Convergence of adaptative sampling schemes

Jean-Michel Marin

Project INRIA SELECT, University Paris-Sud and CEREMADE,
University Paris Dauphine

joint with Randal Douc, Arnaud Guillin and Christian Robert

Introduction

Let $(\mathbf{X}, \mathcal{B}(\mathbf{X}), \Pi)$ be a probability space.

(A1) $\Pi \ll \mu$ and $\Pi(dx) = \pi(x)\mu(dx)$.

(A2) π is known up to a normalizing constant:

- $\pi(x) = \dfrac{\tilde{\pi}(x)}{\displaystyle\int \tilde{\pi}(x)\mu(dx)}$;
- $\tilde{\pi}$ is known;
- the calculation of $\displaystyle\int \tilde{\pi}(x)\mu(dx) < \infty$ is intractable.

Problem: for some Π-measurable applications h, approximate

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

(A3) the calculation of $\displaystyle\int h(x)\tilde{\pi}(x)\mu(dx)$ is intractable.

Applications in Bayesian statistic: $(\pi(\theta|x) \propto f(x|\theta)\pi(\theta))$.

The Monte Carlo framework

1) Monte Carlo methods (MC)

$\Longrightarrow$ Generate an iid sample $x_1, \ldots, x_N$ from Π and to estimate $\Pi(h)$ by

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

$$\hat{\Pi}_N^{MC}(h) \longrightarrow_{as} \Pi(h)$$

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

$$\sqrt{N}(\hat{\Pi}_N^{MC}(h) - \Pi(h)) \longrightarrow_{\mathcal{L}} \mathcal{N}(0, \Pi((h - \Pi(h))^2)).$$

Often impossible to simulate directly from Π!

2) Markov Chain Monte Carlo methods (MCMC)

$\Longrightarrow$ Generate $x^{(1)}, \ldots, x^{(T)}$ from a Markov chain $(x_t)_{t \in \mathbb{N}}$ with stationary distribution Π and estimate $\Pi(h)$ by

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

Convergence to the stationary distribution could be very slow!

3) Importance sampling

Let Q be a probability distribution on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$. Suppose that $\Pi \ll Q$, $Q \ll \mu$ and that $Q(dx) = q(x)\mu(dx)$:

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

$\Longrightarrow$ Generate an iid sample $x_1, \ldots, x_N$ from Q, called the proposal distribution, and to estimate $\Pi(h)$ by

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

$$\hat{\Pi}^{IS}_{Q,N}(h) \longrightarrow_{as} \Pi(h).$$

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

$$\sqrt{N}(\hat{\Pi}^{IS}_{Q,N}(h) - \Pi(h)) \longrightarrow_{\mathcal{L}} \mathcal{N}\left(0, Q((h\pi/q - \Pi(h))^2)\right).$$

For many h, a sufficient condition for $Q((h\pi/q)^2) < \infty$ is that π/q is bounded.

The normalizing constant of Π is unknown, not possible to use $\hat{\Pi}^{IS}_{Q,N}$.

It is natural to use the self-normalized version of the IS estimator,

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

$$\hat{\Pi}^{SNIS}_{Q,N}(h) \longrightarrow_{as} \Pi(h).$$

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

$$\sqrt{N}(\hat{\Pi}^{SNIS}_{Q,N}(h) - \Pi(h)) \longrightarrow_{\mathcal{L}} \mathcal{N}\left(0, Q((\pi/q)^2(h-\Pi(h))^2)\right).$$

The quality of the SNIS approximation depends on the choice of the proposal distribution Q.

PMC algorithms

The notion of importance sampling can actually be greatly generalized to encompass much more adaptive and local schemes than thought previously.

This extension is to learn from experience, that is, to build an importance sampling function based on the performances of earlier importance sampling proposals.

By introducing a temporal dimension to the selection of the importance function, an adaptive perspective can be achieved at little cost, for a potentially large gain in efficiency.

$$\mathcal{F}_{N,t} = \sigma\left\{(x_{i,j}, J_{i,j})_{1\le i\le N, 0\le j\le t}\right\}, \quad \mathcal{F}^J_{N,t} = \mathcal{F}_{N,t} \bigvee \sigma\left\{(x_{i,t+1})_{1\le i\le N}\right\} (t \ge 0)$$

$$Q_0 \ll \mu \quad (q_0) \quad \text{and} \quad Q_{i,t}(\mathcal{F}_{N,t-1}) \ll \mu \quad (q_{i,t}(\mathcal{F}_{N,t-1}))$$

Generic PMC algorithm

- At time 0,

a) Draw $(x_{i,0})_{1\le i\le N}$ iid according to Q_0 and set $\omega_{i,0} = \{\pi/q_0\}(x_{i,0})$;

b) Conditionally on $\sigma\left\{(x_{i,0})_{1\le i\le N}\right\}$,
generate $(J_{i,0})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar\omega_{i,0})_{1\le i\le N})$.

