[21] L. Gurvits, "Stability of discrete linear inclusion," *Linear Algebra Applications*, vol. 231, pp. 47–85, 1995.
[22] N. E. Barabanov, "Lyapunov indicators of discrete inclusions. I.," *Automat. Remote Contr.*, vol. 49, pp. 152–157, Feb. 1988.
[23] J. N. Tsitsiklis and V. D. Blondel, "The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate," *Math. Control, Signals, Syst.*, vol. 10, pp. 31–40, 1997.
[24] V. D. Blondel and J. N. Tsitsiklis, "Boundedness of finitely generated matrix semigroups is undecidable," preprint, Jan. 1999.
[25] M. Maesumi, "An efficient lower bound for the generalized spectral radius of a set of matrices," *Linear Algebra Applications*, vol. 240, pp. 1–7, 1996.
[26] ——, "Spectral radius of sets of matrices," in *Proc. SPIE*, vol. 2303, San Diego, CA, July 1994, pp. 27–29.
[27] D. Colella and C. Heil, "The characterization of continuous, four-coefficient scaling functions and wavelets," *IEEE Trans. Inform. Theory*, vol. 38, pp. 876–881, Mar. 1992.
[28] G. Gripenberg, "Computing the joint spectral radius," *Linear Algebra Applications*, vol. 234, pp. 43–60, 1996.
[29] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*. New York: Cambridge Univ. Press, 1995.
[30] R. Karabed and P. H. Siegel, "Matched spectral-null codes for partial-response channels," *IEEE Trans. Inform. Theory*, vol. 37, pp. 818–855, May 1991.

On the Optimal Markov Chain of IS Simulation

Kenji Nakagawa, *Member, IEEE*

***Abstract*—We investigate the importance sampling (IS) simulation for the sample average of an output sequence from an irreducible Markov chain. The optimal Markov chain used in simulation is known to be a twisted Markov chain, however, the proofs in [2], [3] are very complicated and do not give us a good perspective. We give a simple and natural proof for the optimality of the simulation Markov chain in terms of the Kullback–Leibler (KL) divergence of Markov chains. The performance degradation of the IS simulation by using a not optimal simulation Markov chain, i.e., the difference between the obtained variance and the minimum variance is shown to be represented by the KL divergence. Moreover, we show a geometric relationship between a simulation Markov chain and the optimal one.**

***Index Terms*—Importance sampling simulation, information geometry, Kullback–Leibler (KL) divergence, Markov chain.**

I. Introduction

The importance sampling (IS) simulation technique has been used to obtain quickly an accurate estimate for a very small probability that is not tractable by the ordinary Monte Carlo (MC) simulation. The IS technique is widely used for various types of engineering problems, e.g., the estimation of a blocking probability in queuing system [2], [4], [7], an error rate in communications system [2], [6], etc. See [2] for an overview of the application of IS simulation.

If the target event (blocking in queuing systems, or error in communications systems) is a rare event with small probability of less than about 10^{-6}, it is impossible to obtain an estimate by the ordinary MC simulation because of the limit of simulation time and the precision limit of pseudo-random numbers. To overcome these difficulties, a different probability distribution from the underlying probability distribution is used for simulation in order to generate more samples in the target event. Then the obtained value is modified by the likelihood ratio to obtain an unbiased estimate. This estimate is called an IS estimate. The probability distribution used for simulation 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 MC estimate. The simulation distribution that yields the IS estimate of the minimum variance is referred to as the optimal simulation distribution. When we apply the IS technique to some simulation problem, it is critical to find the optimal simulation distribution.

In the case of a Markov chain, the Markov chain which is used in the IS simulation is called a simulation Markov chain, and the optimal one is called the optimal simulation Markov chain. It has been known [2], [3] that the optimal simulation Markov chain is unique and belongs to the class of twisted Markov chains (TMC), but the proofs in [2], [3] are complicated and do not give a good perspective. In [2], the perturbation technique is used to prove that the optimal simulation Markov chain is a TMC. In [3], Jensen's inequality is used in the proof and the line of argument is elementary but complicated. Both of the proofs in [2] and [3] are lengthy.

