Centre de Recherche

SAMOS

Statistique Appliquée &
Modélisation Stochastique

On constrained simulation and optimization by Metropolis chains

J.F. YAO

Prépublication du SAMOS N° 66 (Version révisée)

novembre 1998

On constrained simulation and optimization by Metropolis chains

J. YAO*   SAMOS, Université Paris 1

Abstract

Given a everywhere positive probability measure π on a finite state space E and the associated energy function H, we propose time-inhomogeneous Metropolis chains to simulate π and to minimize H under some constraints.

Keywords : Constrained simulation; constrained optimization; Metropolis algorithm

1 Introduction and notations

Let π be a everywhere positive probability measure π on a finite state space E given by $\pi(x) = Z^{-1}\exp[-H(x)]$ where $H : E \to \mathbb{R}$ is the associated *energy* function. Assume that we are interested in some subset $E_c \subset E$ defined by a constrained equation $E_c := \{J = 0\}$ where the function of constraints $J : E \to \mathbb{R}^+$ is otherwise positive. Let π_c be the restriction of π on E_c : $\pi_c(x) = 1_{E_c}(x) Z_c^{-1}\exp[-H(x)]$ (throughout the note, the letter Z is used to denote generic normalizing factors). We consider the following two problems:

(i) Problem **[S]**: simulate the distribution π_c.

(ii) Problem **[M]**: minimize H over E_c.

We propose to solve these problems by Markov chains endowed with a time-inhomogeneous Metropolis dynamic. Let (β_k), (λ_k) be two positive non decreasing sequences and q a *symmetric* and *irreducible* Markov transition kernel on E and set

$$H_k(x) := \beta_k[H(x) + \lambda_k J(x)] \, , k \geq 1 \, . \tag{1}$$

Let us consider a inhomogeneous Markov chain $\mathbf{X} := (X(k))_{k\geq 0}$ with state space E whose transition probabilities at time k is given by the following Metropolis rule (we write a^+ for $\max(a, 0)$):

$$P_k(x, y) = \begin{cases} q(x, y) e^{-[H_k(y) - H_k(x)]^+}, & y \neq x, \\ 1 - \displaystyle\sum_{z \neq x} P_k(x, z), & y = x. \end{cases} \tag{2}$$

It is well-known that for each k, P_k is irreducible and reversible with respect to the probability measure (p.m.)

$$\pi_k(x) = Z_k^{-1}\exp[-H_k(x)] \, ,$$

and consequently, π_k is the unique invariant probability measure (i.p.m) of P_k. Let us set

$$H_0 \ := \ \min\{H(x) : x \in E_c\}, \quad E_0 := \{x \in E_c \ : \ H(x) = H_0\},$$

*SAMOS, Université Paris 1, 90 rue de Tolbiac, F-75634 Paris Cedex 13, France (yao@univ-paris1.fr)

and denote by π_0 the uniform distribution on E_0. It is straightforward to find that

$$\text{with } \beta_k \equiv 1, \lambda_k \uparrow \infty, \qquad \lim_{k\to\infty} \pi_k(x) = \pi_c(x), \qquad x \in E, \tag{3}$$

$$\text{with } \beta_k \uparrow \infty, \lambda_k \uparrow \infty, \qquad \lim_{k\to\infty} \pi_k(x) = \pi_0(x), \qquad x \in E. \tag{4}$$

Let us set $P^{(m,k)} = P_{m+1}\cdots P_k$, the transition probabilities from time m to k ($k > m$). The aim of this note is to establish conditions on the control sequences (β_k) and (λ_k) which guarantee that the chain $\mathbf{X}$ is strongly ergodic in the following sense ((Isaacson and Madsen, 1976)): there is some p.m. π_∞ on E such that

$$\text{for all } m \geq 1, \quad \lim_{k\to\infty} \sup_\mu ||\mu P^{m,k} - \pi_\infty|| = 0 \,. \tag{5}$$

Here $||\cdot||$ is the total variation distance and the supremum is taken over all p.m.'s on E. The target measure π_∞ will be π_c or π_0 according to the considered problem.

