monte Carlo Algorithm
   - Choice of the kernels \(Q_{i,t}\)
3. Illustrations
4. Further advances
Sequential importance sampling

**Idea** Apply dynamic importance sampling to simulate a sequence of iid samples

\[ x^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)}) \overset{iid}{\approx} \pi(x) \]

where \( t \) is a simulation iteration index (at sample level)

**Sequential Monte Carlo applied to a fixed distribution \( \pi \)**

[Iba, 2000]

Iterated importance sampling

As in Markov Chain Monte Carlo (MCMC) algorithms, introduction of a *temporal dimension*:

\[ x_i^{(t)} \sim q(t|x_i^{(t-1)}) \quad i = 1, \ldots, n, \quad t = 1, \ldots \]

and

\[ \hat{I}_t \ = \ \frac{1}{n} \sum_{i=1}^{n} \varrho_i^{(t)} h(x_i^{(t)}) \]

is still unbiased for

\[ \varrho_i^{(t)} = \frac{\pi(x_i^{(t)})}{q(t|x_i^{(t-1)})} \quad i = 1, \ldots, n \]

Adaptive IS

**Fact**

IS can be generalized to encompass much more adaptive/local schemes than thought previously

**Adaptivity** means learning from experience, i.e., to design new importance sampling functions based on the performances of earlier importance sampling proposals

**Incentive**

Use previous sample(s) to learn about \( \pi \) and \( q \)

Fundamental importance equality

**Preservation of unbiasedness**

\[
\mathbb{E} \left[ h(X^{(t)}) \frac{\pi(X^{(t)})}{q_t(X^{(t)}|X^{(t-1)})} \right] \\
= \int h(x) \frac{\pi(x)}{q_t(x|y)} q_t(x|y) g(y) \, dx \, dy \\
= \int h(x) \pi(x) \, dx
\]

for any distribution \( g \) on \( X^{(t-1)} \)
**PMCA: Population Monte Carlo Algorithm**

At time $t = 0$

- Generate $(x_{i,0})_{1 \leq i \leq N} \overset{iid}{\sim} Q_0$
- Set $\omega_{i,0} = \{\pi/q_0\}(x_{i,0})$

- Generate $(J_{i,0})_{1 \leq i \leq N} \overset{iid}{\sim} M(1, (\bar{\omega}_{i,0})_{1 \leq i \leq N})$
- Set $\tilde{x}_{i,0} = x_{J_{i,0},0}$

At time $t$ ($t = 1, \ldots, T$),

- Generate $x_{i,t} \overset{iid}{\sim} Q_{i,t}(\tilde{x}_{i,t-1}, \cdot)$
- Set $\omega_{i,t} = \{\pi(x_{i,t})/q_{i,t}(\tilde{x}_{i,t-1}, x_{i,t})\}$

- Generate $(J_{i,t})_{1 \leq i \leq N} \overset{iid}{\sim} M(1, (\bar{\omega}_{i,t})_{1 \leq i \leq N})$
- Set $\tilde{x}_{i,t} = x_{J_{i,t},t}$.

Self-norm’ed weights

#### Links with sequential Monte Carlo (1)

- Hammersley & Morton’s (1954) self-avoiding random walk problem
- Wong & Liang’s (1997) and Liu, Liang & Wong’s (2001) dynamic weighting
- Chopin’s (2001) fractional posteriors for large datasets
- Rubinstein & Kroese’s (2004) cross-entropy method for rare events

#### Links with sequential Monte Carlo (2)

- West’s (1992) mixture approximation is a precursor of smooth bootstrap
- Gilks & Berzuini (2001) SIR+MCMC: the MCMC step uses a $\pi_i$ invariant kernel
- Hürzeler & Künsch’s (1998) and Stavropoulos & Titterington’s (1999) smooth bootstrap
- Warnes’ (2001) kernel coupler
- Mengersen & Robert’s (2002) “pinball sampler” (MCMC version of PMC)
- Del Moral & Doucet’s (2003) sequential Monte Carlo sampler, with Markovian dependence on the past $x^{(t)}$ but (limited) stationarity constraints

**Choice of the kernels $Q_{i,t}$**

After $T$ iterations of the previous algorithm, the PMC estimator of $\Pi(h)$ is given by

$$\hat{\Pi}^{PMC}_{N,T}(h) = \sum_{i=1}^{N} \bar{\omega}_{i,T} h(x_{i,T}).$$

or

$$\hat{\Pi}^{PMC}_{N,T}(h) = \frac{1}{N} \sum_{t=1}^{T} \sum_{i=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}).$$