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

a) Conditionally on $\mathcal{F}_{N,t-1}$, generate independent $x_{i,t} \sim Q_{i,t}(\mathcal{F}_{N,t-1})$ and set $\omega_{i,t} = \{\pi/q_{i,t}(\mathcal{F}_{N,t-1})\}(x_{i,t})$;

b) Conditionally on $\mathcal{F}^J_{N,t-1}$, generate $(J_{i,t})_{1\le i\le N} \overset{\text{iid}}{\sim} \mathcal{M}(1, (\bar\omega_{i,t})_{1\le i\le N})$.

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

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

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

It is also possible to use the cumulated PMC estimator according to the iterations.

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

D-kernel PMC schemes

Idea: construct $Q_{i,t}\left(\mathcal{F}_{N,t-1}\right)$ as a mixture of D different transition kernels whose weights are proportional to their survival rates in the previous resampling step.

Purpose: the algorithm would automatically adapt the mixture to the target distribution

Consider $(Q_d)_{1\leq d\leq D}$ a family of transition kernels on $\mathbf{X}\times\mathcal{B}(\mathbf{X})$:

$$(Q_d(x,\cdot))_{1\leq d\leq D, x\in\mathbf{X}} \ll \mu, \qquad \forall A\in\mathcal{B}(\mathbf{X}), Q_d(x,A)=\int_A q_d(x,x')\mu(dx').$$

D-kernel PMC algorithm

- At time 0, use the same step as in the generic PMC algorithm to produce the sample $(x_{i,0}, J_{i,0})_{1\le i\le N}$ and set $p_d^{1,N} = 1/D$.
- At time t ($t = 1, \dots, T$),

a) Conditionally on $\sigma\left\{(\tilde{K}_{i,t-1})_{1\le i\le N}\right\}$ (if $t \neq 1$), generate

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

b) Conditionally on $\sigma\{(\tilde{x}_{i,t-1}, K_{i,t})_{1\le i\le N}\}$, generate independent

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

and set $\omega_{i,t} = \pi(x_{i,t})/q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t})$;

c) Conditionally on $\sigma\{(\tilde{x}_{i,t-1}, K_{i,t}, x_{i,t})_{1\le i\le N}\}$, 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 $(\tilde{x}_{i,t}, \tilde{K}_{i,t}) = (x_{J_{i,t},t}, K_{J_{i,t},t})\,, \quad p_d^{t+1,N} = N^{-1}\sum_{i=1}^N \mathbb{I}_d(\tilde{K}_{i,t})$.

$$\textbf{(A1)} \quad \forall k \in \{1,\dots,D\}, \Pi \otimes \Pi \left\{\pi(x')/q_k(x,x') < \infty\right\} = 1.$$

We denote by γ_u the uniform distribution on $\{1,\dots,D\}$.

We can then deduce a LLN on the pairs $(x_{i,t}, K_{i,t})$ produced by the previous algorithm

Proposition 1 *Under **(A1)**, for $h \in L^1_{\Pi\otimes\gamma_u}$ and for all $T \geq 1$,*

$$\sum_{i=1}^{N} \bar{\omega}_{i,T} h(x_{i,T}, K_{i,T}) \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}} \Pi \otimes \gamma_u(h).$$

This result is negative because, while it implies that

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

is a convergent estimator of $\pi(h)$, it also shows that, for $T \geq 1$,

$$N^{-1} \sum_{i=1}^{N} \mathbb{I}_{\tilde{K}_{i,T}=k} \xrightarrow[\mathbb{P}]{N\rightarrow\infty} \frac{1}{D}.$$

Therefore, at *each* iteration, the weights of *all* kernels converge to $1/D$ when the number of points in the sample grows to infinity. This translates in the lack of learning properties for the D-kernel PMC algorithm: its properties at iteration 1 and at iteration 10 are the same.