In [10], we studied geometric properties of IS simulation and showed that the Kullback–Leibler (KL) divergence of Markov chains plays an important role in this problem. In this correspondence, we give a geometric view to this problem and provide a simple and natural proof for the optimality of a simulation Markov chain in terms of KL divergence. The performance degradation of the IS simulation by using a not optimal simulation Markov chain, i.e., the difference between the obtained variance and the minimum variance is shown to be represented by the KL divergence. Moreover, we show a geometric relationship between a simulation Markov chain and the optimal one.

II. Importance Sampling Simulation for Markov Chains

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

Let P_0 denote an irreducible Markov chain on the state space $\Omega \equiv \{0, 1, \ldots, K\}$, $K > 0$. 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 $f: \Omega \times \Omega \to Z$, where Z denotes the set of integers. Denote by $E_{P_0}[f]$ the expectation of f with respect to P_0

$$E_{P_0}[f] \equiv \sum_{x, x' \in \Omega} P_0(x, x')f(x, x'). \tag{1}$$

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

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

$$A_n = \left\{ \boldsymbol{x}^n \in \Omega^n \,\middle|\, \frac{1}{n-1} \sum_{i=1}^{n-1} f(x_i, x_{i+1}) > c \right\}, \qquad c > 0. \tag{2}$$

Write

$$\alpha_n \equiv P_0(A_n) = \sum_{\boldsymbol{x}^n \in A^n} P_0(\boldsymbol{x}^n)$$

where

$$P_0(\boldsymbol{x}^n) = p_0(x_1)P_0(x_2|x_1) \ldots P_0(x_n|x_{n-1}).$$

Manuscript received December 28, 1998.

The author is with Nagaoka University of Technology, Nagaoka, Niigata 940-2188 Japan (e-mail: nakagawa@nagaokaut.ac.jp).

Communicated by R. Cruz, Associate Editor for Communication Networks.
Publisher Item Identifier S 0018-9448(01)00580-6.

We see from large deviations theory [2], [6], that the asymptotic behavior of the α_n is given as follows. Let R denote the set of real numbers. For $t \in R$, denote by λ_t the largest eigenvalue of a matrix

$$\left(P_0(x'|x)e^{tf(x,x')}\right)_{x,x'\in\Omega}. \tag{3}$$

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

$$I(y) = \sup_{t\in R}(yt - \log \lambda_t). \tag{4}$$

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

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

$$\lim_{n\to\infty} -\frac{1}{n}\log \alpha_n = I(c) \tag{5}$$

where c is the value in (2). □

We know the asymptotic behavior of α_n by this theorem, however, it is often necessary to have an accurate value of α_n for some finite n. So, we need to use a stochastic simulation, but if α_n is very small, it is not tractable by the ordinary MC simulation. We apply the IS technique to the simulation of α_n.

Consider the IS simulation with a simulation Markov chain Q. Let $\widetilde{X}^n = (\widetilde{X}_1, \ldots, \widetilde{X}_n)$, $n = 1, 2, \ldots$, be a random sequence generated by Q, and $\widetilde{\boldsymbol{x}}_1^n, \ldots, \widetilde{\boldsymbol{x}}_k^n \in \Omega^n$ be k independent realizations of $\widetilde{X}^n$. The IS estimate $\widehat{\alpha}_n(Q)$ for α_n with a simulation Markov chain Q is given by

$$\widehat{\alpha}_n(Q) = \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}_{A_n}(\widetilde{\boldsymbol{x}}_j^n)\,\frac{P_0(\widetilde{\boldsymbol{x}}_j^n)}{Q(\widetilde{\boldsymbol{x}}_j^n)} \tag{6}$$

where $\mathbf{1}_{A_n}$ is the indicator function of the set A_n.

III. The Optimal Simulation Markov Chain

It is readily seen that the IS estimate (6) is unbiased [2]. Among the IS estimates, we would like to have the Markov chain Q that minimizes the variance $V[\widehat{\alpha}_n(Q)]$ of $\widehat{\alpha}_n(Q)$. A simulation Markov chain Q is assumed to be of the same type of irreducible Markov chain as the underlying Markov chain P_0.