When E is a product spaces, $E = \Pi_{s\in S} F_s$, (Geman, 1990) have proposed a solution to these problems by using Gibbs sampling. Our approach, as well as the their, is based on Dobrushin's ergodicity coefficients ((Dobrushin, 1956)). More precisely, the ergodicity coefficient of a Markov kernel Q is defined as

$$a(Q) := \min_{x,x'\in E} \sum_{y\in E} Q(x,y) \wedge Q(x',y). \tag{6}$$

By taking into account the convergences (3)-(4), a well-known result (see e.g. (Isaacson and Madsen, 1976)) implies that the ergodicity (5) is ensured if the following two conditions are satisfied:

(C1) $\sum_{k\geq 0} ||\pi_{k+1} - \pi_k|| < \infty$;

(C2) for some positive integer p, $\sum_{k=1}^{\infty} a(P_{kp+1}\cdots P_{(k+1)p}) = \infty$.

For $\beta \geq 0, \lambda \geq 0$ and $x \in E$, let us define

$$\Pi(x;\beta,\lambda) := \frac{1}{Z_{\beta,\lambda}} \exp^{-\beta[H(x)+\lambda J(x)]} \,,$$

so that $\pi_k(\cdot) = \Pi(\cdot;\beta_k,\lambda_k)$. The underlying expectations are denoted $\mathbb{E}_{\beta,\lambda}$ and $\mathbb{E}_k$ respectively, with $\mathbb{E}_\lambda = \mathbb{E}_{1,\lambda}$.

To solve the problem **[S]**, the sequence (β_k) is kept constant: $\beta_k \equiv 1$ while (λ_k) will be a sequence of positive numbers increasing to infinity. On the other hand, to solve the problem **[M]**, both the sequences are required to increase to infinity (see Eqs. (3)-(4)). We examine in more details the conditions **(C1)-(C2)** for these two problems.

2 Condition (C1) on invariant probability measures (π_k)

The idea is to show that for large enough k, the sequence $[\pi_k(x)]$ is increasing for x belonging to the target set E_c or E_0 and decreasing otherwise. In this case, since

$$\sum_{k\geq 0} ||\pi_{k+1} - \pi_k|| = \sum_{x\in E} |\pi_{k+1} - \pi_k| \,,$$

this sum is thus finite.

The simulation problem [S]. Recall that $\beta_k \equiv 1, \lambda_k \uparrow \infty$ for this problem. For $x \in E$, let us set $\varphi_x(\lambda) := \log \Pi(x; 1, \lambda)$. We have

$$\varphi_x'(\lambda) := \frac{\partial}{\partial \lambda}\varphi_x(\lambda) = \mathbb{E}_\lambda[J - J(x)] = \frac{\sum_{a\in E}[J(a) - J(x)]\exp\{-[H(a)+\lambda J(a)]\}}{\sum_{a\in E}\exp\{-[H(a)+\lambda J(a)]\}}\,.$$

If $\lambda \to \infty$ and $x \notin E_c$, $\varphi_x'(\lambda)$ tends to $-J(x) < 0$. On the other hand, for all $x \in E_c$, $\varphi_x'(\lambda) = \mathbb{E}_\lambda[J] \geq 0$. Hence for large enough k, we have $\pi_{k+1}(x) \leq \pi_k(x)$ for $x \notin E_c$ and $\pi_{k+1}(x) \geq \pi_k(x)$ for $x \in E_c$.

The optimization problem [M]. Here we have $\beta_k \uparrow \infty$ and $\lambda_k \uparrow \infty$. For $x \in E$, let be $\varphi_x(\beta,\lambda) := \log \Pi(x;\beta,\lambda)$. We have

$$\left\{\begin{array}{l} \frac{\partial}{\partial \beta}\varphi_x(\beta,\lambda) = \mathbb{E}_{\beta,\lambda}[H + \lambda J - (H+\lambda J)(x)] \\ \\ \frac{\partial}{\partial \lambda}\varphi_x(\beta,\lambda) = \beta\mathbb{E}_{\beta,\lambda}[J - J(x)] \end{array}\right. .$$

We will show that for large enough β, λ, the function $(\beta,\lambda) \mapsto \varphi_x(\beta,\lambda)$ is coordinatewise decreasing for $x \notin E_0$ and coordinatewise increasing for $x \in E_0$. So, let $\lambda \to \infty$, $\beta \to \infty$:

