Population Monte Carlo and adaptive sampling schemes

Christian P. Robert

Université Paris Dauphine and CREST-INSEE
http://www.ceremade.dauphine.fr/~xian

Joint work with O. Cappé, R. Douc, A. Guillin, J.M. Marin

Outline

General purpose

Given a density π 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 π and estimate $\Pi(h)$ by

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

LLN: $\hat{\Pi}_N^{MC}(h) \xrightarrow{\text{as}} \Pi(h)$

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

$$CLT: \quad \sqrt{N}\left(\hat{\Pi}_N^{MC}(h) - \Pi(h)\right) \overset{\mathscr{L}}{\rightsquigarrow} \mathscr{N}\left(0, \Pi\left\{[h - \Pi(h)]^2\right\}\right).$$

Caveat

Often impossible or inefficient to simulate directly from Π

MCMC basics

Generate

$$x^{(1)}, \ldots, x^{(T)}$$

ergodic Markov chain $(x_t)_{t\in\mathbb{N}}$ with stationary distribution π

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 $q(\cdot|\cdot)$,

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

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

where

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

A generic MH: RWMH

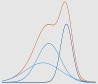
Choose for proposal the **random walk**

$$q(x|y) = g(y - x) = g(x - y)$$

- 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:

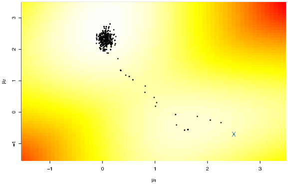
$$\theta^{(t+1)} = \begin{cases} \theta^{(t)} + \omega\varepsilon^{(t)} & \text{if } u^{(t)} < \rho^{(t)} \\ \theta^{(t)} & \text{otherwise} \end{cases}$$

where

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

and ω scaled for **good** acceptance rate

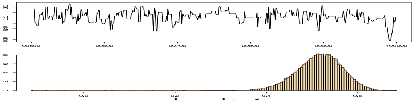
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$



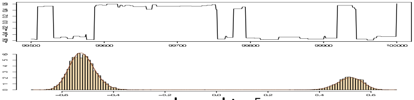
MCMC difficulties

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

Noisy AR_1^2



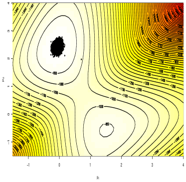
scale equal to $.1$



scale equal to $.5$

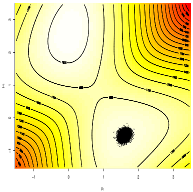
Adaptive MCMC?!

A wee problem with Gibbs on mixtures



Gibbs started at random

Gibbs stuck at the wrong mode



[Marin, Mengersen & Robert, 2005]

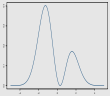
MCMC difficulties

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

Example (Bimodal target)

Density

$$f(x) = \frac{\exp -x^2/2}{\sqrt{2\pi}} \, \frac{4(x-.3)^2+.01}{4(1+(.3)^2)+.01} \, .$$

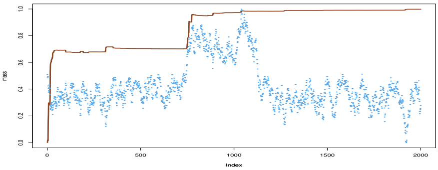


and use of random walk Metropolis–Hastings algorithm with variance .04

Evaluation of the missing mass by

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

[Philippe & Robert, 2001]



MCMC difficulties

Trapping modes may remain undetected

Convergence to the stationary distribution can be very slow or intractable

Difficult adaptivity

Population Monte Carlo and adaptive sampling schemes
Crash course in simulation
MCMC difficulties

Simple adaptive MCMC is not possible

$\lightning$ **Algorithms trained on-line usually invalid:**
using the whole past of the "chain" implies that this is not a Markov chain any longer!

To controlled MCMC

Population Monte Carlo and adaptive sampling schemes
Crash course in simulation
MCMC difficulties

Example (Poly t distribution)

Consider a t-distribution $\mathcal{T}(3, \theta, 1)$ 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

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

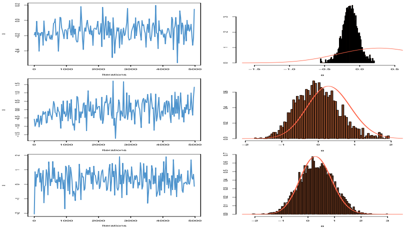
where $\xi \sim \mathcal{N}(\mu_t, \sigma_t^2)$.