A. Twisted Markov Chain and Q-Twisted Markov Chain

For the underlying Markov chain P_0, we define a TMC P_t, $t \in R$. Recalling that λ_t is the largest eigenvalue of the matrix (3), let us denote by $w_t = (w_t(x))_{x\in\Omega}$ a right eigenvector associated with λ_t. Then, a TMC P_t is defined by

$$P_t(x'|x) = P_0(x'|x)e^{tf(x,x')}\,\frac{w_t(x')}{\lambda_t w_t(x)}, \qquad x, x' \in \Omega,\ t \in R. \tag{7}$$

From the definition of λ_t and w_t, one can readily see that

$$\sum_{x'\in\Omega} P_t(x'|x) = 1, \quad x \in \Omega$$

holds.

Next, we define another type of TMC which plays an important role when we evaluate the variance of an IS estimate. Consider the IS simulation with simulation Markov chain Q. For $s \in R$, denote by $\zeta_{Q,s}$ the largest eigenvalue of a matrix

$$\left(\frac{P_0(x'|x)^2}{Q(x'|x)}\,e^{sf(x,x')}\right)_{x,x'\in\Omega} \tag{8}$$

and $v_{Q,s} = (v_{Q,s}(x))_{x\in\Omega}$ its associated right eigenvector. A TMC $P_{Q,s}$ concerned with the simulation Markov chain Q, which we call a Q-TMC for short, is defined by

$$P_{Q,s}(x'|x) = \frac{P_0(x'|x)^2}{Q(x'|x)}\,e^{sf(x,x')}\,\frac{v_{Q,s}(x')}{\zeta_{Q,s}v_{Q,s}(x)}, \qquad x, x' \in \Omega. \tag{9}$$

We can see

$$\sum_{x'\in\Omega} P_{Q,s}(x'|x) = 1,\ x \in \Omega.$$

We next define

$$J_Q(y) = \sup_{s\in R}(ys - \log \zeta_{Q,s}), \qquad y \in R. \tag{10}$$

Then the following theorem holds.

Theorem 2 (see [2]): For the variance $V[\widehat{\alpha}_n(Q)]$ of $\widehat{\alpha}_n(Q)$ by a simulation Markov chain Q, we have

$$\lim_{n\to\infty} -\frac{1}{n}\log V[\widehat{\alpha}_n(Q)] = J_Q(c) \tag{11}$$

where c is the value in (2). □

The *optimal* simulation Markov chain is defined to be the Markov chain Q that maximizes the $J_Q(c)$.

B. Lemmas for TMC and Q-TMC

For preparation, we give some lemmas.

Let $\widetilde{w}_t = (\widetilde{w}_t(x))_{x\in\Omega}$ denote a left eigenvector of the matrix (3) associated with λ_t, and recall that $w_t = (w_t(x))_{x\in\Omega}$ is a right eigenvector. The $\widetilde{w}_t$ is assumed to be normalized as

$$\sum_{x\in\Omega} w_t(x)\widetilde{w}_t(x) = 1.$$

Similarly, let $\widetilde{v}_{Q,s} = (\widetilde{v}_{Q,s}(x))_{x\in\Omega}$ denote a left eigenvector of the matrix (8) associated with $\zeta_{Q,s}$ which is normalized as

$$\sum_{x\in\Omega} v_{Q,s}(x)\widetilde{v}_{Q,s}(x) = 1.$$

Lemma 1: The stationary distribution $p_t = (p_t(x))_{x\in\Omega}$ of a TMC P_t is given by

$$p_t(x) = w_t(x)\widetilde{w}_t(x),\ x \in \Omega.$$

Furthermore, the stationary distribution $p_{Q,s} = (p_{Q,s}(x))_{x\in\Omega}$ of a Q-TMC $P_{Q,s}$ is given by

$$p_{Q,s}(x) = v_{Q,s}(x)\widetilde{v}_{Q,s}(x),\ x \in \Omega.$$ □

Proof: For P_t, see [10, Lemma 1]. Here we give a proof for $P_{Q,s}$. By the definition of $v_{Q,s}$ and $\widetilde{v}_{Q,s}$, we have