- For $x \notin E_c$, $\dfrac{\partial}{\partial \beta}\varphi_x(\beta,\lambda) \sim H_0 - H(x) - \lambda J(x)$ and $\dfrac{\partial}{\partial\lambda}\varphi_x(\beta,\lambda) \sim -\beta J(x)$ which are both negative. Therefore $\varphi_x(\beta,\lambda)$ is coordinatewise decreasing for large enough β and λ.
- For $x \in E_0$, $\dfrac{\partial}{\partial\beta}\varphi_x(\beta,\lambda) = \mathbb{E}_{\beta,\lambda}[H] - H_0 + \lambda\mathbb{E}_{\lambda\beta}[J] \geq 0$ and $\dfrac{\partial}{\partial\lambda}\varphi_x(\beta,\lambda) = \beta\mathbb{E}_{\beta,\lambda}[J] \geq 0$. It follows that $\varphi_x(\beta,\lambda)$ is coordinatewise increasing for large enough β and λ.
- For $x \in E_c \backslash E_0$, , $\dfrac{\partial}{\partial\lambda}\varphi_x(\beta,\lambda) = \beta\mathbb{E}_{\beta,\lambda}[J] \geq 0$ and $\dfrac{\partial}{\partial\beta}\varphi_x(\beta,\lambda) = \mathbb{E}_{\beta,\lambda}[H+\lambda J] - H(x)$ that tends to $H_0 - H(x) < 0$. The situation is here a little more complicated since these two derivatives have an opposite sign. However, we have

$$\mathbb{E}_{\beta,\lambda}[J] \sim \frac{J^*|E^*|}{|E_0|}e^{-\beta(H^*+\lambda J^*)}, \qquad \text{when } \lambda\to\infty, \beta\to\infty, \tag{7}$$

where the constants are:

$$\begin{aligned} J^* &= \min\{J(x) \,:\, x \notin E_c\} && (8)\\ H^* &= \min\{H(x) - H_0 \,:\, x\notin E_c,\ J(x) = J^*\} && (9)\\ E^* &= \{x \,:\, x \notin E_c,\ J(x) = J^*,\ H(x) - H_0 = H^*\}\,. && (10) \end{aligned}$$

Let us take two positive, differentiable and increasing functions $\beta(\cdot)$, $\lambda(\cdot)$ defined on $[0,\infty)$ such that $\beta_k = \beta(k)$ and $\lambda_k = \lambda(k)$. Then

$$\varphi_x'(u) := \frac{\partial}{\partial u}\varphi_x(\beta(u),\lambda(u)) = \beta'(u)\left[\mathbb{E}_{\beta,\lambda}(H) - H(x) + \left\{\lambda(u) + \beta(u)\frac{\lambda'(u)}{\beta'(u)}\right\}\mathbb{E}_{\lambda\beta}(J)\right]. \tag{11}$$

Note that $\mathbb{E}_{\lambda\beta}(H) - H(x) \sim H_0 - H(x) < 0$ and let us define

$$H_1 = \min\ \{H(x) - H_0 \,:\, x \in E_c\backslash E_0\}\,. \tag{12}$$

Now assume that the following condition is fulfilled:

$$\limsup_{u\to\infty} \beta(u)\frac{\lambda'(u)}{\beta'(u)}e^{-\beta(u)\{H^*+\lambda(u)J^*\}} < \frac{H_1|E_0|}{J^*|E^*|}\,. \tag{13}$$

In this case, the derivative $\varphi'_x(u)$ is negative for large enough u. Hence $k \mapsto \varphi_x(\beta(k), \lambda(k))$ is decreasing for large enough k.

To summarize, for the minimization problem **[M]**, under the condition (13) and for large enough k, $\pi_k(x)$ is decreasing for $x \notin E_0$ and increasing for $x \in E_0$.

Remark 1 *For logarithmic functions $\beta(u) \sim \log^a(u)$ and $\lambda(u) \sim \log^b(u)$ (when $u \to \infty$) with $a > 0,\ b > 0$, we have:*

$$\lim_{u\to\infty} \beta(u)\frac{\lambda'(u)}{\beta'(u)}e^{-\beta(u)\{H^*+\lambda(u)J^*\}} \to 0. \tag{14}$$

The condition (13) is thus satisfied. The same is true for functions $\lambda(\cdot)$ and $\beta(\cdot)$ satisfying for some positive constants $a,\ A,\ b,\ B$:

$$\limsup_{u\to\infty} \frac{\lambda'(u)}{\lambda^a(u)} \le A, \qquad \limsup_{u\to\infty} \frac{1}{\beta'(u)\beta^b(u)} \le B. \tag{15}$$

3 Condition (C2) on the ergodicity coefficients

We write $Q > 0$ for a Markov kernel Q on E satisfying $Q(x, y) > 0,\ \forall(x, y)$. Let be:

$$\begin{aligned}
\delta_H &:= \max\left\{[H(y) - H(x)]^+ : \quad x, y \in E, \quad q(x, y) > 0\right\},\\
\delta_J &:= \max\left\{[J(y) - J(x)]^+ : \quad x, y \in E, \quad q(x, y) > 0\right\},\\
\gamma &:= \min\left\{q(x, y) : \quad x, y \in E, \quad q(x, y) > 0\right\},\\
\partial x &:= \left\{y \in E : \quad q(x, y) > 0\right\}.
\end{aligned}$$

We first prove two auxiliary lemmas.