Population Monte Carlo and adaptive sampling schemes
Crash course in simulation
MCMC difficulties

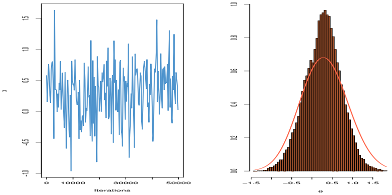
Example (Poly t distribution (2))

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 $x_j \sim \mathcal{T}_3$ 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...

▸ out-of-control MCMC

Controlled MCMC

Optimal choice of a parameterised proposal $\mathfrak{K}(x, dy; \theta)$ against a proposal **minimisation problem**

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

where

$$\eta(\theta) = \int_{\mathcal{X}} \mathfrak{H}(\theta, x)\mu_\theta(dx)$$

[Andrieu & Robert, 2001]

- Coerced acceptance
- Autocorrelations
- Moment matching

Two-time scale stochastic approximation

Set $\xi_i = (\eta_i, \dot\eta_i)$
Corresponding recursive system

$$\begin{aligned} x_{i+1} &\sim \mathfrak{K}(x_i, dx_{i+1}; \theta_i) \\ \xi_{i+1} &= (1-\gamma_{i+1})\xi_i + \gamma_{i+1}\xi(\theta_i, x_{i+1}) \\ \theta_{i+1} &= \theta_i - \gamma_{i+1}\varepsilon_{i+1}\dot\eta_i \Psi'(\eta_i) \end{aligned}$$

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

Two-time scale stochastic approximation (2)

Convergence conditions on $\{\gamma_i\}$ and $\{\varepsilon_i\}$

- $\{\gamma_i\}$ and $\{\varepsilon_i\}$ go to 0 at infinity
- slow decrease to 0:

$$\sum_i \gamma_i \varepsilon_i = \infty \qquad \sum_i \gamma_i^2 < \infty$$

[Andrieu & Moulines, 2002]

Warning

Hidden difficulties...

Importance Sampling

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

$$\Pi(h) = \int h(x)\{\pi/q\}(x)q(x)\mu(dx).$$

Principle

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

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

Then

$LLN: \quad \hat\Pi^{IS}_{Q,N}(h) \xrightarrow{\text{as}} \Pi(h) \qquad$ and if $Q((h\pi/q)^2) < \infty$,

$$CLT: \quad \sqrt{N}(\hat\Pi^{IS}_{Q,N}(h) - \Pi(h)) \overset{\mathscr{L}}{\rightsquigarrow} \mathscr{N}\left(0, Q\{(h\pi/q - \Pi(h))^2\}\right).$$

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

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

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

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

$CLT:\quad \sqrt{N}(\hat{\Pi}^{SNIS}_{Q,N}(h) - \Pi(h)) \overset{\mathcal{L}}{\rightsquigarrow} \mathscr{N}\left(0, \pi((\pi/q)(h-\Pi(h))^2)\right).$

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 π

[Rubin, 1987]

Theorem (Bootstraped importance sampling)

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

$$x_i^\star \sim \pi(x)$$

where $\bar\omega_{i,t} = \omega_{i,t}/\sum_{j=1}^{N}\omega_{j,t}$

Note

Obviously, the $x_i^\star$'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

Sequential importance sampling

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

$$\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)}) \stackrel{iid}{\approx} \pi(x)$$

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

Sequential Monte Carlo applied to a fixed distribution π

[Iba, 2000]

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 π and q

Iterated importance sampling

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

$$x_i^{(t)} \sim q_t(x|x_i^{(t-1)}) \qquad i = 1, \ldots, n, \quad t = 1, \ldots$$

and

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

is still unbiased for

$$\varrho_i^{(t)} = \frac{\pi_t(x_i^{(t)})}{q_t(x_i^{(t)}|x_i^{(t-1)})}, \qquad i = 1, \ldots, n$$

Fundamental importance equality

Preservation of unbiasedness

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

for **any distribution** g on $X^{(t-1)}$

PMCA: Population Monte Carlo Algorithm

At time $t = 0$

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

Generate $(J_{i,0})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar\omega_{i,0})_{1\le i\le N})$
Set $\tilde x_{i,0} = x_{J_i,0}$

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

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

Generate $(J_{i,t})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar\omega_{i,t})_{1\le i\le N})$
Set $\tilde x_{i,t} = x_{J_{i,t},t}$.

