Adaptive importance sampling in general mixture classes

Christian P. Robert

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

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

Efficient Monte Carlo, Sandbjerg Gods, July 15, 2008

Outline

Basics

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.

Importance Sampling

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

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

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

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

Then

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

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

Self-Normalised Importance Sampling

Self normalized version

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

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

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

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

Self-Normalised Importance Sampling

Self normalized version

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

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

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

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

The quality of the SNIS approximation depends on the choice of Q

Population Monte Carlo Algorithm

Repeated importance sampling

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

Repeated importance sampling

Idea Apply repeated importance sampling to simulate a sequence of iid samples with a fixed target π

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

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

Repeated importance sampling

Idea Apply repeated importance sampling to simulate a sequence of iid samples with a fixed target π

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

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

Sequential Monte Carlo applied to a static distribution π

[Iba, 2000]

Adaptive IS

Fact

IS can be generalized to encompass adaptive/local schemes

Adaptive IS

Fact

IS can be generalized to encompass adaptive/local schemes

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

Adaptive IS

Fact

IS can be generalized to encompass adaptive/local schemes

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

Incentive

Use previous sample(s) to learn about π and to update 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

$$\mathbb{E}\left[h(X^{(t)})\ \frac{\pi(X^{(t)})}{q_t(X^{(t)}|X^{(t-1)})}\right]$$

Fundamental importance equality

Preservation of unbiasedness

$$\mathbb{E}\left[h(X^{(t)})\ \frac{\pi(X^{(t)})}{q_t(X^{(t)}|X^{(t-1)})}\right]$$

$$= \int h(x)\,\frac{\pi(x)}{q_t(x|y)}\,q_t(x|y)\,g(y)\,dx\,dy$$

Fundamental importance equality

Preservation of unbiasedness

$$\mathbb{E}\left[h(X^{(t)})\ \frac{\pi(X^{(t)})}{q_t(X^{(t)}|X^{(t-1)})}\right]$$

$$= \int h(x)\,\frac{\pi(x)}{q_t(x|y)}\,q_t(x|y)\,g(y)\,dx\,dy$$

$$= \int h(x)\,\pi(x)\,dx$$

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

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

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

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

Notes

1. $Q_{i,t}$'s chosen arbitrarily under support constraint

Notes

1. $Q_{i,t}$'s chosen arbitrarily under support constraint
2. $Q_{i,t}$'s may depend on whole sequence of simulations

Notes

1. $Q_{i,t}$'s chosen arbitrarily under support constraint
2. $Q_{i,t}$'s may depend on whole sequence of simulations
3. alternatives to multinomial sampling reduce variance/preserve "unbiasedness"

[Kitagawa, 1996 / Carpenter, Clifford & Fearnhead, 1997]

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

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

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

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

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

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

or (better)

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

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

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

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

or (better)

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

Given the past, how to construct $Q_{i,t}$?

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**

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**

Darwinian survival of the fittest:

The algorithm should automatically fit the mixture to the target distribution

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

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

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

An initial LLN

Under the assumption

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

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

Proposition

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

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

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

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

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]

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

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

RBDPMCA: Rao-Blackwellised D-kernel PMC Algorithm

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

Generate

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

and

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

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

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

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

LLN (2) and convergence

Proposition

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

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

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

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

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

Kullback divergence

For $\alpha \in S$,

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

Kullback divergence between Π and the mixture.

Kullback divergence

For $\alpha \in S$,

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

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

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

and

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

Connection with RBDPMCA ??

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

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

Proposition

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

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

Connection with RBDPMCA ??

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

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

Proposition

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

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

©The Kullback divergence decreases at every iteration of RB-DPMCA!!!