In importance sampling as well as in MCMC settings, the conditioning improvement brought by Rao-Blackwellization may be significant.

In the context of the D-kernel PMC scheme, the Rao-Blackwellization argument is that it is not necessary to use the mixture component in the computation of the importance weight but rather the whole mixture.

The importance weight is therefore

$$\pi(x_{i,t}) \Big/ \sum_{d=1}^{D} p_d^{t,N} q_d(\tilde{x}_{i,t-1}, x_{i,t}) \quad \text{rather than} \quad \pi(x_{i,t}) / q_{K_{i,t}}(\tilde{x}_{i,t-1}, x_{i,t}).$$

Rao-Blackwellized D-kernel PMC algorithm

- At time t ($t = 2, \ldots, T$),

a) Conditionally on $\sigma\left\{(\tilde{K}_{i,t-1})_{1\leq i\leq N}\right\}$, generate

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

b) Conditionally on $\sigma\left\{(\tilde{x}_{i,t-1}, K_{i,t})_{1\leq i\leq N}\right\}$, generate independent

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

and set $\omega_{i,t} = \pi(x_{i,t}) \quad \displaystyle\sum_{k=1}^{D} p_k^{t,N} q_k(\tilde{x}_{i,t-1}, x_{i,t})$;

c) Conditionally on $\sigma\left\{(\tilde{x}_{i,t-1}, \tilde{K}_{i,t-1}, x_{i,t})_{1\leq i\leq N}\right\}$, 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 $(\tilde{x}_{i,t}, \tilde{K}_{i,t}) = (x_{J_{i,t},t}, K_{J_{i,t},t})$, $p_d^{t+1,N} = N^{-1}\sum_{i=1}^{N} \mathbb{I}_d(\tilde{K}_{i,t})$.

(A2) $\forall (i,j,k,l) \in \{1,\dots,D\}^4, \int \left[\frac{q_k(x,x')q_l(x,x')}{q_i(x,x')q_j(x,x')}\right] \Pi \otimes \Pi(dx,dx') < \infty.$

Proposition 2 *Under **(A1)** and **(A2)**, for $h \in L^1_\Pi$ and for all $T \geq 1$, then*

$$\frac{1}{N}\sum_{k=1}^{N} \bar{\omega}_{i,T} h(x_{i,T}) \overset{N\to\infty}{\longrightarrow_{\mathbb{P}}} \Pi(h)$$

$$p_d^{T,N} \overset{N\to\infty}{\longrightarrow_{\mathbb{P}}} p_d^T$$

and the limiting coefficients $(p_d^T)_{1\leq d\leq D}$ are defined recursively by

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

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

$\forall \alpha \in S$, let us denote by $\mathrm{KL}(\alpha)$ the Kullback-Leibler divergence between $\Pi(dx) \displaystyle\sum_{d=1}^{D} \alpha_d Q_d(x, dx')$ and $\Pi \otimes \Pi(dx, dx')$:

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

Kullback-Leibler divergence criterion: the best mixture of transition kernels is the one that minimizes $\mathrm{KL}(\alpha)$.

Link between this criteria and the Rao-Blackwellized PMC algorithm?

We define F as the function on S such that

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

and construct the sequence on S

$$\begin{cases} \alpha^0 = (1/D, \ldots, 1/D) & \\ \alpha^{t+1} = F(\alpha^t) & \text{for } t \ge 0 \end{cases}$$

F corresponds to

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

(A3) $\forall i \in \{1, \ldots, D\}, -\infty < \int [\log(q_i(x, x'))] \, \Pi \otimes \Pi(dx, dx') < \infty.$

Proposition 3 *Under **(A3)**, for all $\alpha \in S$,*

$$KL \otimes F(\alpha) \leq KL(\alpha).$$

Therefore, the Kullback Leibler divergence criterion decreases at each step. This property is closely linked with the EM algorithm.

$$\alpha^{\max} = \arg\max_{\alpha \in S} KL(\alpha)$$

Proposition 4 *Under **(A3)**,*

$$\lim_{t \to \infty} \alpha^t = \alpha^{\max}$$

$$\begin{cases} \alpha^0 = (1/D, \ldots, 1/D) & \\ \alpha^{t+1} = F(\alpha^t) & \text{for } t \geq 0 \end{cases}$$

Proposition 5 *Under* ***(A1)*** *and* ***(A2)****, for* $h \in L^2_\Pi$ *and for all* $T \geq 1$*,*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\bar{\omega}_{i,T} h(x_{i,T}) - \Pi(h)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_T^2)$$

where

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

A toy Bayesian study: a Gaussian mixture model

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

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

where $p \neq 1/2$ and σ^2 are known parameters.

