PAPER Special Section on Information Theory and Its Applications

On the Twisted Markov Chain of Importance Sampling Simulation

Kenji NAKAGAWA†, *Member*

SUMMARY The importance sampling simulation technique has been exploited to obtain an accurate estimate for a very small probability which is not tractable by the ordinary Monte Carlo simulation. In this paper, we will investigate the simulation for a sample average of an output sequence from a Markov chain. The optimal simulation distribution will be characterized by the Kullback-Leibler divergence of Markov chains and geometric properties of the importance sampling simulation will be presented. As a result, an effective computation method for the optimal simulation distribution will be obtained.

key words: *importance sampling, Markov chain, Kullback-Leibler divergence, large deviations theory, information geometry*

1. Introduction

The Importance Sampling (IS) simulation technique has been exploited to obtain an accurate estimate for a very small probability which is not tractable by the ordinary Monte Carlo (MC) simulation. The IS technique is widely used for various types of engineering problems, e.g., for estimation of the blocking probability in queueing system [2], [4], [7], [12], [13], [15], the error rate in communications system [2], [6], [8], etc. See [2] for overview of the application of the IS simulation.

In case that the target event (a blocking in queueing or an error in communications system) is a rare event with small probability of the order less than 10^{-6}, it is impossible to obtain an estimate by the ordinary MC method. The MC method requires considerable amount of computation time. To overcome this difficulty, the underlying probability distribution is modified to generate more samples in the target event. Then the samples obtained by the modified distribution form an unbiased estimate of the true probability. This estimate is called an IS estimate. The modified distribution is called a simulation distribution. If the simulation distribution is appropriately chosen, the variance of the IS estimate can be smaller than that of the estimate by the ordinary MC simulation. The simulation distribution that yields the IS estimate with the minimum variance is referred to as the optimal simulation distribution. When we apply the IS technique to some simulation problems, it is critical to find the optimal simulation distribution. It has been known [2] that, for Markov chains, the optimal simulation distribution exists in the class of twisted Markov chains (TMC). But, in general, in order to have the optimal simulation distribution, we must compute an eigen value and an eigen vector of a matrix and solve an optimization problem concerned with the eigen value. Thus, it is very hard to obtain the optimal simulation distribution in this way even if we use the numerical computation.

In this paper, we will characterize the TMC's by the Kullback-Leibler (KL) divergence of Markov chains and then clarify the relation between the underlying Markov chain and the optimal TMC. Further the geometric properties of the importance sampling simulation will be presented. As a result, an effective computation method will be obtained for the optimal TMC.

Manuscript received December 21, 1995.

†The author is with the Department of Electrical Engineering, Nagaoka University of Technology, Nagaoka-shi, 940–21 Japan.

2. Importance Sampling Simulation

We will show a fundamental form of the importance sampling (IS) simulation technique.

Let X denote the Gaussian random variable with mean 0 and variance 1, i.e., $X \sim \mathcal{N}(0,1)$. For $c > 0$, let us consider the estimation of the probability $P(X > c)$ by the ordinary Monte Carlo (MC) method. Denote by $\boldsymbol{R}$ the set of real numbers. Define the indicator function $I_c(x)$ of the set $\{x \in \boldsymbol{R} \mid x > c\}$ by

$$I_c(x) = \begin{cases} 1, & x > c, \\ 0, & x \leq c. \end{cases} \tag{1}$$

Let $x_1, x_2, \cdots, x_k$ be k independent realizations of X. We have the MC estimate of $P(X > c)$ as follows;

$$P(X > c) \approx \frac{1}{k} \sum_{i=1}^{k} I_c(x_i). \tag{2}$$

If c is large, the set $\{X > c\}$ is a rare event. Hence, to obtain a stable estimate of $P(X > c)$, the number k of generating random numbers should be large. For example, if $c = 10$, $P(X > 10) = 7.6 \times 10^{-24}$. We must have, at least, $k = 10^{25}$, however, the generation of 10^{25} random numbers is impractical.