Given $\mathcal{F}_{N,t-1}$, how to construct $Q_{i,t}(\mathcal{F}_{N,t-1})$?
Population Monte Carlo and adaptive sampling schemes

D kernel PMC

**Idea:**
Take for $Q_{i,t}$ a mixture of $D$ fixed transition kernels

$$\sum_{d=1}^{D} \alpha_{d}^{t} q_{d}(x, \cdot)$$

and set the weights $\alpha_{d}^{t+1}$ equal to previous survival rates

**Survival of the fittest:**
The algorithm should automatically fit the mixture to the target distribution

**An initial LLN**
Under the assumption

(A1) $\forall d \in \{1, \ldots, D\}, \Pi \otimes \Pi \{q_{d}(x, x') = 0\} = 0$

with $\gamma_{u}$ the uniform distribution on $\{1, \ldots, D\}$,

**Proposition**
If (A1) holds, for $h \in L_{\Pi \otimes \gamma_{u}}^{1}$ and every $t \geq 1$,

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}, K_{i,t}) \xrightarrow{N \to \infty} \Pi \otimes \gamma_{u}(h).$$

**Bad!!!**
Even very bad because, while

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}) \xrightarrow{N \to \infty} \Pi(h),$$

convergence to $\gamma_{u}$ implies that

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} I_{K_{i,t}=d} \xrightarrow{N \to \infty} \frac{1}{D}.$$
Population Monte Carlo and adaptive sampling schemes

Population Monte Carlo Algorithm

Choice of the kernels $Q_{i,t}$

Saved by Rao-Blackwell !!

**Idea:**

Use Rao-Blackwellisation by deconditioning the chosen kernel 

\[ \text{[Gelfand & Smith, 1990]} \]

Use the whole mixture in the importance weights

\[
\pi(x_{i,t}) \sum_{D} \alpha_{t,N}^{d} q_{d}(\tilde{x}_{i,t-1}, x_{i,t})
\]

instead of

\[
\pi(x_{i,t}) q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})
\]

and in the kernels weights $\alpha_{t,N}^{d}$

**RBDPMCA: Rao-Blackwellised $D$-kernel PMC Algorithm**

At time $t$ ($t = 1, \ldots, T$),

Generate $(K_{i,t})_{1 \leq i \leq N} \overset{iid}{\sim} \mathcal{M}(1, (\alpha_{t,N}^{d})_{1 \leq d \leq D})$

and $(x_{i,t})_{1 \leq i \leq N} \overset{iid}{\sim} Q_{i,t}(\tilde{x}_{i,t-1}, \cdot)$

Set $\omega_{i,t} = \pi(x_{i,t}) / \sum_{D} \alpha_{t,N}^{d} q_{d}(\tilde{x}_{i,t-1}, x_{i,t})$

Generate $(J_{i,t})_{1 \leq i \leq N} \overset{iid}{\sim} \mathcal{M}(1, (\bar{\omega}_{i,t})_{1 \leq i \leq N})$

and set $\tilde{x}_{i,t} = x_{J_{i,t},t}$, $\alpha_{t+1,N}^{d} = \sum_{i=1}^{N} \bar{\omega}_{i,t} \alpha_{t,N}^{d}$.

**Proposition**

Under (A1), for $h \in L_{1}^{1}$ and for every $t \geq 1$,

\[
\frac{1}{N} \sum_{k=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}) \xrightarrow{N \to \infty} \mathbb{P}(h)
\]

where $(1 \leq d \leq D)$

\[
\alpha_{t,N}^{d} \xrightarrow{N \to \infty} \alpha_{t}^{d}
\]

\[
\alpha_{t}^{d} = \alpha_{t-1}^{d} \int \left( \frac{q_{d}(x, x')}{\sum_{j=1}^{D} \alpha_{t-1}^{j} q_{j}(x, x')} \right) \mathbb{P} \otimes \mathbb{P}(dx, dx').
\]

**LLN (2) and convergence**

**Kullback divergence**

For $\alpha \in S$,

\[
\text{KL}(\alpha) = \int \left[ \log \left( \frac{\pi(x) \pi(x')}{{{\pi}(x) \sum_{d=1}^{D} \alpha_{d} q_{d}(x, x')}} \right) \right] \mathbb{P} \otimes \mathbb{P}(dx, dx')
\]

Kullback divergence between $\Pi$ and the mixture.