$$\sum_{x\in\Omega} P_{Q,s}(x'|x)v_{Q,s}(x)\widetilde{v}_{Q,s}(x) \tag{12}$$

$$= \sum_{x\in\Omega} \frac{P_0(x'|x)^2}{Q(x'|x)}e^{sf(x,x')}\zeta_{Q,s}^{-1}v_{Q,s}(x')\widetilde{v}_{Q,s}(x) \tag{13}$$

$$= v_{Q,s}(x')\widetilde{v}_{Q,s}(x'), \qquad x' \in \Omega. \tag{14}$$

Lemma 2: For $t = t^*$ that attains

$$\sup_{t\in R}(ct - \log \lambda_t)$$

we have $E_{P_{t^*}}[f] = c$. Furthermore, for $s = s^*(Q)$ that attains

$$\sup_{s\in R}(cs - \log \zeta_{Q,s})$$

we have $E_{P_{Q,s^*(Q)}}[f] = c$. □

Proof: For P_t, see [10, Lemma 2]. Here we give a proof for $P_{Q,s}$. First, we have

$$\frac{d}{ds}(cs - \log \zeta_{Q,s})\Big|_{s=s^*(Q)} = c - \zeta_{Q,s}^{-1}(Q) \frac{d\zeta_{Q,s}}{ds}\Big|_{s=s^*(Q)} = 0. \tag{15}$$

Since $v_{Q,s}$ is a right eigenvector associated with $\zeta_{Q,s}$, we have

$$\sum_{x' \in \Omega} \frac{P_0(x'|x)^2}{Q(x'|x)} e^{sf(x,x')} v_{Q,s}(x') = \zeta_{Q,s} v_{Q,s}(x), \qquad x \in \Omega. \tag{16}$$

By differentiating the both sides of (16) with respect to s, multiplying by $\widetilde{v}_{Q,s}(x)$, and taking summation with respect to $x \in \Omega$, we have from Lemma 1

$$\zeta_{Q,s^*} \sum_{x,\, x' \in \Omega} P_{Q,s}(x,\, x') f(x,\, x') = \frac{d\zeta_{Q,s}}{ds}. \tag{17}$$

Substituting $s = s^*(Q)$ into (17), we see from (15) that $E_{P_{Q,s^*(Q)}}[f] = c$ holds. □

Lemma 3: For a TMC P_t, we have the following:

i) $\zeta_{P_t,s} = \lambda_t \lambda_{s-t}$,

ii) $v_{P_t,s}(x) = w_t(x) w_{s-t}(x)$, $x \in \Omega$,

iii) $P_{P_t,s} = P_{s-t}$. □

Proof: We can confirm by a straightforward computation that $\lambda_t \lambda_{s-t}$ is an eigenvalue of the matrix

$$\left(\frac{P_0(x'|x)^2}{P_t(x'|x)} e^{sf(x,x')} \right)_{x,\, x' \in \Omega} \tag{18}$$

and $w_t(x) w_{s-t}(x)$ is its associated left eigenvector. iii) is easily checked by i) and ii). □

For two Markov chains

$$P = (P(x'|x))_{x,\, x' \in \Omega}$$

and

$$Q = (Q(x'|x))_{x,\, x' \in \Omega}$$

we 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

$$\sum_{x' \in \Omega} P(x,\, x') = \sum_{x' \in \Omega} P(x',\, x) = p(x), \qquad x \in \Omega. \tag{19}$$

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 KL divergence $D(P\|Q)$ is defined by

$$D(P\|Q) = \begin{cases} \displaystyle\sum_{x,\, x' \in \Omega} P(x,\, x') \log \frac{P(x'|x)}{Q(x'|x)}, & P \prec Q \\ \infty, & \text{otherwise.} \end{cases} \tag{20}$$

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 function of hypothesis testing for Markov chains [9], [11], or the error exponent of the source coding of Markov chains [5], [11], etc.

The following is a fundamental property of the KL divergence.

Lemma 4 (see [9]): For any Markov chains P and Q, we have $D(P\|Q) \geq 0$. The equality holds if and only if $P = Q$. □

Next, we characterize the rate function $I(c)$ by the KL divergence.