Lemma 1 *Consider the Metropolis algorithm with transition probabilities (P_k) defined in (2) For both the problems **[S]** and **[M]**, there are two positive integers p, k_0 such that*

$$\textit{for all}\ \ k \ge k_0,\ \ P^{(k,k+p)} > 0\ . \tag{16}$$

Proof. *Case 1.* If the kernel q is aperiodic, q is positive recurrent since E is finite. Therefore there is an integer $p > 0$ such that $q^p > 0$. Since for all x, y and j,

$$P_j(x, y) \ge q(x, y)\ e^{-[H_j(y)-H_j(x)]^+} \ge q(x, y)\ e^{-\beta_j(\delta_H+\lambda_j\delta_J)}\ ,$$

we get for all k,

$$P^{(k,k+p)}(x, y) \ge q^p(x, y)\ e^{-(\beta_{k+1}+\cdots+\beta_{k+p})(\delta_H+\lambda_k\delta_J)} > 0.$$

Case 2. Otherwise, let q be a periodic kernel with period d. Since q is irreducible, there is a partition $E = E_1 + \cdots + E_d$ such that the iterated kernel q^d is positive recurrent on each subclass E_j. We can then find a s_0 such that for all $s \geq s_0$, $q^{sd}(x, y) > 0$ if (and only if) x, y are two points in a same subclass.

On the other hand, let us take a point $a \in E_c$ for which there is a $b \in \partial a$ but not in E_c. This is always possible since q is irreducible and $E_c \neq E$. Therefore, $H_k(a) = \beta_k H(a) < \beta_k[H(b) + \lambda_k J(b)]$ for large enough k for both the two problems. Hence there is a k_0 such that $P_k(a, a) > 0$ for all $k \geq k_0$. It follows that for $p = 2s_0 d + d - 1$ and $k \geq k_0$, we have $P^{(k,k+p)}(x, y) > 0$, for all x, y. ■

In the sequel, we will write p for the smallest integer satisfying (16).

Lemma 2 *With the same assumptions made in Lemma 1, there is an integer k_1 such that for all $k \geq k_1$, we have*

$$a\left(P^{(k,k+p)}\right) \geq \gamma^p |E| \; e^{-p\beta_{k+p}\left(\delta_H + \lambda_{k+p}\delta_J\right)}. \tag{17}$$

Proof. By Lemma 1, for $k \geq k_0$ and $(a, b) \in E^2$, there is a path $a = a_0 \to a_1 \to \cdots \to a_p = b$ for which $P_{k+p-j}(a_j, a_{j+1}) > 0,\ 0 \leq j < p$. It is thus enough to prove that for large enough k

$$P_k(x, y) > 0 \quad \text{implies that} \quad P_k(x, y) \geq \gamma \; e^{-\beta_k(\delta_H + \lambda_k \delta_J)}\ , \tag{18}$$

since in this case, we actually have $P^{(k,k+p)}(a, b) \geq \gamma^p \; e^{-p\beta_{k+p}\left(\delta_H + \lambda_{k+p}\delta_J\right)}$, and the required result (17) follows.

Let us prove (18). For $x \neq y$ or $x = y$ with $q(x, x) > 0$, we have for all k,

$$P_k(x, y) \geq q(x, y) e^{-[H_k(y) - H_k(x)]^+} \geq \gamma \; e^{-\beta_k(\delta_H + \lambda_k \delta_J)}.$$

A more intricate case concerns those transition probabilities $P_k(x, x)$ for which $q(x, x) = 0$. Let us recall that

$$0 < P_k(x, x) = \sum_{z \neq x} q(x, z) \; \left(1 - e^{-[H_k(z) - H_k(x)]^+}\right).$$