An integrated EM interpretation

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

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

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 (First toy example)

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

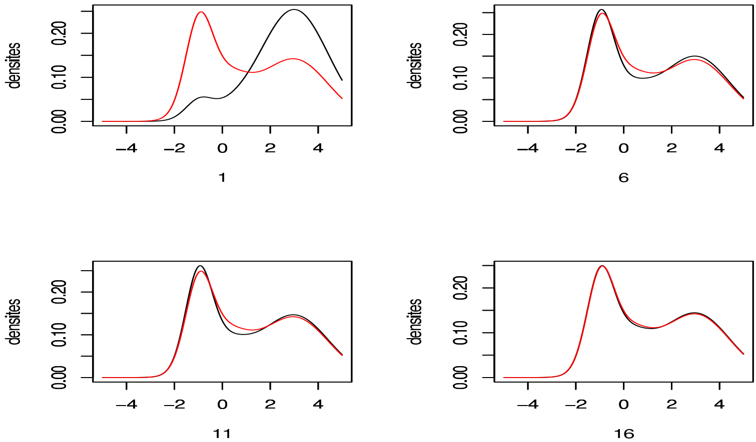


Figure: Target and mixture evolution

Example (Second toy example)

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

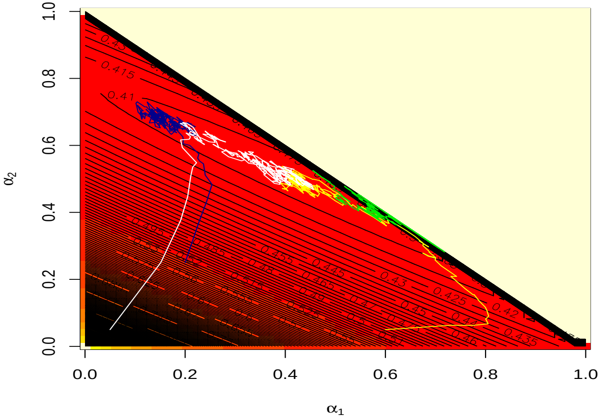
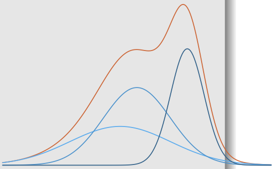


Figure: A few examples of convergence on the divergence surface

Example (Gaussian mixtures)

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

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

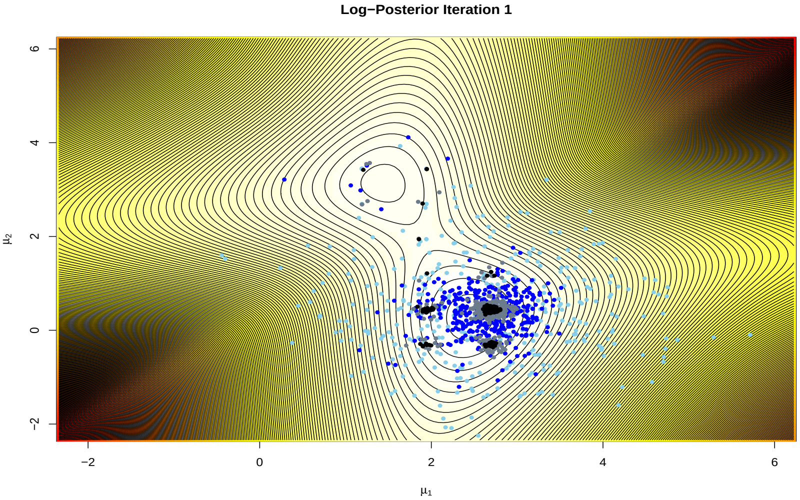


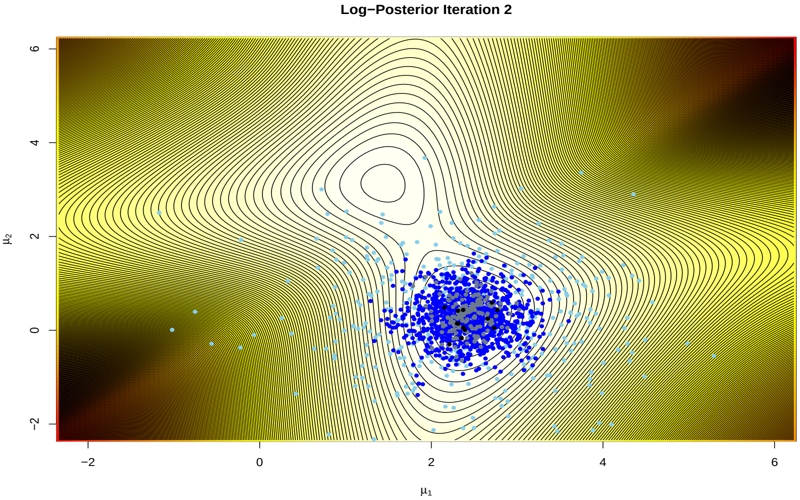
where $p \neq 1/2$ and σ^2 are fixed and