There is another limit of the ordinary MC simulation owing to the precision limit of pseudo random numbers. Let us consider to generate $X \sim \mathcal{N}(0,1)$ by the well-known Box-Müller method;

$$X = \sqrt{-2 \log u_1} \cos 2\pi u_2, \tag{3}$$

where u_1 and u_2 are the uniform random numbers on the interval $(0, 1]$ and the log is the natural logarithm. Let U denote the pseudo random integer generated by a computer with range $0 \leqq U \leqq 32767$. Putting $u_1 = (U + 1)/(32767 + 1)$, we apply the Box-Müller method. It is readily seen that $|X| < 4.57 (= \sqrt{-2\log(1/32768)})$. Thus, if $c > 4.57$, no sample X_i is in the range $\{X > c\}$. It is absolutely impossible to have an estimate of $P(X > c)$ by the ordinary MC method.

We here apply the IS technique. Write $X^* = X + c$, i.e., $X^* \sim \mathcal{N}(c, 1)$. Let $p(x)$ and $p^*(x)$ denote the probability density functions of X and X^*, respectively. Let $x_1^*, x_2^*, \cdots, x_k^*$ be k independent realizations of X^*. The IS estimate of $P(X > c)$ is given by

$$P(X > c) \approx \frac{1}{k} \sum_{i=1}^{k} I_c(x_i^*) \frac{p(x_i^*)}{p^*(x_i^*)}. \tag{4}$$

Since many samples x_i^* are in the range $\{X^* > c\}$, the difficulties of the ordinary MC method are overcome. We can see that the IS estimate (4) is an unbiased estimate. It has been proved that $p^*(x)$ is the optimal simulation distribution among the Gaussian distributions [2].

3. IS for Markov Chains

We will investigate the simulation for a sample average of an output sequence from a finite state Markov chain.

Let P_0 denote an irreducible Markov chain on the state space $\Omega \equiv \{0, 1, \cdots, K\}$, $K > 0$. The state transition probability matrix of P_0 is denoted by $P_0 = (P_0(x'|x))_{x,x' \in \Omega}$, and the initial distribution is given by the stationary distribution $p_0 = (p_0(x))_{x \in \Omega}$ of P_0. The joint probability of $x, x' \in \Omega$ is denoted by $P_0(x, x') \equiv p_0(x) P_0(x'|x)$. Consider a mapping $f : \Omega \times \Omega \to \boldsymbol{Z}$, where $\boldsymbol{Z}$ is the set of integers. Denote by $E_{P_0}[f]$ the expectation of f with respect to P_0;

$$E_{P_0}[f] \equiv \sum_{x,x' \in \Omega} P_0(x, x') f(x, x'). \tag{5}$$

Assume $E_{P_0}[f] = 0$ without loss of generality.

Let $\boldsymbol{x}^n = (x_1, x_2, \cdots, x_n) \in \Omega^n$ be a sample sequence generated by the Markov chain P_0. We consider the probability of the following set A_n;

$$A_n = \{\boldsymbol{x}^n \in \Omega^n \mid \frac{1}{n-1} \sum_{i=1}^{n-1} f(x_i, x_{i+1}) > c\}, \quad c > 0. \tag{6}$$

Write $\alpha_n \equiv P_0(A_n) = \sum_{\boldsymbol{x}^n \in A^n} P_0(\boldsymbol{x}^n)$, where $P_0(\boldsymbol{x}^n) = p_0(x_1) P_0(x_2|x_1) \cdots P_0(x_n|x_{n-1})$. We see from large deviations theory [2], [6], that the asymptotic of the α_n is given as follows. For $t \in \boldsymbol{R}$, denote by λ_t the largest eigen value of the matrix

$$(P_0(x'|x) e^{t f(x,x')})_{x,x' \in \Omega}. \tag{7}$$

From the Perron-Frobenius theorem [2], we can see that λ_t is a positive eigen value with multiplicity 1. For $y \in \boldsymbol{R}$, the *rate function* $I(y)$ is defined by

$$I(y) = \sup_{-\infty < t < \infty} (yt - \log \lambda_t). \tag{8}$$

The rate of the exponential decay of α_n is represented by the rate function.