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

$$\mu(\mu_1, \mu_2) = \pi\left(\mu_1, \mu_2 | \underline{y}\right) \propto \left(f_{\mathcal{N}(\alpha,\sigma^2/\delta)}(\mu_1) f_{\mathcal{N}(\alpha,\sigma^2/\delta)}(\mu_2) \prod_{i=1}^{n} f\left(y_i | \mu_1, \mu_2\right) \right)$$

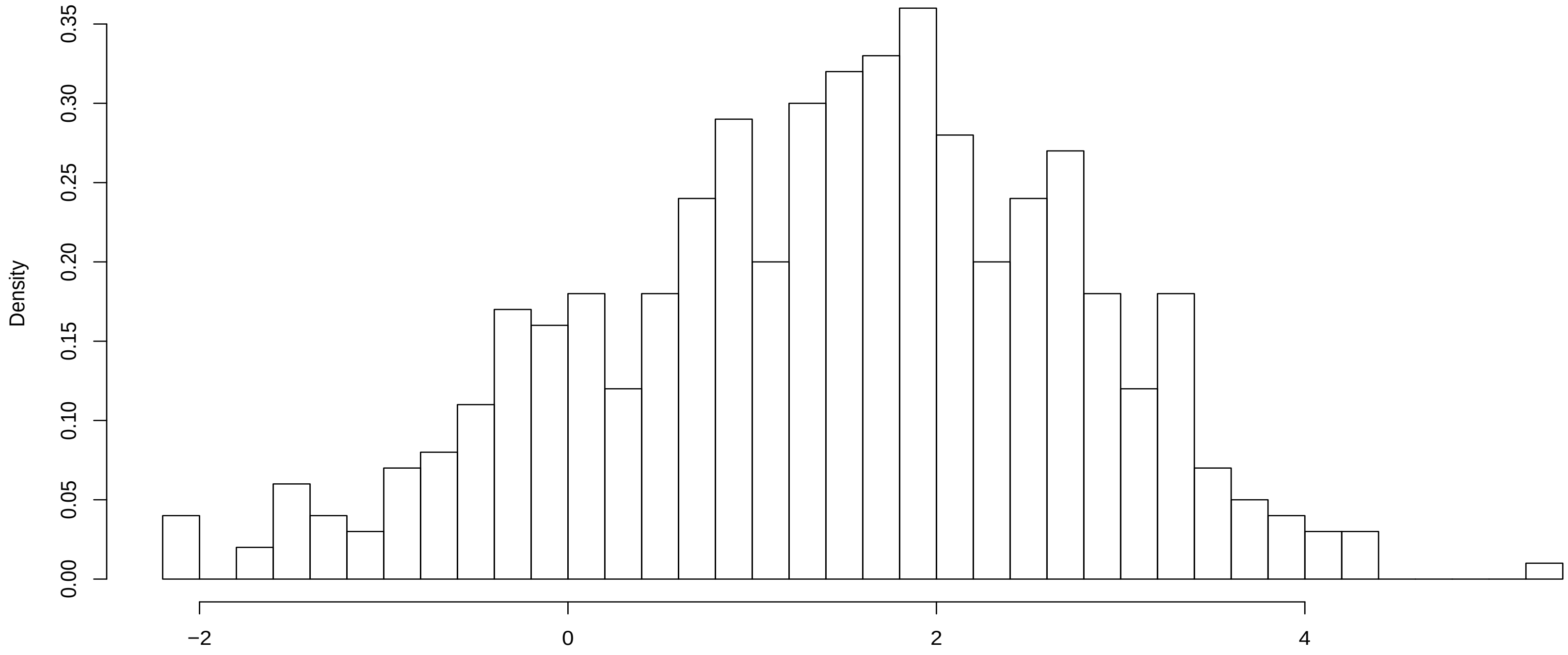

Figure 1: Simulated data $n = 500$, $p = 0.3$, $\sigma^2 = 1$, $\mu_1 = 0$ and $\mu_2 = 2$

Use of the D-kernels PMC algorithm.