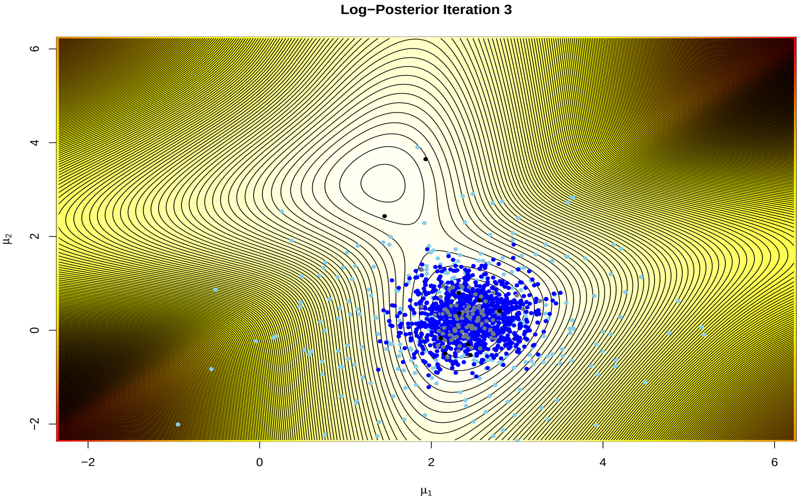
$$\mu_1, \mu_2 \sim \mathscr{N}\left(\alpha, \sigma^2/\delta\right)$$

Use of the random walk RBDPMC with D different scales

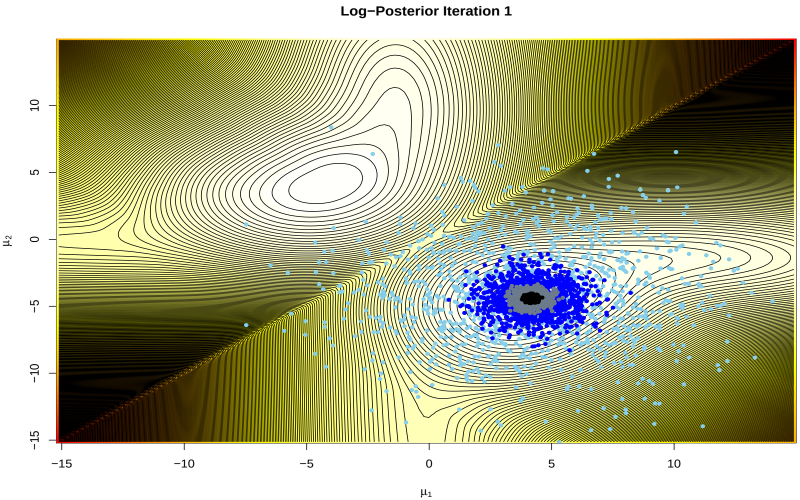
$$\mathscr{N}\left((\mu)_i^{(t-1)}, v_i\right)$$