Theorem 1: (Large Deviations Theorem [2]) We have

$$\lim_{n \to \infty} -\frac{1}{n} \log \alpha_n = I(c). \qquad \square$$

The asymptotic behavior of α_n is described by this theorem, however, it is often necessary to have the accurate value of α_n for some finite n. We need to use a stochastic simulation to have α_n.

A *twisted Markov chain (TMC)* [2] for the underlying Markov chain P_0 is the Markov chain P_t defined by

$$P_t(x'|x) = P_0(x'|x) e^{t f(x,x')} \frac{w_t(x')}{\lambda_t w_t(x)}, \quad x, x' \in \Omega, \ t \in \boldsymbol{R}, \tag{9}$$

where $w_t = (w_t(x))_{x \in \Omega}$ is a right eigen vector associated with λ_t.

Consider the IS simulation with a Markov chain P. Let $X^{*n} = (X_1^*, \cdots, X_n^*)$, $n = 1, 2, \cdots$, be a random sequence generated by the Markov chain P, and $\boldsymbol{x}_1^{*n}, \boldsymbol{x}_2^{*n}, \cdots, \boldsymbol{x}_k^{*n} \in \Omega^n$ be k independent realizations of the random variable X^{*n}. The IS estimate $\widehat{\alpha}_n(P)$ of α_n with simulation distribution P is given by

$$\widehat{\alpha}_n(P) = \frac{1}{k} \sum_{j=1}^{k} I_{A_n}(\boldsymbol{x}_j^{*n}) \frac{P_0(\boldsymbol{x}_j^{*n})}{P(\boldsymbol{x}_j^{*n})}, \ n = 1, 2, \cdots, \tag{10}$$

where I_{A_n} is the indicator function of the set A_n. We will see in next chapter that the TMC is a candidate of the optimal simulation distribution.

4. The Optimal Simulation Distribution

Since the estimate $\widehat{\alpha}_n(P)$ of (10) is an unbiased estimate of α_n, we will next investigate the variance $V[\widehat{\alpha}_n(P)]$ of $\widehat{\alpha}_n(P)$. We would like to have the Markov chain P that minimizes $V[\widehat{\alpha}_n(P)]$. We here constrain the range of minimization of $V[\widehat{\alpha}_n(P)]$ to the same type of irreducible Markov chains P as the underlying Markov chain P_0.

The unconstrained minimization of the estimator variance is easily obtained by elementary calculus, however, the resultant simulation distribution is not Markov, nor stationary. Moreover, obtaining the unconstrained optimal simulation distribution requires the knowledge of the target probability itself [2].

For a Markov chain P and $s \in \boldsymbol{R}$, denote by ζ_s the largest eigen value of the matrix

$$\left(\frac{P_0(x'|x)^2}{P(x'|x)}e^{sf(x,x')}\right)_{x,x'\in\Omega}. \tag{11}$$

Define

$$J_P(y) = \sup_{-\infty<s<\infty}(ys - \log\zeta_s),\ y\in \boldsymbol{R}. \tag{12}$$

The following theorem holds from large deviations theory.

Theorem 2: (See [2]) For the variance $V[\hat{\alpha}_n(P)]$ of the IS estimate $\hat{\alpha}_n(P)$, we have

$$\lim_{n\to\infty} -\frac{1}{n}\log kV[\hat{\alpha}_n(P)] = J_P(c), \tag{13}$$

where c is the value in (6). □

The *optimal simulation distribution* P is the Markov chain that minimizes the asymptotic of the estimator variance, more precisely, that maximizes $J_P(c)$.

Theorem 3: (See [2]) We have

$$J_P(c) \leqq 2I(c). \tag{14}$$

The equality holds if and only if $P = P_{t^*}$ that is the TMC with t^* attaining the suprimum of $\sup_{-\infty<t<\infty}(ct - \log\lambda_t)$. □

In summary, we have the optimal simulation distribution as the TMC with t^* determined by Theorem 3. We can see that t^* satisfies

$$\left.\frac{d\lambda_t}{dt}\right|_{t=t^*} = c\lambda_t^*, \tag{15}$$

where the differentiability of λ_t is shown by Lemma 3 in [10].