Self-norm'ed weights

Population Monte Carlo and adaptive sampling schemes
Population Monte Carlo Algorithm
Population Monte Carlo Algorithm

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

Population Monte Carlo and adaptive sampling schemes
Population Monte Carlo Algorithm
Population Monte Carlo Algorithm

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 π_t 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 $\mathbf{x}^{(t)}$ 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

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

Given $\mathcal{F}_{N,t-1}$, how to construct $Q_{i,t}(\mathcal{F}_{N,t-1})$?

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 α_d^{t+1} equal to previous **survival rates**

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

Population Monte Carlo and adaptive sampling schemes
└Population Monte Carlo Algorithm
└Choice of the kernels $Q_{i,t}$

DPMCA: D-kernel PMC Algorithm

At time $t = 0$, use PMCA.0 and set $\alpha_d^{1,N} = 1/D$
At time t ($t = 1, \ldots, T$),

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

Generate $(x_{i,t})_{1\le i\le N} \overset{\text{ind}}{\sim} Q_{K_{i,t}}(\tilde{x}_{i,t-1}, \cdot)$
and set $\omega_{i,t} = \pi(x_{i,t})/q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})$;

Generate $(J_{i,t})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar{\omega}_{i,t})_{1\le i\le N})$
and set $\tilde{x}_{i,t} = x_{J_{i,t},t}$, $\alpha_d^{t+1,N} = \sum_{i=1}^N \bar{\omega}_{i,t} \mathbb{I}_d(K_{i,t})$.

Population Monte Carlo and adaptive sampling schemes
└Population Monte Carlo Algorithm
└Choice of the kernels $Q_{i,t}$

An initial LLN

Under the assumption

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

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

Proposition
If **(A1)** *holds, for* $h \in L^1_{\Pi\otimes\gamma_u}$ *and every* $t \ge 1$,

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} h(x_{i,t}, K_{i,t}) \xrightarrow[\mathbb{P}]{N\to\infty} \Pi \otimes \gamma_u(h).$$

Population Monte Carlo and adaptive sampling schemes
└Population Monte Carlo Algorithm
└Choice of the kernels $Q_{i,t}$

Bad!!!

Even very bad because, while

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

convergence to γ_u implies that

$$\sum_{i=1}^{N} \bar{\omega}_{i,t} \mathbb{I}_{K_{i,t}=d} \xrightarrow[\mathbb{P}]{N\to\infty} \frac{1}{D}.$$

**At each iteration, every weight converges to $1/D$:
the algorithm fails to learn from experience!!!**

Saved by Rao-Blackwell !!

Idea:
Use Rao-Blackwellisation by deconditioning the chosen kernel

[Gelfand & Smith, 1990]

Use the whole mixture in the importance weights

$$\frac{\pi(x_{i,t})}{\sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde{x}_{i,t-1}, x_{i,t})} \quad \text{instead of} \quad \frac{\pi(x_{i,t})}{q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})}$$

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

RBDPMCA: Rao-Blackwellised D-kernel PMC Algorithm

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

Generate

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

and

$$(x_{i,t})_{1 \le i \le N} \overset{\text{ind}}{\sim} Q_{K_{i,t}}(\tilde{x}_{i,t-1}, \cdot)$$

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

Generate

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

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

LLN (2) and convergence

Proposition

Under **(A1)**, *for* $h \in L^1_\Pi$ *and for every* $t \ge 1$,

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

$$\alpha_d^{t,N} \xrightarrow[\mathbb{P}]{N \to \infty} \alpha_d^t$$

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

$$\alpha_d^t = \alpha_d^{t-1} \int \left(\frac{q_d(x, x')}{\sum_{j=1}^{D} \alpha_j^{t-1} q_j(x, x')} \right) \Pi \otimes \Pi(dx, dx').$$

Kullback divergence

For $\alpha \in S$,

$$\text{KL}(\boldsymbol{\alpha}) = \int \left[\log \left(\frac{\pi(x)\pi(x')}{\pi(x) \sum_{d=1}^{D} \alpha_d q_d(x, x')} \right) \right] \Pi \otimes \Pi(dx, dx')$$

Kullback divergence between Π and the mixture.

Goal
Obtain the mixture of q_d's closest to Π for the Kullback divergence

Recursion on the weights

Define

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

on the simplex

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

and

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

Connection with RBDPMCA ??

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

$$\textbf{(A2)} \qquad -\infty < \int \log(q_d(x,x'))\Pi \otimes \Pi(dx,dx') < \infty$$

Assumption automatically satisfied when all π/q_d's are bounded.