Log−Posterior Iteration 1
μ2
μ1
−2
0
2
4
6
−2
0
2
4
6

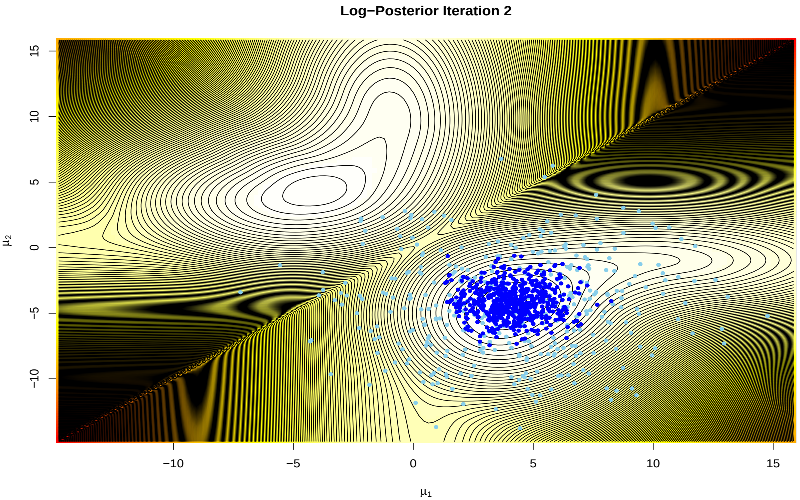

Log−Posterior Iteration 2
μ2
μ1
6
4
2
0
−2
−2
0
2
4
6


Log−Posterior Iteration 3
μ2
μ1
−2
0
2
4
6
−2
0
2
4
6

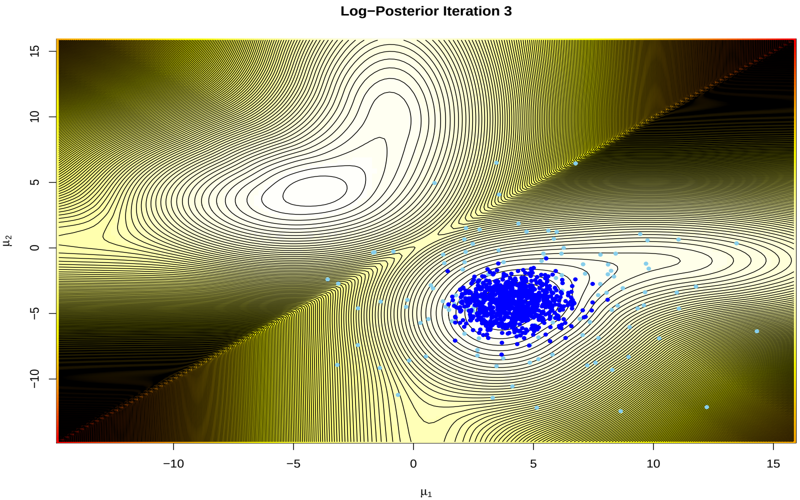
But...



But...



But...



Discrimination against origin

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)

Discrimination against origin

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

Solution

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

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

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

Double Rao–Blackwellisation

Solution

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

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

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

Better recovery in multimodal situations but $\mathrm{O}(N^2)$ cost

Double Rao–Blackwellisation

Solution

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

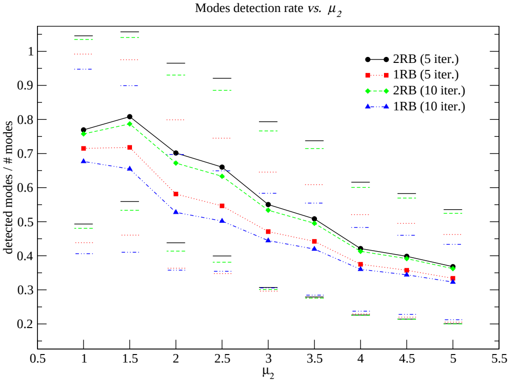
$$\omega_{i,t}^{2RB} = \pi(x_{i,t}) \Big/ \sum_{j=1}^{N} \bar{\omega}_{j,t-1}^{2RB} \sum_{d=1}^{D} \alpha_d^{t,N} q_d(\tilde{x}_{j,t-1}, x_{i,t})$$

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

Better recovery in multimodal situations but $\mathrm{O}(N^2)$ cost

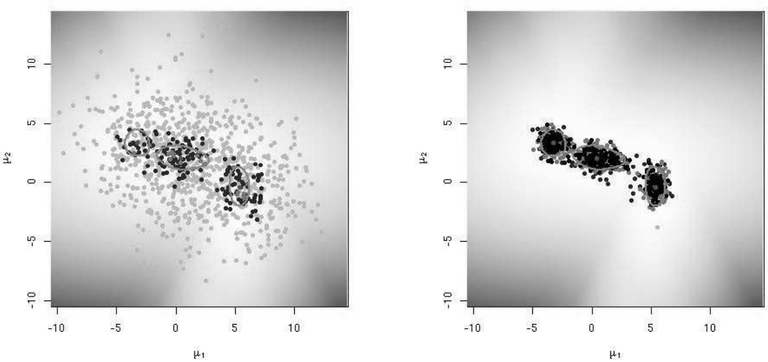
Can be lowered to $\mathrm{O}(N \log N)$ cost by preliminary subsampling based on $\bar{\omega}_{j,t-1}^{2RB}$'s