**Goal**

Obtain the mixture of $q_{d}$'s closest to $\Pi$ for the Kullback divergence
Population Monte Carlo and adaptive sampling schemes

Population Monte Carlo Algorithm

Choice of the kernels $Q_{i,t}$

Recursion on the weights

Define

$$
\Psi(\alpha) = \left( \alpha_d \int \left[ \frac{q_d(x, x')}{\sum_{j=1}^D \alpha_j q_j(x, x')} \right] \prod \Pi(dx, dx') \right)_{1 \leq d \leq D}
$$
on the simplex

$$
S = \left\{ \alpha = (\alpha_1, \ldots, \alpha_D); \alpha_d \geq 0, 1 \leq d \leq D \text{ and } \sum_{d=1}^D \alpha_d = 1 \right\}.
$$

and

$$
\alpha^{t+1} = \Psi(\alpha^t)
$$

Connection with RBDPMCA ??

Under the assumption $(1 \leq d \leq D)$

$$(A2) \quad - \infty < \int \log(q_d(x, x')) \prod \Pi(dx, dx') < \infty$$

Assumption automatically satisfied when all $\pi/q_d$'s are bounded.

Proposition

Under $(A1)$ and $(A2)$, for every $\alpha \in S$,

$$KL(\Psi(\alpha)) \leq KL(\alpha).$$

©The Kullback divergence decreases at every iteration of RBDPMCA!!

An integrated EM interpretation

For $\bar{x} = (x, x')$ and $K \sim M(1, (\alpha_d)_{1 \leq d \leq D})$,

$$
\alpha^{\text{min}} = \arg \min_{\alpha \in S} KL(\alpha) = \arg \max_{\alpha \in S} \int \log p_{\alpha}(\bar{x}) \prod \Pi(d\bar{x})
$$

$$=
\arg \max_{\alpha \in S} \int \log \int p_{\alpha}(\bar{x}, K) dK \prod \Pi(d\bar{x})
$$

Then $\alpha^{t+1} = \Psi(\alpha^t)$ means

$$
\alpha^{t+1} = \arg \max_{\alpha} \int \int \mathbb{E}_{\alpha^t}(\log p_{\alpha}(\bar{X}, K)|\bar{X} = \bar{x}) \prod \Pi(d\bar{x})
$$

and

$$
\lim_{t \to \infty} \alpha^t = \alpha^{\text{min}}
$$

CLT

Proposition

Under $(A1)$, for every $h$ such that

$$
\min_{d \in \{1, \ldots, D\}} \int h^2(x') \pi(x)/q_d(x, x') \prod \Pi(dx, dx') < \infty
$$

$$
\frac{1}{\sqrt{N}} \sum_{i=1}^N (\bar{\omega}_{i,t} h(x_{i,t}) - \Pi(h)) \overset{N}{\sim} \mathcal{N}(0, \sigma^2_t)
$$

where

$$
\sigma^2_t = \int \left\{ (h(x') - \Pi(h))^2 \frac{\pi(x')}{\sum_{d=1}^D \alpha_d q_d(x, x')} \right\} \prod \Pi(dx, dx').
$$
1 Crash course in simulation

2 Population Monte Carlo Algorithm

3 Illustrations
   - Toy (1)
   - Toy (2)
   - Mixtures

4 Further advances

Example (A toy example (1))

Target \( \frac{1}{4} N(-1,0.3)(x) + \frac{1}{4} N(0,1)(x) + \frac{1}{2} N(3,2)(x) \)

3 proposals: \( N(-1,0.3), N(0,1) \) and \( N(3,2) \)

<table>
<thead>
<tr>
<th></th>
<th>0.0500000</th>
<th>0.0500000</th>
<th>0.9000000</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.2605712</td>
<td>0.0997029</td>
<td>0.6397259</td>
</tr>
<tr>
<td>6</td>
<td>0.2740816</td>
<td>0.1916018</td>
<td>0.5343166</td>
</tr>
<tr>
<td>10</td>
<td>0.2989651</td>
<td>0.1920090</td>
<td>0.5090259</td>
</tr>
<tr>
<td>16</td>
<td>0.2651511</td>
<td>0.2412904</td>
<td>0.4935585</td>
</tr>
</tbody>
</table>

Table: Weight evolution

Example (A toy example (2))

Target \( N(0,1) \).