Proposition

Under **(A1)** *and* **(A2)**, *for every* $\boldsymbol{\alpha} \in S$,

$$KL(\Psi(\boldsymbol{\alpha})) \le KL(\boldsymbol{\alpha}).$$

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

An integrated EM interpretation

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

$$\begin{aligned}\boldsymbol{\alpha}^{\min} = \arg\min_{\boldsymbol{\alpha}\in S} KL(\boldsymbol{\alpha}) &= \arg\max_{\boldsymbol{\alpha}\in S} \int \log p_{\boldsymbol{\alpha}}(\bar{x})\Pi \otimes \Pi(d\bar{x}) \\ &= \arg\max_{\boldsymbol{\alpha}\in S} \int \log \int p_{\boldsymbol{\alpha}}(\bar{x},K)dK\, \Pi \otimes \Pi(d\bar{x})\end{aligned}$$

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

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

and

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

CLT

Proposition

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

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

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

where

$$\sigma_t^2 = \int \left\{ (h(x') - \Pi(h))^2 \frac{\pi(x')}{\sum_{d=1}^D \alpha_d^T q_d(x,x')} \right\} \Pi \otimes \Pi(dx,dx').$$

Example (A toy example (1))

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

3 proposals: $\mathcal{N}(-1,0.3)$, $\mathcal{N}(0,1)$ and $\mathcal{N}(3,2)$

1	0.0500000	0.05000000	0.9000000
2	0.2605712	0.09970292	0.6397259
6	0.2740816	0.19160178	0.5343166
10	0.2989651	0.19200904	0.5090259
16	0.2651511	0.24129039	0.4935585

Table: Weight evolution

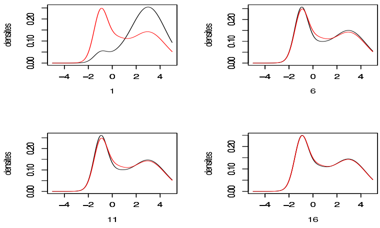


Figure: Target and mixture evolution

Example (A toy example (2))

Target $\mathcal{N}(0,1)$.

3 Gaussian random walks proposals:
$q_1(x,x') = f_{\mathcal{N}(x,0.1)}(x')$,
$q_2(x,x') = f_{\mathcal{N}(x,2)}(x')$
and $q_3 = f_{\mathcal{N}(x,10)}(x')$

Use of the Rao-Blackwellised 3-kernel algorithm with $N = 100,000$

1	0.33333	0.33333	0.33333
2	0.24415	0.43145	0.32443
3	0.19525	0.52445	0.28031
4	0.10725	0.72955	0.16324
5	0.08223	0.83092	0.08691
6	0.06155	0.88355	0.05490
7	0.04255	0.92950	0.02795
8	0.03790	0.93760	0.02450
9	0.03130	0.94505	0.02365
10	0.03460	0.94875	0.01665

Table: Evolution of the weights

Population Monte Carlo and adaptive sampling schemes
Illustrations
Toy (2)

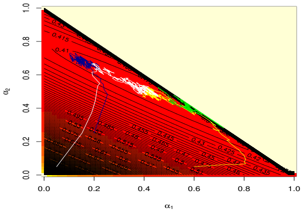


Figure: A few examples of convergence on the divergence surface

Example (Back to Gaussian mixtures)

iid sample $\underline{y} = (y_1, \ldots, y_n)$ from

$$p\mathcal{N}\left(\mu_1, \sigma^2\right) + (1-p)\mathcal{N}\left(\mu_2, \sigma^2\right)$$

where $p \neq 1/2$ and σ^2 are fixed and

$$\mu_1, \mu_2 \sim \mathcal{N}\left(\alpha, \sigma^2/\delta\right)$$

Use of the random walk RBDPMC with D different scales

$$\mathcal{N}((\mu)_i^{(t-1)}, v_i)$$

Population Monte Carlo and adaptive sampling schemes
Illustrations
Mixtures

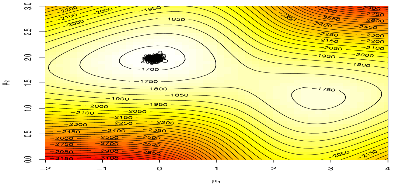


Figure: PMC sample (N=1000) after 10 iterations.

Origin discrimination

Simple RB weight

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

still too local (dependent on i)

Paradox

Same value + different origin = different weight!