Illustration for the mixture example



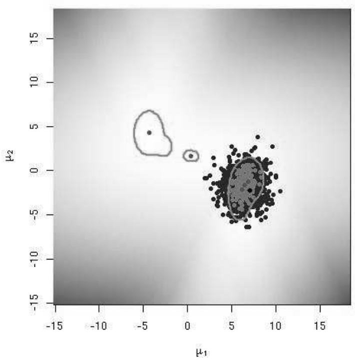
Rate of recovery of modes against μ_2

Illustration for the mixture example (cont'd)



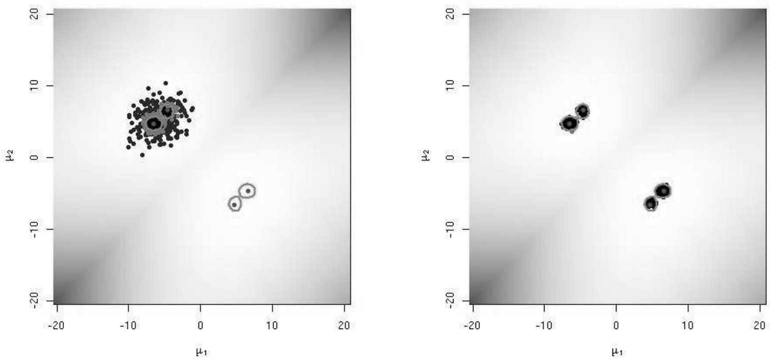
Comparison between single and double Rao-Blackwellisation

Illustration for the mixture example (cont'd)



Comparison between single and double Rao-Blackwellisation

Illustration for the mixture example (cont'd)



Comparison between single and double Rao-Blackwellisation

A corresponding optimality criterion

Marginal divergence

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

More rational Kullback divergence between Π and the integrated mixture

Weight actualisation

Theoretical EM-like step

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

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 (=true MC) perspective for approximating

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

Aiming at variance reduction

Estimation (=true MC) perspective for approximating

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

Proposition (Optimal importance distribution)

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

achieves the minimal variance for estimating $\mathfrak{I}$

Aiming at variance reduction

Estimation (=true MC) perspective for approximating

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

Proposition (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

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

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 (Variance decrease)

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

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

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

Variance reduction in action

Proposition (Variance decrease)

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)
- $\pm\sqrt{\mathscr{E}xp(0.5)}$ where $s \sim \mathscr{B}(1,0.5)$ (Bernoulli distribution with parameter 1/2)

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)
- $\pm\sqrt{\mathscr{E}xp(0.5)}$ 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	.8	.1	.982
2	.00061	.112	.715	.173	.926
3	-.00124	.116	.607	.276	.863
5	.00248	.108	.357	.534	.742
10	.00332	.049	.062	.888	.650
15	.00284	.026	.015	.958	.640
20	.00062	.019	.004	.976	.638

Table: PMC estimates for $N = 100,000$ and $T = 20$.

Example (Cox-Ingersol-Ross model (1))

Diffusion

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

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

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

Example (Cox-Ingersol-Ross model (1))

Diffusion

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

(wrongly) 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

$$\mathfrak{P} = \mathbb{E}_{\mathbb{P}}\left[\exp\left(-\int_0^T r_t dt\right) M \max(r_T - K, 0)\right]$$

Example (Cox-Ingersol-Ross model (2))

Requires the simulation of the whole path using independent

1. exact Gaussian distribution

Example (Cox-Ingersol-Ross model (2))

Requires the simulation of the whole path using independent

1. exact Gaussian distribution
2. exact Gaussian distribution shifted by a_i $(i = 2, 3)$

$$X_t^n = X_t^n + (\eta - kX_t^n + a_i \sigma \sqrt{X_t^n}) \frac{T}{n} + \sigma \sqrt{X_t^n} \epsilon_p \, ,$$

with weights $\alpha_1^{t,N}$, $\alpha_2^{t,N}$ and $\alpha_3^{t,N}$.