Lemma 5 (see [10, Theorem 4]): For $t = t^*$ that attains

$$\sup_{t \in R} (ct - \log \lambda_t)$$

we have

$$I(c) = D(P_{t^*}\|P_0) = \inf_{P:\, E_P[f] = c} D(P\|P_0) \tag{21}$$

i.e., P_{t^*} is the dominating point [2] of the set $\{P|E_P[f] = c\}$ with respect to P_0. □

Proof: For any P with $E_P[f] = c$, we see from (7) and Lemma 4

$$D(P\|P_0) = \sum_{x,\, x' \in \Omega} P(x,\, x') \log \frac{P(x'|x)}{P_{t^*}(x'|x)} \frac{P_{t^*}(x'|x)}{P_0(x'|x)} \tag{22}$$

$$= D(P\|P_{t^*}) + ct^* - \log \lambda_{t^*} \tag{23}$$

$$\geq I(c). \tag{24}$$

The equality holds if and only if $P = P_{t^*}$. □

We can now give a new proof for the following theorem with the KL divergence.

Theorem 3: For an arbitrary simulation Markov chain Q, we have

$$J_Q(c) \leq 2I(c). \tag{25}$$

The equality holds if and only if $Q = P_{t^*}$. □

Proof: From (9), we have

$$\log \frac{P_{Q,s}(x'|x)}{P_0(x'|x)} = \log \frac{P_0(x'|x)}{Q(x'|x)} + sf(x,\, x') + \log v_{Q,s}(x') - \log v_{Q,s}(x) - \log \zeta_{Q,s}. \tag{26}$$

Multiplying both sides of (26) by $P_{t^*}(x,\, x')$ and taking summation with $x,\, x' \in \Omega$, we have, from Lemma 2 and (19)

$$\sum_{x,\, x' \in \Omega} P_{t^*}(x,\, x') \log \frac{P_{Q,s}(x'|x)}{P_0(x'|x)} = \sum_{x,\, x' \in \Omega} P_{t^*}(x,\, x') \log \frac{P_0(x'|x)}{Q(x'|x)} + sc - \log \zeta_{Q,s}. \tag{27}$$

Hence, from (27) and Lemma 5

$$sc - \log \zeta_{Q,s} = 2D(P_{t^*}\|P_0) - D(P_{t^*}\|Q) - D(P_{t^*}\|P_{Q,s}) = 2I(c) - D(P_{t^*}\|Q) - D(P_{t^*}\|P_{Q,s}). \tag{28}$$

From (28) and Lemma 4, we have

$$J_Q(c) \leq 2I(c). \tag{29}$$

The equality in (29) holds only if $Q = P_{t^*}$. If $Q = P_{t^*}$, from Lemma 3

$$\inf_{s \in R} D(P_{t^*}\|P_{Q,s}) = \inf_{s \in R} D(P_{t^*}\|P_{s-t^*}) \tag{30}$$

$$= 0 \qquad (\text{by putting } s = 2t^*). \tag{31}$$

This completes the proof. □

Corollary 1: The degradation of the variance by using a not necessarily optimal simulation Markov chain Q, i.e., the difference $2I(c) - J_Q(c)$ is represented by the KL divergence

$$2I(c) - J_Q(c) = D(P_{t^*}\|Q) + \inf_{s \in R} D(P_{t^*}\|P_{Q,s}). \tag{32}$$

IV. Geometric Relation of P_0 and Q

Let us consider the set Δ of irreducible Markov chains on the state space $\Omega = \{0, 1, \ldots, K\}$. On the set Δ, two mutually dual coordinate systems η and θ are introduced [1], [8], [10], namely

$$\eta(x, x') = p(x)P(x'|x) \equiv P(x, x'), \qquad x, x' \in \Omega, x' \neq 0 \quad (33)$$

$$\theta(x, x') = \log \frac{P(x'|x)P(0|x')}{P(0|x)P(0|0)}, \qquad x, x' \in \Omega, x' \neq 0 \quad (34)$$

where $p = (p(x))_{x \in \Omega}$ is the stationary distribution of P and $P(x, x')$ is the joint probability of $x, x' \in \Omega$.