5. Characterization of the Optimal Simulation Distribution by Kullback-Leibler Divergence of Markov Chains

When applying the IS technique to simulation problems, it is the most important to determine the optimal simulation distribution. But in Markov chain case, it is very difficult to have the optimal P_{t^*} by Theorem 3 because of the computation of the eigen value and the eigen vector and the optimization of the eigen value. Less complex algorithm is necessary to obtain P_{t^*}. The optimal simulation distribution of the IS for i.i.d. case is easily obtained [2]. The explicit form of the optimal simulation distribution for Gaussian, Laplacian, and Bernoulli cases are given in [2]. The relation between a twisted distribution and the Kullback-Leibler (KL) divergence in i.i.d. case is discussed in [4].

We will provide, in this chapter, a characterization of the optimal simulation distribution P_{t^*} by the KL divergence of Markov chains. Further, the geometric aspect of the TMC will be discussed in next chapter.

First, we have

Lemma 1: Let $\widetilde{w}_t = (\widetilde{w}_t(x))_{x\in\Omega}$ denote a left eigen vector of the matrix (7) associated with λ_t and recall that $w_t = (w_t(x))_{x\in\Omega}$ is a right eigen vector. The $\widetilde{w}_t$ is assumed to be normalized as $\sum_{x\in\Omega}\widetilde{w}_t(x)w_t(x) = 1$. Consider the TMC P_t defined by (9). The stationary distribution $p_t = (p_t(x))_{x\in\Omega}$ of P_t is given by $p_t(x) = \widetilde{w}_t(x)w_t(x),\ x\in\Omega$. □

Proof. By the definition of w_t and $\widetilde{w}_t$, we have

$$\sum_{x\in\Omega} P_t(x'|x)\widetilde{w}_t(x)w_t(x) \tag{16}$$

$$= \sum_{x\in\Omega} P_0(x'|x)e^{tf(x,x')}\lambda_t^{-1}\widetilde{w}_t(x)w_t(x') \tag{17}$$

$$= \widetilde{w}_t(x')w_t(x'). \tag{18}$$

So, $\widetilde{w}_t(x)w_t(x)$ gives the stationary distribution of P_t. □

Lemma 2: For the optimal simulation distribution P_{t^*}, we have $E_{P_{t^*}}[f] = c$. □

Proof. Since w_t is a right eigen vector associated with λ_t, we have

$$\sum_{x'\in\Omega} P_0(x'|x)e^{tf(x,x')}w_t(x') = \lambda_t w_t(x),\ x\in\Omega. \tag{19}$$

By differentiating the both sides of (19) with respect to t, we have

$$\sum_{x'\in\Omega} P_0(x'|x)\{f(x,x')e^{tf(x,x')}w_t(x') + e^{tf(x,x')}\frac{dw_t(x')}{dt}\} = \frac{d\lambda_t}{dt}w_t(x) + \lambda_t\frac{dw_t(x)}{dt},\ x\in\Omega. \tag{20}$$

Multiplying the both sides of (20) by $\widetilde{w}_t(x)$ and taking the summation with respect to $x\in\Omega$,

$$\sum_{x,x'\in\Omega} P_t(x'|x)\lambda_t w_t(x)f(x,x')\widetilde{w}_t(x) + \lambda_t\sum_{x'\in\Omega}\frac{dw_t(x')}{dt}\widetilde{w}_t(x') = \frac{d\lambda_t}{dt}\sum_{x\in\Omega} w_t(x)\widetilde{w}_t(x) + \lambda_t\sum_{x\in\Omega}\frac{dw_t(x)}{dt}\widetilde{w}_t(x). \tag{21}$$

Hence from Lemma 1, we have

$$\lambda_t\sum_{x,x'\in\Omega} P_t(x,x')f(x,x') = \frac{d\lambda_t}{dt}. \tag{22}$$

