□ 論 文 □

중요 표본추출 기법 이용

重要標本抽出 技法 利用한 交通網 區間의 混雜確率 推定

교통망 구간 혼잡 확율 추정

ESTIMATION OF THE CONGESTION PROBABLITY ON A TREE−TYPE TRANSPORATATION NETWORK BY IMPORTANCE SAMPLING

음 성 직
陰 盛 稷
(國土開發硏究院)

박 영 도
朴 泳 郁
(國土開發硏究院)

目 次

要 約

본 논문의 목적은 교통망 분석에 있어서 중요한 그러나 흔히 발생하지 않는 사건의 발생확률을 추정하는 방법론 개발에 있다. 예를 들어, 안정적(stable) 교통망에서 일시적인 혼잡현상이 발생할 확률을 씨뮬레이숀을 통해 추정하는 방법에 관한 것이다. 이 분야에 활발한 연구([3], [12])가 있어 왔으나 개괄적(Heuristic)방법에 제한되어 있었다.

본 논문은 위 문제에 대하여 포괄적(unified)이며 이론적인 방법론을 제시하였다. 이를 위해 대 분산이론(Large Deviation Theory)과 중요표본추출(Importance Sampling)기법이 이용되었으며 예로서 사용된 망은 두개의 구간이 이어진 교통망이다. 부수적으로 혼잡현상의 가장 대표적 형태를 구하는 방법이 제시되었다.

Ⅰ. Introduction

Consider a stable system of a tree−type network of queues. Let {Xn} denote the embedded r−dimensional Markov chain repersenting the system. Our goal is to estimate the probability that $\underset{1\leq i\leq r}{\text{Max}}\ X_{ni}$ or $\sum_{1\leq i\leq r} X_{ni}$ reaches a large number K within a busy cycle by simulation.

Direct simulation is very expensive, inefficient, and even impossible in cases, since the events of interest are extremely rare.

Instead, the change of measure for importance sampling is considered. It is known that if a Markov chain obeys a large deviation principle(LDP) with a rate function, the efficient change of measure for importance sampling can be obtained successfully.

After reviewing the use of importance sampling in simulating a M/M/1 queue, the paper apply this idea to the case of two queues in tandem by providing the LDP of the embedded Markov chain of the system.

Ⅱ. Review of M/M/1 Queue and Importance Sampling

Before stating the new results, a brief summary of the relevent theory relating to M/M/1 queue will be given. Consider a stable M/M/1 system, $\bar{S}$, with arrival rate μ and service rate λ with $\mu/\lambda \langle 1$. We are interested in estimating $\alpha = Pr[A]$, where $\{X_n\}$ is a sequence of successive population in the queue and A denotes the event that X_n hits K before becoming 0, assuming it is initially zero. Our concern is to understand how these values can be obtained by efficient sumulation.

The estimated value of α, $\hat{\alpha}$, may be obtained by direct simulation as follows;

Let us call a "cycle" a movement from 0 to the fist time either 0 is reached again, or K is reached. Define $V_m = 1_{\{X_n \text{ reaches } K \text{ in the } m^{th} \text{ cycle}\}}$. Since $\{X_n\}$ is a regenerative process with a regenerating point=0, V_m's are independent and identically distributed wth $E[V_m] = \alpha$. Therefore, $\hat{\alpha} = \frac{V_1+\cdots+Vp}{P}$ is an unbiased estimator of α.

To guarantee the following accuracy.

$$P_r[\frac{|\hat{\alpha}-\alpha|}{\alpha}\langle a]\rangle b$$

we have to generate $N\bar{s}$ cycles

$$N_s \rangle \frac{\gamma(b)^2 Var(V_m)}{a^2\alpha^2}$$

where $P[N(0,1)\langle\gamma(b)] = b$.

Sine we are interested in a very samll α, $Var(V_m) = \alpha - \alpha^2 \sim \alpha$. Therefore,

$$Ns \rangle \frac{\gamma(b)^2}{a^2\alpha}$$

It shows that the minimum $N\bar{s}$ depends on α. Since α has a very small value, it implies that we have to generate a large number of cycles to obtain precision to some degree. It is the main source of inefficiency in direct simulation.

To overcome this inefficiency, we consider "Importance Sampling". Main idea of importance sampling is to modify the orginal system so that its population grows much faster and, in consequence, the event happens more frequently and translate the estimate from the modified system in terms of the original one. For example, if a modification S of the orignal $\bar{S}$ is defined as M/M/1 with arrival rate μ and service rate λ, the event is more frequent in S than in $\bar{S}$.

Suppose $P\bar{s}$, Ps are the probability measures that govern $\bar{S}$, S, repectively and L_m's are the liklihood ratio between Ps and Ps in the m^{th} cycle. (Lm's can be easily computed for the m^{th} cycle.) Since the event is more likely under Ps than under $P\bar{s}$, Lm $\langle 1$ whenever $V_m=1$. Therefore, $V_{ar\bar{s}}(LmVm) \langle Vars(V_m)$. It implies that, to obtain a same error bound, we have to generate more cycles for $\bar{S}$ than for S.