Differential geometric structures are defined on Δ and statistical quantities are represented in terms of geometric quantities, see [1], [8] for a reference of geometry on Δ.

A straight line in the θ coordinate is called a *geodesic*, which is indeed a geodesic defined by an affine connection introduced on Δ [1]. A subset H of Δ of the form

$$H = \{P \in \Delta | \sum_{x, x' \in \Omega} P(x, x')a(x, x') = b, a(x, x') \in R, b \in R\}$$

is called a *hyperplane*. For a geodesic l and a hyperplane H, we say that l *intersects orthogonally to* H *at* P^* if $P^* \in l \cap H$ and

$$\sum_{x, x' \in \Omega, x' \neq 0} \frac{d\theta_{P(t)}(x, x')}{dt} (\eta_{Q_1}(x, x') - \eta_{Q_2}(x, x')) = 0 \quad (35)$$

hold for any $P(t) \in l$ with the θ coordinate $\theta_{P(t)}$ and any $Q_1, Q_2 \in H$ with η coordinates η_{Q_1}, η_{Q_2}.

Then we have the following lemma.

Lemma 6 (Pythagoras, [1], [8]): Let l be a geodesic that intersects orthogonally to a hyperplane H at P^*. Then for any $P \in l$ and $Q \in H$, we have

$$D(Q\|P) = D(Q\|P^*) + D(P^*\|Q). \quad (36)$$

We will show some geometric properties of the IS simulation. First, we notice that the set of P which satisfies $E_P[f] = c$ forms a hyperplane. Let us denote it by $H_f \equiv \{P \in \Delta | E_P[f] = c\}$. Then we have the following theorem.

Theorem 4: The TMC P_t is a geodesic and it intersects orthogonally to the hyperplane H_f at P_{t^*}. Furthermore, the Q-TMC $P_{Q,s}$ is a geodesic and it intersects orthogonally to the hyperplane H_f at $P_{Q, s^*(Q)}$. □

Proof: For P_t, see [10, Theorem 5]. Here we give a proof for $P_{Q,s}$. Let $\theta_{Q,s}$ denote the θ coordinate of $P_{Q,s}$. Then by the definition (34) of θ coordinate, it is easy to see that

$$\theta_{Q,s}(x, x') = \theta_{Q,0}(x, x') + s(\theta_{Q,1}(x, x') - \theta_{Q,0}(x, x')), \qquad x, x' \in \Omega, x' \neq 0. \quad (37)$$

From (37), we see that $\theta_{Q,s}$ is a straight line in the θ coordinate, hence $P_{Q,s}$ is a geodesic. For any $P_{Q,s}$ with θ coordinate $\theta_{Q,s}$ and for any $Q_1, Q_2 \in H_f$ with η coordinates η_{Q_1}, η_{Q_2}, we have from (33), (34), and (37) that

$$\sum_{x, x' \in \Omega, x' \neq 0} \frac{d\theta_{Q,s}(x, x')}{ds} (\eta_{Q_1}(x, x') - \eta_{Q_2}(x, x')) = 0. \quad (38)$$

Thus, from Lemma 2, we see that the geodesic $P_{Q,s}$ intersects orthogonally to the hyperplane H_f at $P_{Q, s^*(Q)}$. □

Corollary 2: The infimum in (32) is attained by $s = s^*(Q)$. Thus, we have

$$2I(c) - J_Q(c) = D(P_{t^*}\|Q) + D(P_{t^*}\|P_{Q, s^*(Q)}). \quad (39)$$

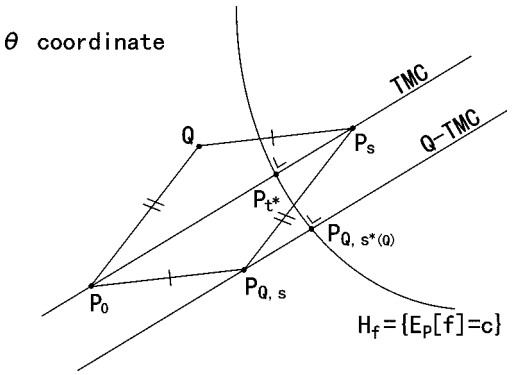

Fig. 1. Geometric relation of $P_0, Q, P_{t^*}, P_{Q,s}$ in the θ coordinate.