Substituting $t = t^*$ into (22), we see that

$$E_{P_t^*}[f] = \sum_{x,x'\in\Omega} P_{t^*}(x,x')f(x,x') = c \tag{23}$$

holds from (15). □

For two Markov chains $P = (P(x'|x))_{x,x' \in \Omega}$ and $Q = (Q(x'|x))_{x,x' \in \Omega}$, we will define the KL divergence $D(P\|Q)$. Let $p = (p(x))_{x \in \Omega}$ denote the stationary distribution of P and $P(x,x') \equiv p(x)P(x'|x)$ denote the joint probability of $x, x' \in \Omega$. Note that

$$\sum_{x' \in \Omega} P(x,x') = \sum_{x' \in \Omega} P(x',x) = p(x), \ x \in \Omega. \quad (24)$$

If $P(x'|x) = 0$ holds for any $x, x' \in \Omega$ with $Q(x'|x) = 0$, we write $P \prec Q$ and say that P is dominated by Q. The *Kullback-Leibler (KL) divergence* $D(P\|Q)$ is defined by

$$D(P\|Q) = \begin{cases} \displaystyle\sum_{x,x' \in \Omega} P(x,x') \log \frac{P(x'|x)}{Q(x'|x)}, & P \prec Q, \\ \infty, \quad \text{otherwise.} \end{cases}$$

The KL divergence is a kind of distance measure introduced on the space of Markov chains which plays a fundamental role in stochastic problems concerning Markov chains. For example, the KL divergence determines the power exponent of the most powerful test functions of hypothesis testing for Markov chains [10], [11], or the error exponent of the source coding of Markov chains [5], [11], [14], etc.

The following is a fundamental property of the KL divergence.

Lemma 3: (See [10]) For any Markov chains P and Q, $D(P\|Q) \geqq 0$ holds with equality if and only if $P = Q$. □

We then have a characterization of the optimal simulation distribution by the KL divergence.

Theorem 4: The optimal simulation distribution P_{t^*} of the IS simulation for $\alpha_n = P_0(A_n)$ is the unique Markov chain P that minimizes $D(P\|P_0)$ subject to the constraint $E_P[f] = c$. □

Proof. For any P with $E_P[f] = c$, we see from Lemma 3 and (24),

$$\begin{aligned} D(P\|P_0) &= \sum_{x,x' \in \Omega} P(x,x') \log \frac{P(x'|x)}{P_{t^*}(x'|x)} \frac{P_{t^*}(x'|x)}{P_0(x'|x)} \quad (25) \\ &= D(P\|P_{t^*}) + \sum_{x,x' \in \Omega} P(x,x') \{ t^* f(x,x') \\ &\quad + \log w_{t^*}(x') - \log w_{t^*}(x) - \log \lambda_{t^*} \} \quad (26) \\ &= D(P\|P_{t^*}) + ct^* - \log \lambda_{t^*} \quad (27) \\ &\geqq ct^* - \log \lambda_{t^*}. \quad (28) \end{aligned}$$

From Lemma 2, P_{t^*} satisfies $E_{P_{t^*}}[f] = c$. So, the equality in (28) holds for $P = P_{t^*}$. The uniqueness of P_{t^*} is guaranteed by Lemma 3. □

From Theorem 4, we see that P_{t^*} is the *dominating point* [1], [2] of the set $\{P \in \Delta | E_P[f] = c\}$ with respect to P_0.