There is then some $z_0 \in \partial x$ s.t. $H_k(z_0) - H_k(x) = \beta_k \left\{H(z_0) - H(x) + \lambda_k[J(z_0) - J(x)]\right\} > 0$. On the other hand, if (x', z') are some neighboring points belonging to the set

$$G := \{(x', z') \ : \ z' \in \partial x',\ J(z') < J(x') \quad \text{or} \quad \{J(z') = J(x') \quad \text{and} \quad H(z') \leq H(x')\} \ \},$$

there is an integer $m(x', z')$ such that $H_k(z') \leq H_k(x')$ for all $k \geq m(x', z')$. Let be $m^* := \max_{(x',z') \in G} m(x', z')$. For z_0 defined above and $k \geq m^*$, since $H_k(z_0) - H_k(x) > 0$, $(x, z_0) \notin G$ and hence one of the following holds

$$(i).\ J(z_0) > J(x), \quad \text{or} \quad (ii).\ J(z_0) = J(x) \quad \text{with} \quad H(z_0) > H(x).$$

In the first case, we get $H_k(z_0) - H_k(x) \to \infty$ and

$$P_k(x, x) \geq q(x, z_0)[1 - e^{-[H_k(z_0) - H_k(x)]}] \to q(x, z_0).$$

And in the second case, we have $H_k(z_0) - H_k(x) \geq \beta_1[H(z_0) - H(x)]$ and

$$P_k(x, x) \geq q(x, z_0)[1 - e^{-[H_k(z_0) - H_k(x)]}] \geq q(x, z_0)[1 - e^{-\beta_1[H(z_0) - H(x)]}].$$

Since these lower bounds are positive and independent of k, and $\gamma\ e^{-\beta_k(\delta_H+\lambda_k\delta_J)} \to 0$, there is an integer $n(x) \geq m^*$ such that for all $k \geq n(x)$,

$$P_k(x,x) \geq \gamma\ e^{-\beta_k(\delta_H+\lambda_k\delta_J)}\ .$$

Taking $k_1 = k_0 \vee \max_x n(x)$ proves the inequality (18). ■

An immediate consequence of the last lemma is that the condition (**C2**) is fulfilled if

$$\sum_{k=1}^{\infty} e^{-p\beta_{(k+1)p}[\delta_H+\lambda_{(k+1)p}\delta_J]} = \infty. \tag{19}$$

4 Main results

Summarizing the results established in the previous sections gives the following main results of this note.

Theorem 1 *Consider the Metropolis algorithm with (P_k) defined in (2) with $\beta_k \equiv 1$ and $\lambda_k \uparrow \infty$. Let p be the smallest integer satisfying (16). The underlying inhomogeneous Markov chain $\mathbf{X} = (X(k))_{k\geq 0}$ is strongly ergodic with limit p.d.m. π_c if the control sequence (λ_k) fulfills the following condition :*

$$\sum_{k=1}^{\infty} e^{-p\lambda_{kp}\delta_J} = \infty. \tag{20}$$

Theorem 2 *Consider the Metropolis algorithm with (P_k) defined in (2) with $\beta_k \uparrow \infty$ and $\lambda_k \uparrow \infty$. Let p be the smallest integer satisfying (16). The underlying inhomogeneous Markov chain $\mathbf{X} = (X(k))_{k\geq 0}$ is strongly ergodic with limit p.d.m. π_0 if the control sequences (β_k) and (λ_k) fulfill the conditions (13) et (19).*

Remark 2 *(20) is satisfied for a logarithmic sequence $\lambda_k = c\log(k+D)$ with some constants $D > 0$ and $0 < c \leq (p\delta_J)^{-1}$.*

Remark 3 *The condition (19) is satisfied for sequences $\beta_k \uparrow \infty$, $\lambda_k \uparrow \infty$ satisfying $\beta_k\lambda_k = c\log(k+D)$ with some constants $D > 0$ and $0 < c < (p\delta_J)^{-1}$. Furthermore, the condition (13) is satisfied for e.g. $\beta_k = [c\log(k+D)]^{1/2}$, and $\lambda_k = [c\log(k+D)]^{1/2}$.*

ACKNOWLEDGEMENT. *I thank* Xavier GUYON *and* Alain TROUVÉ *for their helpful comments on this work.*

References

Dobrushin, R. (1956). Central limit theorems for non-stationary markov chains ii. *Theory of Probability and its Applications*, 15:41–51. English translation.

Geman, D. (1990). Random fields and inverse problems in imaging. In Hennequin, P., editor, *École d'Été de Probabilités de Saint-Flour XVIII- 1988*, number 1427 in L.N.M., pages 113–193. Springer-Verlag.

Isaacson, D. and Madsen, R. (1976). *Markov Chains Theory and Applications*. John Wiley & Sons.