(See Lemma 2 for the definition of $s^*(Q)$.) □

Proof: By Lemmas 2, 6, and Theorem 4, for any $s \in R$ we have

$$D(P_{t^*}\|P_{Q,s}) = D(P_{t^*}\|P_{Q, s^*(Q)}) + D(P_{Q, s^*(Q)}\|P_{Q,s}) \quad (40)$$

$$\geq D(P_{t^*}\|P_{Q, s^*(Q)}). \quad (41)$$

The equality in (41) holds if and only if $s = s^*(Q)$. □

Theorem 5: For the underlying Markov chain P_0 and a simulation Markov chain Q and $s \in R$, denote by $\theta_0, \theta_s, \theta_Q, \theta_{Q,s}$ the θ coordinates of $P_0, P_s, Q, P_{Q,s}$, respectively. Then we have a relation

$$\theta_{Q,s} = \theta_0 + \theta_s - \theta_Q \quad (42)$$

which can be depicted in Fig. 1. □

Proof: (42) is easily checked by the definition (34) of the θ coordinate. □

From Fig. 1, we can see by geometric consideration that the optimal simulation Markov chain Q is $Q = P_{t^*}$ and $s = 2t^*$.

V. Conclusion

We investigated the IS simulation for the sample average of an output sequence from an irreducible Markov chain. We provided a new proof for the optimality of a simulation Markov chain by means of the KL divergence of Markov chains. Furthermore, we gave a geometric view to the relation of the underlying Markov chain and a simulation Markov chain.

In a more complicated simulation problem, usually the optimal simulation distribution is searched only from among the twisted distributions. This is because the set of twisted distributions is one-dimensional and the optimization is easy. The introduction of the geometric view to the IS simulation can yield a simple proof for the proposition that the optimal simulation distribution among a wider class is necessarily a twisted distribution.

References

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

[2] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley, 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. Probab.*, vol. 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. Automat. Contr.*, 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. Inform. Theory*, vol. IT-27, pp. 431–438, July 1981.

[6] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Application*. Boston, MA: 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, pp. 293–305, June 1993.
[8] H. Itoh and S. Amari, "Geometry of information sources," in *Proc. 11th SITA*, 1988, pp. 57–60.
[9] K. Nakagawa and F. Kanaya, "On the converse theorem in statistical hypothesis testing for Markov chains," *IEEE Trans. Inform. Theory*, vol. 39, pp. 629–633, Mar. 1993.
[10] K. Nakagawa, "On the twisted Markov chain of importance sampling simulation," *IEICE Trans. Fund.*, vol. E79-A, no. 9, pp. 1423–1428, Sept. 1996.
[11] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 360–365, May 1985.

Quasi-Convexity and Optimal Binary Fusion for Distributed Detection with Identical Sensors in Generalized Gaussian Noise

Wei Shi, *Member, IEEE*, Thomas W. Sun, *Student Member, IEEE*, and Richard D. Wesel, *Member, IEEE*

***Abstract*—In this correspondence, we present a technique to find the optimal threshold τ for the binary hypothesis detection problem with n identical and independent sensors. The sensors all use an identical and single threshold τ to make local decisions, and the fusion center makes a global decision based on the n local binary decisions. For generalized Gaussian noises and some non-Gaussian noise distributions, we show that for any admissible fusion rule, the probability of error is a quasi-convex function of threshold τ. Hence, the problem decomposes into a series of n quasi-convex optimization problems that may be solved using well-known techniques.**

Assuming equal *a priori* probability, we give a sufficient condition of the non-Gaussian noise distribution $g(x)$ for the probability of error to be quasi-convex. Furthermore, this technique is extended to Bayes risk and Neyman–Pearson criteria. We also demonstrate that, in practice, it takes fewer than twice as many binary sensors to give the performance of infinite precision sensors in our scenario.