Example (Cox-Ingersol-Ross model (3))

t	$\widehat{\mathfrak{P}}_{t,N}^{PMC}$	$\alpha_1^{t,N}$	$\alpha_2^{t,N}$	$\alpha_3^{t,N}$	$\sigma_{1,t}^2$
1	9.2635	0.3333	0.3333	0.3334	27.0664
2	9.2344	0.4748	0.3703	0.1549	13.4474
3	9.2785	0.5393	0.3771	0.0836	9.7458
4	9.2495	0.5672	0.3835	0.0493	8.5258
5	9.2444	0.5764	0.3924	0.0312	7.8595
6	9.2400	0.5780	0.4014	0.0206	7.5471
7	9.2621	0.5765	0.4098	0.0137	7.2214
8	9.2435	0.5727	0.4183	0.0090	7.1354
9	9.2553	0.5682	0.4260	0.0058	7.0289
10	9.2602	0.5645	0.4320	0.0035	6.8854

Table: Approximation of $\mathfrak{P}$ for $K = 0.07$ with $a_2 = 1$ and $a_3 = 2$

Another Kullback criterion

Given the optimal choice $g^\sharp$, another possibility is to minimize the $\mathfrak{h}$-entropy Optimal g

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

Another Kullback criterion

Given the optimal choice $g^\sharp$, another possibility is to minimize the $\mathfrak{h}$-entropy ▸ Optimal g

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

Plusses

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

Solution

Weight update

$$\alpha_d^{t+1} \propto \alpha_d^t \mathbb{E}_{\bar{\pi}} \left[\frac{g^\sharp(x')}{\left(\sum_d \alpha_d^t q_d(x,x')\right)^2} q_d^t(x,x') \right]$$

Solution

Weight update

$$\alpha_d^{t+1} \propto \alpha_d^t \mathbb{E}_{\bar{\pi}} \left[\frac{g^\sharp(x')}{\left(\sum_d \alpha_d^t q_d(x,x')\right)^2} q_d^t(x,x') \right]$$

...and proposal update

$$q_d^{t+1}(x,x') \propto \frac{g^\sharp(x') q_d^t(x,x')}{\sum_d \alpha_d^t q_d^t(x,x')}$$

Independent mixture proposal

Use of proposals

$$q_{(\alpha,\theta)}(x) = \sum_{d=1}^{D} \alpha_d q_d(x;\theta_d)\,,$$

when adapting both the α_d's and the θ_d's

Optimality criterion

Kullback-Leibler divergence between the target density π and the mixture $q_{(\alpha,\theta)}$:

$$\mathfrak{E}(\pi, q_{(\alpha,\theta)}) = D(\pi \| q_{(\alpha,\theta)}) = \int \log \left(\frac{\pi(x)}{\sum_{d=1}^{D} \alpha_d q_d(x; \theta_d)} \right) \pi(x) \mathrm{d}x \,.$$

equivalent to maximising

$$\int \log \left(\sum_{d=1}^{D} \alpha_d q_d(x; \theta_d) \right) \pi(x) \, \mathrm{d}x \,,$$

EM foundations

Latent variable $Z \in \{1, \ldots, D\}$ as component indicator

$$(x, z) \sim \alpha_z q_z(x; \theta_x)$$

Maximise entropy via integrated-EM:
E step uses expected complete log-likelihood:

$$\mathbb{E}_\pi^X \left[\mathbb{E}^Z_{(\alpha^t, \theta^t)} \left\{ \log \left(\alpha_Z q_Z(X; \theta_Z) \right) | X \right\} \right] ,$$

where inner expectation computed under

$$f(z|x) = \alpha_z^t q_z(x; \theta_z^t) \bigg/ \sum_{d=1}^D \alpha_d^t q_d(x; \theta_d^t) \, ,$$

and outer expectation under $X \sim \pi$.

M-PMC

At iteration t,

1. Generate a sample $(X_{i,t})$ from the mixture IS proposal parameterised by $(\alpha^{t,N}, \theta^{t,N})$ and compute IS weights

$$\bar{\omega}_{i,t} \propto \pi(X_{i,t}) / \sum_{d=1}^{D} \alpha_d^{t,N} q_d(X_{i,t}; \theta_d^{t,N})$$

and posterior probabilities

$$\rho_d(X_{i,t}; \alpha^{t,N}, \theta^{t,N}) \propto \alpha_d^{t,N} q_d(X_{i,t}; \theta_d^{t,N})$$

M-PMC (2)