Double Rao–Blackwellisation

Replace $\omega_{i,t}$ with 2×Rao–Blackwellised version

$$\omega_{i,t}^{2RB} = \pi(x_{i,t}) \Big/ \sum_{j=1}^{N} \bar{\omega}_{j,t-1}^{2RB} \sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde{x}_{j,t-1}, x_{i,t})$$

©If $x_{i,t} = x_{\ell,t}$, then $\omega_{i,t}^{2RB} = \omega_{\ell,t}^{2RB}$

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

New criterion

Marginal divergence

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

More rational Kullback divergence between Π and the integrated mixture

Weight actualisation

Theoretical EM-like step

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

Implementation

$$\alpha_d^{t+1} = \alpha_d^t \sum_{i=1}^N \bar\omega_{i,t}^{2RB} \frac{\sum_{j=1}^N \bar\omega_{j,t-1}^{2RB} q_d(x_{j,t-1}, x_{i,t})}{\sum_{j=1}^N \bar\omega_{j,t-1}^{2RB} \sum_{d=1}^D \alpha_d^t q_d(x_{j,t-1}, x_{i,t})}$$

[$\mathrm{O}(N^2 d^2)$]

Aiming at variance reduction

Estimation perspective for approximating

$$\mathfrak{I} = \int f(y)\pi(y)\,\mathrm{d}y$$

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 $\mathfrak{I}$

A formal result: requires exact knowledge of $\int |f(y)|\pi(y)\,\mathrm{d}y$

SIS version

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

$$g^\sharp(x) = \frac{|f(x) - \mathfrak{I}|\pi(x)}{\int |f(y) - \mathfrak{I}|\pi(y)\,\mathrm{d}y}$$

Still not available!

Weight update

Try instead to get a guaranteed variance reduction, using recursion

$$\alpha_d^{t+1,N} = \frac{\sum_{i=1}^N \bar\omega_{i,t}^2 \left(h(x_{i,t}) - N^{-1} \sum_{j=1}^N \bar\omega_{j,t} h(x_{j,t}) \right)^2 \mathbb{I}_d(K_{i,t})}{\sum_{i=1}^N \bar\omega_{i,t}^2 \left(h(x_{i,t}) - N^{-1} \sum_{j=1}^N \bar\omega_{j,t} h(x_{j,t}) \right)^2}.$$

Theoretical version

...with theoretical equivalent

$$\Psi(\alpha) = \left(\frac{\nu_h \left(\frac{\alpha_d q_d(x,x')}{(\sum_{l=1}^{D} \alpha_l q_l(x,x'))^2} \right)}{\sigma_h^2(\alpha)} \right)_{1\leq d\leq D}$$

where

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

and

$$\sigma_h^2(\alpha) = \nu_h \left(\frac{1}{\sum_{d=1}^{D} \alpha_d q_d(x,x')} \right)$$

Variance reduction in action

Proposition

Under **(A1)**, *for all* $\alpha \in \mathscr{S}$,

$$\sigma_h^2(\Psi(\alpha)) \leq \sigma_h^2(\alpha)\,,$$

$$\lim_{t\to\infty} \alpha^t = \alpha^{min} \quad \textit{and} \quad \alpha_d^{t,N} \xrightarrow[\mathbb{P}]{N\to\infty} \alpha_d^t$$

©The variance decreases at every iteration of RBDPMCA

Illustration

Example

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

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

t	$\delta^{t,N}$	$\alpha_1^{t,N}$	$\alpha_2^{t,N}$	$\alpha_3^{t,N}$	var($\delta^{t,N}$)
1	.00126	.1	0.8	0.1	0.982
2	.00061	.112	0.715	0.173	0.926
3	-.00124	.116	0.607	0.276	0.863
5	.00248	.108	0.357	0.534	0.742
10	.00332	.049	0.062	0.888	0.650
15	.00284	.026	0.015	0.958	0.640
20	.00062	.019	0.004	0.976	0.638

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

1. exact Gaussian distribution shifted by a_1
2. exact Gaussian distribution
3. exact Gaussian distribution shifted by a_3

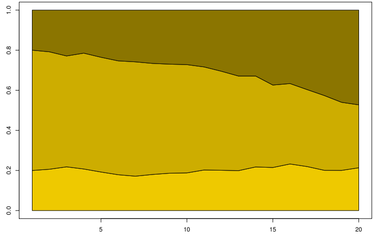


Figure: Cumulated α_d's

Another Kullback criterion

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

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

Plusses

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