3 Gaussian random walks proposals:

\[ q_1(x,x') = f_{N(x,0.1)}(x'), \]

\[ q_2(x,x') = f_{N(x,2)}(x'), \]

and \( q_3 = f_{N(x,10)}(x') \)

Use of the Rao-Blackwellised 3-kernel algorithm with \( N = 100,000 \)
Table: Evolution of the weights

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.33333</td>
<td>0.33333</td>
</tr>
<tr>
<td>2</td>
<td>0.24415</td>
<td>0.43145</td>
</tr>
<tr>
<td>3</td>
<td>0.19525</td>
<td>0.52445</td>
</tr>
<tr>
<td>4</td>
<td>0.10725</td>
<td>0.72955</td>
</tr>
<tr>
<td>5</td>
<td>0.06155</td>
<td>0.88355</td>
</tr>
<tr>
<td>6</td>
<td>0.04255</td>
<td>0.92950</td>
</tr>
<tr>
<td>7</td>
<td>0.03790</td>
<td>0.93760</td>
</tr>
<tr>
<td>8</td>
<td>0.03130</td>
<td>0.94505</td>
</tr>
<tr>
<td>9</td>
<td>0.03460</td>
<td>0.94875</td>
</tr>
</tbody>
</table>

Figure: A few examples of convergence on the divergence surface

Example (Back to Gaussian mixtures)

iid sample \( y = (y_1, \ldots, y_n) \) from

\[
p \mathcal{N}(\mu_1, \sigma^2) + (1-p) \mathcal{N}(\mu_2, \sigma^2)
\]

where \( p \neq 1/2 \) and \( \sigma^2 \) are fixed and

\[
\mu_1, \mu_2 \sim \mathcal{N}(\alpha, \sigma^2/\delta)
\]

Use of the random walk RBDPMC with \( D \) different scales

\[
\mathcal{N}((\mu_i)^{(t-1)}, v_i)
\]

Figure: PMC sample (\( N=1000 \)) after 10 iterations.
Origin discrimination

Simple RB weight

\[ \omega_{i,t} = \frac{\pi(x_{i,t})}{\sum_{d=1}^{D} \alpha_{d}^{t,N} q_{d}(\tilde{x}_{i,t-1}, x_{i,t})} \]

still too local (dependent on \(i\))

Paradox

Same value + different origin = different weight!

Double Rao–Blackwellisation

Replace \(\omega_{i,t}\) with 2x Rao–Blackwellised version

\[ \omega_{2RB}^{i,t} = \frac{\pi(x_{i,t})}{\sum_{j=1}^{N} \omega_{2RB}^{j,t-1} \sum_{d=1}^{D} \alpha_{d}^{t,N} q_{d}(\tilde{x}_{j,t-1}, x_{i,t})} \]

\(\text{\textcopyright} \) If \(x_{i,t} = x_{\ell,t}\), then \(\omega_{2RB}^{i,t} = \omega_{2RB}^{\ell,t}\)

Better recovery in multimodal situations but \(O(N^2)\) cost

New criterion

Marginal divergence

\[ KL(\alpha) = \int \left[ \log \left( \frac{\pi(x')}{\int \Pi(dx) \sum_{d=1}^{D} \alpha_{d} q_{d}(x, x')} \right) \right] \Pi(dx'). \]

More rational Kullback divergence between \(\Pi\) and the integrated mixture
Population Monte Carlo and adaptive sampling schemes
Further advances

Weight actualisation

Theoretical EM-like step

\[ \alpha_{t+1}^d = \mathbb{E}^\pi \left[ \alpha_t^d \frac{\int \Pi(dx)q_d(x, x')}{\int \Pi(dx) \sum_{d=1}^D \alpha_t^d q_d(x, x')} \right] \]

Implementation

\[ \alpha_{t+1}^d = \alpha_t^d \sum_{i=1}^N \bar{\omega}_{i,t}^2 \frac{\sum_{j=1}^N \bar{\omega}_{j,t-1}^{2RB} q_d(x_{j,t-1}, x_{i,t})}{\sum_{j=1}^N \bar{\omega}_{j,t-1}^{2RB} \sum_{d=1}^D \alpha_t^d q_d(x_{j,t-1}, x_{i,t})} \]

[O(N^2d^2)]

Aiming at variance reduction

Estimation perspective for approximating

\[ J = \int f(y) \pi(y) dy \]

Proposition

The optimal importance distribution

\[ g^\star(x) = \frac{|f(x)\pi(x)|}{\int |f(y)|\pi(y) dy} \]

achieves the minimal variance for estimating \( J \)

A formal result: requires exact knowledge of \( \int |f(y)|\pi(y) dy \)

SIS version

For the self-normalised version, the optimum importance function is

\[ g^\star(x) = \frac{|f(x) - J|\pi(x)}{\int |f(y) - J|\pi(y) dy} \]

Still not available!