Concerning the uniqueness of P_{t^*}, we can show the uniqueness of t that satisfies (15) or $\dfrac{d}{dt} \log \lambda_t = c$ as follows;

Proposition 1: The function $\log \lambda_t$ is strictly convex for $-\infty < t < \infty$. □

Proof. See Lemma 3 in [10]. □

The characterization of the optimal simulation distribution P_{t^*} by Theorem 4 is more convenient than that by Theorem 3 from the view point of computation. The variables of the maximization problem $\max_P J_P(c)$ in Theorem 3 are $P(x'|x)$, on the other hand, the variables of the minimization problem $\min_P D(P\|P_0)$ in Theorem 4 are $P(x,x')$. To give $(P(x'|x))_{x,x' \in \Omega}$ and to give $(P(x,x'))_{x,x' \in \Omega}$ are equivalent, however, the computation based on $P(x,x')$ is more advantageous than $P(x'|x)$. In fact, to compute $P(x,x')$ from $P(x'|x)$ is difficult but to compute $P(x'|x)$ from $P(x,x')$ is easy by using the Eq. (24). Hence, we can say that $P(x,x')$ is a more natural coordinate system than $P(x'|x)$ when we consider stochastic problems concerning Markov chains. We will investigate in detail in next chapter the coordinate system of the space of Markov chains.

6. Geometric Properties of the TMC

We will consider the set $\Delta = \{P\}$ of irreducible Markov chains on $\Omega = \{0, 1, \cdots, K\}$ and study the geometric structure of Δ. The transition probability matrix of a Markov chain $P = (P(x'|x))_{x,x' \in \Omega}$ consists of $(K+1)^2$ elements with $K+1$ constraints $\sum_{x' \in \Omega} P(x'|x) = 1$, $x \in \Omega$. Since $(K+1)^2 - (K+1) = K(K+1)$ elements specify a Markov chain, Δ forms a $K(K+1)$ dimensional manifold. On the manifold Δ, two coordinate systems η and θ are introduced [1], [9], namely,

$$\eta(x,x') = p(x)P(x'|x) \equiv P(x,x'), \quad (29)$$

$$\theta(x,x') = \log \frac{P(x'|x)P(0|x')}{P(0|x)P(0|0)}, \quad x, x' \in \Omega, \ x' \neq 0, \quad (30)$$

where $p = (p(x))_{x \in \Omega}$ is the stationary distribution of P and $P(x,x')$ is the joint probability of x, $x' \in \Omega$. The η- and θ- coordinate systems are dual to each other with respect to the Riemannian metric determined by the Fisher information matrix of Markov chains. The η is called the -1 affine coordinate system and θ the $+1$ affine coordinate system. An affine straight line with respect to the η or θ coordinate system is called a -1 or $+1$ geodesic, respectively. A set of the form $\{P \in \Delta | \sum_{x,x' \in \Omega} P(x,x')a(x,x') = b, \ a(x,x') \in \boldsymbol{R}, b \in \boldsymbol{R}\}$ is called a *hyperplane* in Δ.

Let us consider a smooth curve $P_t \in \Delta$, $t \in \boldsymbol{R}$ with -1 affine coordinate η_t and a smooth curve $\widetilde{P}_t \in \Delta$, $t \in \boldsymbol{R}$ with $+1$ affine coordinate $\widetilde{\theta}_t$. The P_t and $\widetilde{P}_t$ are assumed to intersect at P^*, i.e., $P_{t_1} = \widetilde{P}_{t_2} = P^*$ for some $t_1, t_2 \in \boldsymbol{R}$. The two curves P_t and $\widetilde{P}_t$ are said to be *orthogonal* at $P_{t_1} = \widetilde{P}_{t_2} = P^*$ if

$$\sum_{x,x' \in \Omega, x' \neq 0} \left. \frac{d\eta_t(x,x')}{dt} \right|_{t=t_1} \left. \frac{d\widetilde{\theta}_t(x,x')}{dt} \right|_{t=t_2} = 0. \quad (31)$$

Let H be a hyperplane in Δ and $\widetilde{P}_t$ be a $+1$ geodesic with $+1$ affine coordinate $\widetilde{\theta}_t$, $t \in \boldsymbol{R}$. H and $\widetilde{P}_t$ are assumed to intersect at a $P^* = \widetilde{P}_{t_2}$, $t_2 \in \boldsymbol{R}$ whose -1 affine coordinate is η^*. H and $\widetilde{P}_t$ are said to be orthogonal if for any $P \in H$ with -1 affine coordinate η, the following holds;

$$\sum_{x,x' \in \Omega, x' \neq 0} \{\eta(x,x') - \eta^*(x,x')\} \left. \frac{d\widetilde{\theta}(x,x')}{dt} \right|_{t=t_2} = 0. \tag{32}$$

For reference of the geometric structure of the set Δ of Markov chains, see [1], [9].