After an arbitrary initialization, use of the previous (importance) sample (after resampling) to build random walk proposals,

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

with a multi-scale variance v_i within a predetermined set of D scales, whose importance is proportional to its survival rate in the resampling step.

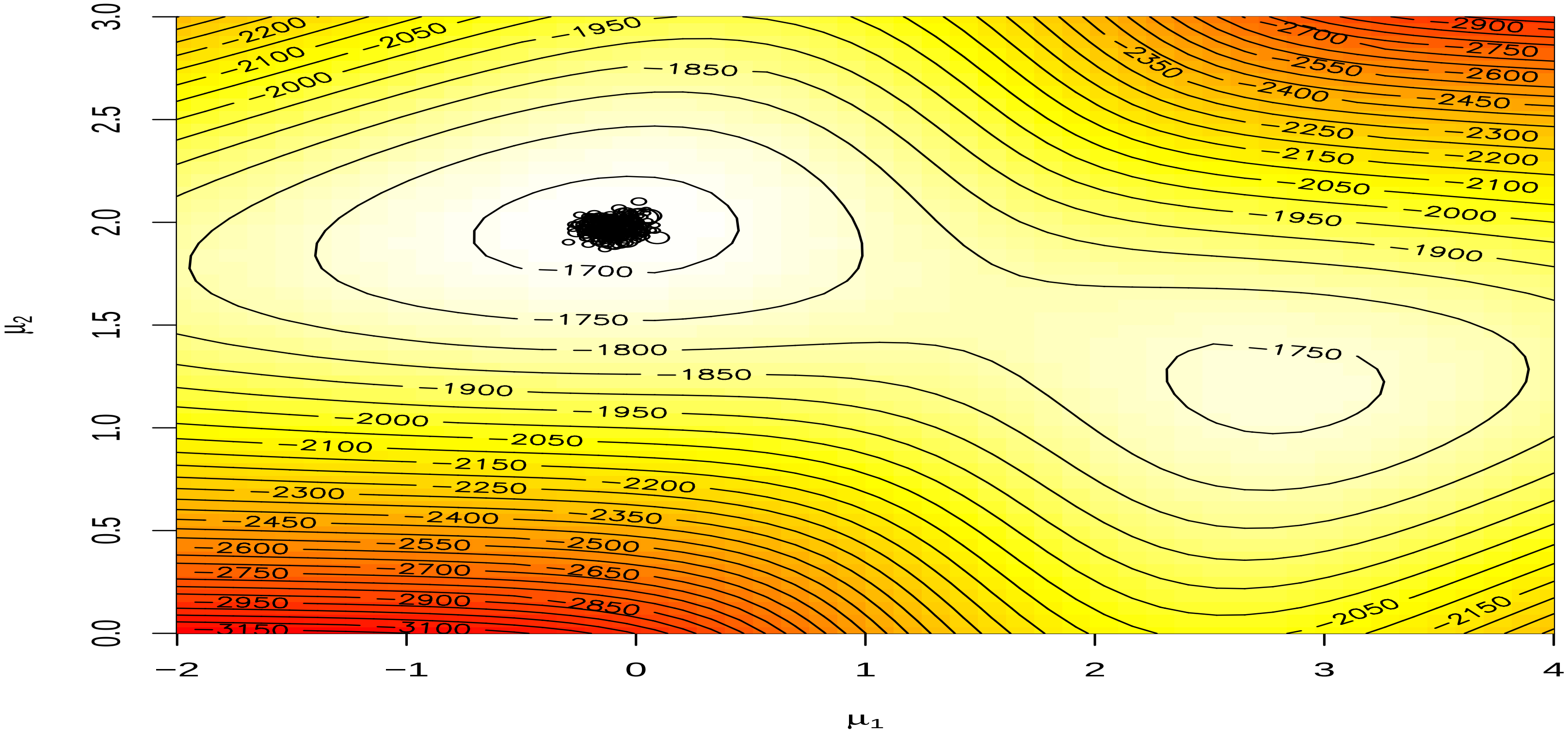

Figure 2: PMC weighted sample (n=1000) after 10 iterations (the weights are proportional to the circles at each point).

Conclusion

- The PMC scheme is a viable alternative to MCMC schemes.

- The iterative nature of PMC erodes the dependence to the importance function by offering a wide range of adaptive kernels that can take advantage of the previously simulated samples.