2. Update the parameters α_d and θ_d as

$$
\begin{aligned}
\alpha_d^{t+1,N} &= \sum_{i=1}^{N} \bar{\omega}_{i,t} \rho_d \left(X_{i,t}; \alpha^{t,N}, \theta^{t,N} \right) , \\
\theta_d^{t+1,N} &= \arg\max_{\theta_d} \left[\sum_{i=1}^{N} \bar{\omega}_{i,t} \rho_d \left(X_{i,t}; \alpha^{t,N}, \theta^{t,N} \right) \right. \\
&\qquad \left. \times \log \left\{ q_d \left(X_{i,t}; \theta_d^{t,N} \right) \right\} \right]
\end{aligned}
$$

Rao–Blackwellisation

Original update in Douc *et al.* (2007)

$$\alpha_d^{t+1,N} = \sum_{i=1}^{N} \bar{\omega}_{i,t} \mathbb{1}\{Z_{i,t} = d\} \, ,$$

$$\theta_d^{t+1,N} = \arg\max_{\theta_d} \left[\sum_{i=1}^{N} \bar{\omega}_{i,t} \mathbb{1}\{Z_{i,t} = d\} \log \left\{ q_d \left(X_{i,t}; \theta_d^{t,N} \right) \right\} \right]$$

but Rao-Blackwellisation = stabilisation with insignificant additional computation cost [O$(D \times N)$]

Convergence properties

Convergence of the estimated parameters follows same approach as in Douc et al. (2007a,b)

Under conditions
$\pi(q_d(\cdot;\theta_d)=0)=0$, $\rho_d(\cdot;\alpha,\theta)\log q_d(\cdot,\theta_d)\in L^1(\pi)$,
and some regularity conditions on $q_d(x;\theta)$

$$\alpha_d^{t+1,N} \xrightarrow{\mathbb{P}} \alpha_d^{t+1}, \qquad \theta_d^{t+1,N} \xrightarrow{\mathbb{P}} \theta_d^{t+1}$$

when N goes to infinity.

Monitoring by perplexity

Normalised perplexity $\exp(H^{t,N})/N$ where

$$H^{t,N} = -\sum_{i=1}^{N} \bar{\omega}_{i,t} \log \bar{\omega}_{i,t}$$

Shannon entropy of the normalised IS weights.
Provides an estimate of

$$\exp[-\mathfrak{E}(\pi, q_{(\alpha^{t,N},\theta^{t,N})})]$$

Interrupt adaptation when perplexity stabilises and/or becomes sufficiently close to 1

Mixtures of t's

Setting of

$$\sum_{d=1}^{D} \alpha_d \mathcal{T}(\nu_d, \mu_d, \Sigma_d)$$

[West, 1992; Oh and Berger, 1993]

Mixtures of t's

Setting of

$$\sum_{d=1}^{D} \alpha_d \mathcal{T}(\nu_d, \mu_d, \Sigma_d)$$

[West, 1992; Oh and Berger, 1993]

Completion based on

$$\begin{aligned} f(x,y,z) &\propto \alpha_z |\Sigma_z|^{-1/2} \exp\left\{-(x-\mu_z)^{\mathsf{T}}\Sigma_z^{-1}(x-\mu_z)y/2\nu_z\right\} y^{(\nu_z+p)/2-1} e^{-y/2} \\ &\propto \alpha_z\, \varphi(x;\mu_z,\nu_z\Sigma_z/y)\, \varsigma(y;\nu_z/2,1/2)\,, \end{aligned}$$

using classical normal/χ^2 decomposition of the t distribution

Parameter update

If

$$\rho_d(X;\alpha^t,\theta^t) = \mathbb{P}_{\alpha^t,\theta^t}(Z=d|X) = \frac{\alpha_d^t t(x;\nu_d,\mu_d^t,\Sigma_d^t)}{\sum_{\ell=1}^D \alpha_\ell^t t(x;\nu_\ell,\mu_\ell^t,\Sigma_\ell^t)}\,,$$

and

$$\gamma_d(X;\theta^t) = \mathbb{E}_{\theta^t}^Y\left\{Y/\nu_d|X,Z=d\right\} = \frac{\nu_d+p}{\nu_d+(X-\mu_d^t)^\mathsf{T}(\Sigma_d^t)^{-1}(X-\mu_d^t)}\,.$$