Now, let P_0 and P_1 be irreducible Markov chains whose $+1$ affine coordinates are θ_0 and θ_1, respectively. Let P_t be the $+1$ geodesic passing P_0 and P_1. The $+1$ affine coordinate θ_t of P_t is given by

$$\theta_t(x,x') = \theta_0(x,x') + t(\theta_1(x,x') - \theta_0(x,x')), \quad x,x' \in \Omega, x' \neq 0, \ t \in \boldsymbol{R}. \tag{33}$$

Note that particularly P_t for $t = 0$ and $t = 1$ coincide with the given P_0 and P_1, respectively. The parameter t is called the *affine parameter* of the geodesic P_t.

In the previous sections and this section, we have used the same notation P_t for both the $+1$ geodesic and the TMC (9). We will show that they are indeed the same object. We have

Theorem 5: The TMC P_t defined by (9) is the $+1$ geodesic passing P_0 whose affine parameter is t. The optimal simulation distribution P_{t^*} is the intersection of the geodesic P_t and the hyperplane $H \equiv \{P \in \Delta | E_P[f] = c\}$. Furthermore, at the P_{t^*}, the geodesic P_t and the hyperplane H are orthogonal. □

Proof. Substituting $t = 1$ into (9), we have

$$P_1(x'|x) = P_0(x'|x) e^{f(x,x')} \frac{w_1(x')}{\lambda_1 w_1(x)}, \quad x,x' \in \Omega. \tag{34}$$

We will show that P_t is the $+1$ geodesic passing P_0 and P_1. Let θ_t, θ_0, and θ_1 denote the $+1$ affine coordinates of the Markov chains P_t, P_0, and P_1, respectively. Then from (30), (9), and (34), we have

$$\theta_t(x,x') = \log \frac{P_t(x'|x) P_t(0|x')}{P_t(0|x) P_t(0|0)} \tag{35}$$

$$= \log \frac{P_0(x'|x) e^{tf(x,x')} P_0(0|x') e^{tf(x',0)}}{P_0(0|x) e^{tf(x,0)} P_0(0|0) e^{tf(0,0)}} \tag{36}$$

$$= \theta_0(x,x') + t(\theta_1(x,x') - \theta_0(x,x')), \quad x,x' \in \Omega, x' \neq 0. \tag{37}$$

Hence, from (33), the TMC (9) is the $+1$ geodesic. From Lemma 2, P_{t^*} is lying on the hyperplane H as well as on the $+1$ geodesic P_t, in other words, P_{t^*} is the intersection of P_t and H. We next show the orthogonality of P_t and H at P_{t^*}. From (37), we have

$$\frac{d\theta_t(x,x')}{dt} = \theta_1(x,x') - \theta_0(x,x'), \quad x,x' \in \Omega, x' \neq 0. \tag{38}$$

Let P be an arbitrary Markov chain lying on the hyperplane H. Let η_{t^*} and η denote the -1 affine coordinate of P_{t^*} and P, respectively. Then from (38) and (30), we have by tedious but straightforward computation

$$\sum_{x,x' \in \Omega, x \neq 0} \{\eta(x,x') - \eta_{t^*}(x,x')\} \frac{d\theta(x,x')}{dt} \tag{39}$$

$$= \sum_{x,x' \in \Omega, x \neq 0} \{P(x,x') - P_{t^*}(x,x')\} \times \{\theta_1(x,x') - \theta_0(x,x')\} \tag{40}$$

$$= 0. \tag{41}$$

This completes the proof of Theorem 5. □

7. Example

Let $\Omega = \{0,1\}$, $f(0,x') = 1/2$, $f(1,x') = -1/2$, $x' \in \Omega$ and consider a Markov chain P_0 with the following transition probability matrix [2], p.142;

$$\begin{pmatrix} p & q \\ q & p \end{pmatrix}, \quad 0 < p < q, \ p + q = 1. \tag{42}$$

In [2], the optimal simulation distribution P_{t^*} is obtained by numerical computation because the explicit form of P_{t^*} represented by p and q is very complicated.