***Index Terms*—Distributed detection, fusion rule, generalized Gaussian noise, hard decision, non-Gaussian noise.**

I. Introduction

Consider distributed detection of $s \in \{-m, m\}$, where the ith of n local sensors observes $x_i = s + z_i$ with independent and identically distributed (i.i.d.) noise z_i. The ith sensor compares x_i to a threshold τ to compute a binary decision u_i as

$$u_i = \begin{cases} 0, & \text{when } x_i < \tau \\ 1, & \text{when } x_i \geq \tau. \end{cases}$$

Manuscript received November 3, 1999; revised July 18, 2000. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Sorrento, Italy, June 2000 and at the 33rd Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, October 1999.

The authors are with the Electrical Engineering Department, University of California, Los Angeles, CA 90095-1594 USA (e-mail: wshi@ee.ucla.edu; wsun@ee.ucla.edu; wessel@ee.ucla.edu).

Communicated by P. A. Chou, Associate Editor for Source Coding.
Publisher Item Identifier S 0018-9448(01)00582-X.

Each binary decision u_i is transmitted to a fusion center, which applies a fusion rule F to $k = \sum_{i=1}^{n} u_i$ to produce the final decision $F(k)$. The case with $n = 2$ and Gaussian noise is considered as an example in [1]. This correspondence extends this specific case to the general case with any number of sensors n for the family of generalized Gaussian noise and describes a technique for identifying the pair (τ, F) that minimizes the probability of error. For equal *a priori* probabilities, a sufficient condition of the non-Gaussian noise distribution $g(x)$ for the probability of error to be quasi-convex is given and some non-Gaussian noise distributions which satisfy this condition are listed. This technique extends to Bayes risk and Neyman–Pearson criterion.

The identical threshold τ in local sensors generally does not result in an optimum system. However, the identical threshold assumption reduces the complexity dramatically. For binary hypothesis detection, Irving and Tsitsiklis [2] showed that no optimality is lost with identical local detectors in a two-sensor system. Tsitsiklis [3] as well as Chen and Papamarcou [4] showed that identical local detectors are asymptotically optimum when the number of sensors n tends to infinity. These results provide some justification for restricting attention to identical quantizers.

Even with identical local thresholds, the problem is still complicated by the existence of multiple local minima. Furthermore, minimizing probability of error is difficult because the Bayesian error probability is not a smooth function, i.e., its first derivative is a discontinuous function. Hashlamoun and Varshney [1], [5] have overcome this difficulty by using a smooth bound on the Bayesian error probability to approximately determine the optimal pair (τ, F). Avi-Itzhak and Diep [6] provided an even tighter bound leading to a very good approximation. Even so, these techniques still produce a minimization problem that may have local minima, since neither convexity nor quasi-convexity was shown in these references.

In this correspondence, we show that for any admissible fusion rule F_i (i.e., any F_i that is optimal for at least one τ), $P_e(\tau, i)$ defined as the probability of error for (τ, F_i) is a quasi-convex function of τ. Quasi-convexity [7] means that every sublevel set $S_\alpha = \{\tau\colon P_e(\tau, i) \leq \alpha\}$ is convex. The admissible functions F_i are simply threshold tests of the form

$$s = F_i(k) = \begin{cases} -m, & \text{if } k < i \\ m, & \text{if } k \geq i. \end{cases} \tag{1}$$

Hence, the problem decomposes into a series of n quasi-convex optimization problems (one for each F_i) that may be solved exactly using a variety of techniques including, for example, the ellipsoid algorithm [8].

Section II shows the proof of the quasi-convexity of the probability of error versus τ for every admissible fusion rule F_i with generalized Gaussian noise. In Section III, with equal *a priori* probability, a sufficient condition of the non-Gaussian noise distribution $g(x)$ for the probability of error to be quasi-convex is given and some non-Gaussian noise distributions that satisfy this condition are listed. Section IV shows some illustrative examples. Section V extends this technique to Bayes risk and Neyman-Pearson criterion. Section VI concludes this correspondence.

II. Generalized Gaussian Noise

A. Admissible Fusion Rules

This section identifies the admissible fusion rules for generalized Gaussian noise and shows that the probability of error versus τ is quasi-convex for these rules.