then

$$\alpha_d^{t+1,N} = \sum_{i=1}^N \bar\omega_{i,t}\,\rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\,,$$

$$\mu_d^{t+1,N} = \frac{\sum_{i=1}^N \bar\omega_{i,t}\,\rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\,\gamma_d(X_{i,t};\theta^{t,N})\,X_{i,t}}{\sum_{i=1}^N \bar\omega_{i,t}\,\rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\,\gamma_d(X_{i,t};\theta^{t,N})}\,,$$

$$\Sigma_d^{t+1,N} = \frac{\sum_{i=1}^N \bar\omega_{i,t}\,\rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\,\gamma_d(X_{i,t};\theta^{t,N})\,(X_{i,t}-\mu_d^{t+1,N})(X_{i,t}-\mu_d^{t+1,N})^\mathsf{T}}{\sum_{i=1}^N \bar\omega_{i,t}\,\rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})}\,,$$

Pima Indian Example

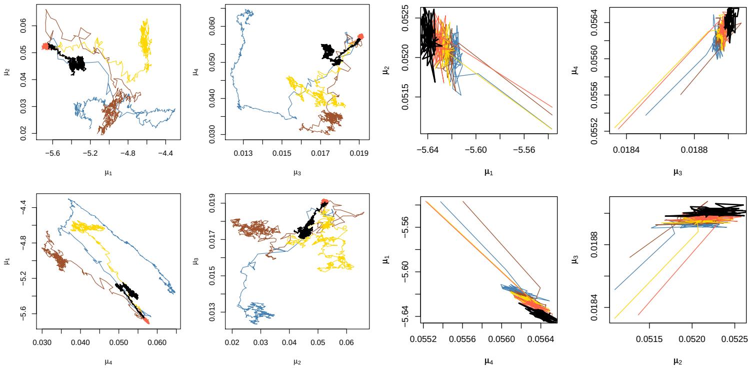
Posterior probit for R pima benchmark:

$$\mathbb{P}_\beta(y=1|\mathbf{x}) = 1 - \mathbb{P}_\beta(y=0|\mathbf{x}) = \Phi(\beta_0 + \mathbf{x}^\mathsf{T}(\beta_1,\beta_2,\beta_3,\beta_4))$$

i.e., target

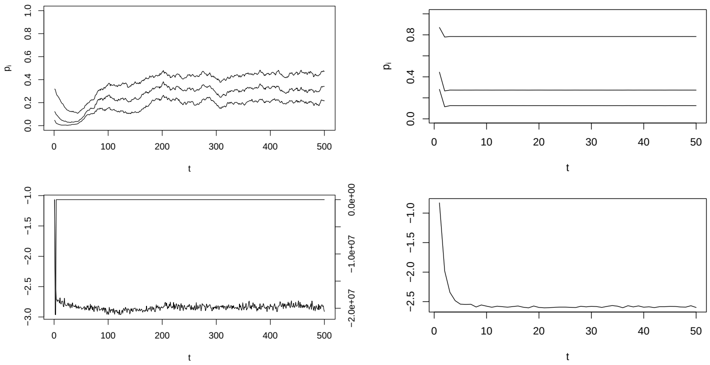
$$\pi(\beta|\mathbf{y},\mathbf{X}) \propto \prod_{i=1}^{532} \left[\Phi\{\beta_0 + (\mathbf{x}^i)^\mathsf{T}(\beta_1,\beta_2,\beta_3,\beta_4)\}\right]^{y_i} \left[1 - \Phi\{\beta_0 + (\mathbf{x}^i)^\mathsf{T}(\beta_1,\beta_2,\beta_3,\beta_4)\}\right[^{1-y_i}$$

Pima Indian Example (cont'd)



Evolution of the components of the five μ_d's over 500 iterations The colour code is blue for μ_1, yellow for μ_2, brown for μ_3 and red for μ_4. The additional dark path corresponds to the estimate of β. All μ_d's were started in the vicinity of the MLE $\hat{\beta}$. *(Left) No RB, (Right) RB*

Pima Indian Example (cont'd)



Evolution of the cumulated weights *(top)* and of the estimated entropy divergence $\mathbb{E}^{\pi}[\log(q_{\alpha,\theta}(\beta))]$ *(bottom)*