Weight update

Try instead to get a guaranteed variance reduction, using recursion

\[ \alpha_{t+1, N}^d = \frac{\sum_{i=1}^N \bar{\omega}_{i,t}^2 \left( h(x_{i,t}) - N^{-1} \sum_{j=1}^N \bar{\omega}_{j,t} h(x_{j,t}) \right)^2}{\sum_{i=1}^N \bar{\omega}_{i,t}^2 \left( h(x_{i,t}) - N^{-1} \sum_{j=1}^N \bar{\omega}_{j,t} h(x_{j,t}) \right)^2} \cdot \mathbb{I}_d(K_{i,t}) \]
Theoretical version

...with theoretical equivalent

\[
\Psi(\alpha) = \left( \nu_h \left( \frac{\sum_{d=1}^{D} \alpha_d q_d(x,x')}{\sigma^2_h(\alpha)} \right) \right)_{1 \leq d \leq D}
\]

where

\[
\nu_h(dx, dx') = \pi(x')(h(x') - \pi(h))^2 \pi(dx) \pi(dx')
\]

and

\[
\sigma^2_h(\alpha) = \nu_h \left( \frac{1}{\sum_{d=1}^{D} \alpha_d q_d(x, x')} \right)
\]

Variance minimisation

Proposition

Under (A1), for all \( \alpha \in \mathcal{S} \),

\[
\sigma^2_h(\Psi(\alpha)) \leq \sigma^2_h(\alpha),
\]

\[
\lim_{t \to \infty} \alpha_t = \alpha_{\min} \quad \text{and} \quad \alpha_d \xrightarrow{N \to \infty} P \alpha_d
\]

© The variance decreases at every iteration of RBDPMCA

Illustration

Example

Case of a \( \mathcal{N}(0, 1) \) target, \( h(x) = x \) and mixture of \( D = 3 \) independent proposals

- \( \mathcal{N}(0, 1) \)
- \( \mathcal{C}(0, 1) \) (a standard Cauchy distribution)
- \( \pm \sqrt{a} (0.5, 0.5) \) where \( s \sim \mathcal{B}(1, 0.5) \) (Bernoulli distribution with parameter 1/2) [This is the optimal choice, \( g^* \!]

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \delta_{t,N} )</th>
<th>( \alpha_{1,t} )</th>
<th>( \alpha_{2,t} )</th>
<th>( \alpha_{3,t} )</th>
<th>( \text{var}(\delta_{t,N}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.00126</td>
<td>.1</td>
<td>0.8</td>
<td>0.1</td>
<td>0.982</td>
</tr>
<tr>
<td>2</td>
<td>.00061</td>
<td>.112</td>
<td>0.715</td>
<td>0.173</td>
<td>0.926</td>
</tr>
<tr>
<td>3</td>
<td>-.00124</td>
<td>.116</td>
<td>0.607</td>
<td>0.276</td>
<td>0.863</td>
</tr>
<tr>
<td>5</td>
<td>.00248</td>
<td>.108</td>
<td>0.357</td>
<td>0.534</td>
<td>0.742</td>
</tr>
<tr>
<td>10</td>
<td>.00332</td>
<td>.049</td>
<td>0.062</td>
<td>0.888</td>
<td>0.650</td>
</tr>
<tr>
<td>15</td>
<td>.00284</td>
<td>.026</td>
<td>0.015</td>
<td>0.958</td>
<td>0.640</td>
</tr>
<tr>
<td>20</td>
<td>.00062</td>
<td>.019</td>
<td>0.004</td>
<td>0.976</td>
<td>0.638</td>
</tr>
</tbody>
</table>

Table: PMC estimates for \( N = 100,000 \) and \( T = 20 \).
Example (Cox-Ingersol-Ross model)

Diffusion

\[ dX_t = k(a - X_t)dt + \sigma \sqrt{X_t} dW_t \]

discretised as (\( \delta > 0 \))

\[ X_{t+1} = X_t + k(a - X_t)\delta + \sigma \sqrt{\delta} X_t \epsilon_t \]

Computation of a European option price \( \mathbb{E}[(K - X_T)^+] \)

Requires the simulation of the whole path using independent

- exact Gaussian distribution shifted by \( a_1 \)
- exact Gaussian distribution
- exact Gaussian distribution shifted by \( a_3 \)

Figure: Cumulated \( \alpha_d \)'s

Another Kullback criterion

Given the optimal choice \( g^\sharp \), another possibility is to minimize the \( h \)-divergence

\[
\tilde{KL}(\alpha) = \int \left[ \log \left( \frac{g^\sharp(x')}{\int \Pi(dx) \sum_{d=1}^{D} \alpha_d q_d(x, x')} \right) \right] \Pi(dx').
\]

Plusses

Gets closer to the minimal variance solution and can be extended to parameterised kernels \( q_d \)'s