On the other hand, the computation of P_{t^*} based on Theorem 4 is simple. The optimal P_{t^*} is given by

$$P_{t^*} = \begin{pmatrix} \dfrac{2z}{1+2c} & \dfrac{1+2c-2z}{1+2c} \\ \dfrac{1+2c-2z}{1-2c} & \dfrac{-4c+2z}{1-2c} \end{pmatrix}, \tag{43}$$

where z is the positive solution of the quadratic equation $(2q-4p^2)z^2 + (4p^2(1+2c) - 4qc)z - p^2(1+2c)^2 = 0$.

Acknowledgment

This work was supported by a research grant from NTT (Nippon Telegraph and Telephone Corporation).

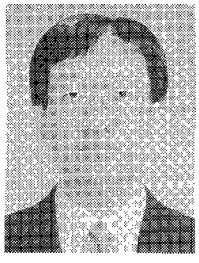
References

[1] S. Amari, "Differential-Geometrical Methods in Statistics," (Lecture Notes in Statistics), Springer-Verlarg, New York, 1985.

[2] J.A. Bucklew, "Large Deviation Techniques in Decision, Simulation, and Estimation," Wiley, New York, 1990.

[3] J.A. Bucklew, P. Ney, and J.S. Sadowsky, "Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains," J. Appl. Prob. 27, pp.44–59, 1990.

[4] M. Cottrell, J-C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," IEEE Trans. Autom. Control, vol.AC-28, no.9, pp.907–920, Sept. 1983.
[5] L.D. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," IEEE Trans. Inf. Theory, vol.27, no.4, pp.431–438, July 1981.
[6] A. Dembo and O. Zeitouni, "Large Deviations Techniques and Application," Jones and Bartlett, 1993.
[7] M. Devetsikiotis and J.K. Townsend, "Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks," IEEE/ACM Trans. Networking, vol.1, no.3, pp.293–305, June 1993.
[8] P.M. Hahn and M.C. Jeruchim, "Developments in the theory and application of importance sampling," IEEE Trans. Commun., vol.35, no.7, pp.706–714, July 1987.
[9] H. Itoh and S. Amari, "Geometry of information sources," Proc. 11th SITA, pp.57–60, 1988.
[10] K. Nakagawa and F. Kanaya, "On the converse theorem in statistical hypothesis testing for Markov chains," IEEE Trans. Inf. Theory, vol.39, no.2, pp.629–633, March 1993.
[11] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains," IEEE Trans. Inf. Theory, vol.31, no.3, pp.360–365, May 1985.
[12] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in the networks of queues," IEEE Trans. Autom. Control, vol.34, no.1, pp.54–66, Jan. 1989.
[13] J.S. Sadowsky, "Large deviations theory and efficient simulation of excessive backlogs in $GI/GI/m$ queues," IEEE Trans. Autom. Control, vol.36, no.12, pp.1383–1394, Dec. 1991.
[14] K. Vašek, "On the error exponent for ergodic Markov chains," Kybernetika, vol.16, no.4, pp.318–329, 1980.
[15] Q. Wan and V.S. Frost, "Efficient estimation of cell blocking probability for ATM systems," IEEE/ACM Trans. Networking, vol.1, no.2, pp.230–235, April 1993.



Kenji Nakagawa received the B.S., M.S. and D.S. degrees from Tokyo Institute of Technology in 1980, 1982, and 1985, respectively. In 1985, he joined NTT (Nippon Telegraph and Telephone Corp.). Since 1992, he has been an associate professor of Dept. of Electrical Engineering, Nagaoka University of Technology. His research interests include information theory, queueing theory, and statistics. Dr. Nakagawa is a member of the IEEE, the SITA, and the Mathematical Society of Japan.