Therefore, the center of importance sampling is on the question "what is the modification that minimizes Lm when Vm=1 ?" In other words, we have to find "the most likely way" that S reaches K and then modify the system to make the most likely way exceedingly likely. In case of M/M/1, there are various motheds to find the most likely way. (cf. [3], [12]).

Unfortunately, it does not seem to be possible to apply those methods used here to more complicated examples, such as the system of a pair of queues in tandem and a 3 dimensional tree-type network, etc.

Ⅲ. Markov Chain and Large Deviation Theory

In this section, we present a large deviation theorem due to Wentzel[8], regarding certain Markov chains. Consider the parameterized Markov chain $\{X_n^\epsilon\}$ given by

$$X_0^\epsilon = x_0$$

$$X_{n+1}^\epsilon = X_n^\epsilon + \epsilon V(X_n^\epsilon, \zeta_n),$$

where $\epsilon \rangle 0$ is the parameter which defines the Markov chain $\{X_n^\epsilon\}$, x_0 is the initial value, $V(\bullet, \bullet)$ is a function from $R^r \times R^l$ to R^r and ζ m's are i. i. d. r. v. 's. We are interested in analyzing $\{X_n^\epsilon\}$ when $\epsilon \to 0$.

Let F_x denote the d. f. of $V(x, \zeta n)$.

Let $m(x) = \int R^r zdFx(z)$

$$M_x(s) = \int R^r \exp\langle s, z\rangle\, dF_x(z)$$

$$lx(s) = \log Mx(s)$$

$$h_x(u) = \sup_{s \epsilon R^r} [\langle s, u\rangle - l_x(s)].$$

Assume the following :

(a) $Mx(s) \langle \infty$ in a neighborhood of 0 for each $x \epsilon R^r$

(b) $d(F_{x_1}, F_{x_2}) \langle c \mid X_1 - X_2 \mid$, where d is the Prohorov distance(see [1]), $c \rangle 0$ is a constant and $\mid \bullet \mid$ is a Euclidean norm.

Next, construct continuous-time paths from the realization of $\{X_n^\epsilon\}$. To do this, at the epochs $t = n \bullet \epsilon$, let $X^\epsilon(t) = X_n^\epsilon$ and interpolate piecewise linearly. Let C_T denote the set of the piecewise continuously differentiable functions $\phi : [0, T] \to R^r$ such that $\phi(0) = x_0$ is fixed. Let P^ϵ denote the measure induced by the $\{X^\epsilon(t)\}$ on the Borel -field of C_T, endowed with the Skorohod topology [1].

Define the rate function

$$I(x_0, \phi) = \begin{cases} \int_0^T h_{\phi(t)}(\phi'(t))dt, & \phi(0) = x_0 \\ \infty & , o. w. . . \end{cases}$$

Under (a), (b) with a couple of more technical assumptions we have the following results.

Theorm 1.

ⅰ) For each closed subset F of C_T,

$$\lim_{\epsilon \to 0} \sup - \epsilon \log P_x^\epsilon(F) \leq \inf_{\Psi \epsilon F} I(x, \Psi).$$

ⅱ) for each open subet G of C_T,

$$\lim_{\epsilon \to 0} \inf - \epsilon \log P^\epsilon(G) \rangle \inf_{\Psi \epsilon G} I(x, \Psi).$$

Therefore, if $\inf_{\Psi \epsilon A} I(\Psi) = \inf_{\Psi \epsilon AO} I(\Psi)$,

then $P_x^\epsilon(A) \sim \exp[-\frac{1}{\epsilon} \inf_{\Psi \epsilon A} I(x, \Psi)]$ (UTLE), as $\epsilon \to 0$.

(where UTLE is the acronym for up to logrithmic equivalence.)

Suppoes that we want to estimate the probability of event A,

A= $\{\mathrm{Max}_i\ X_{ni}$ exceeds K before hitting 0, given $X_0 = x_0\}$.

We consider the importance sampling method for this purpose.

Our goal is to find the optimal modification S^*, which statisfies

$\mathrm{Var}\ s^*(Ls^*mVm) \langle \mathrm{Var}\ s(LsmVm)$, for all modification S.

In other words, we want to solve the following optimization problem.

$$\underset{S}{\mathrm{Min}} \int_A L_s^2 \, d\, P_{\bar{s}},$$

when $L_s = \dfrac{dP_{\bar{s}}}{dP_s}$ is the likelihood ratio.

The following two observations are cirtical to find the optimal S^*;

Observation 1

If S^* is the optimal soution of the problem, then it also minimize

$$\int_A L_s^2 \, dP_s^{\epsilon}$$

Observation 2

If ϕ^* satisfies

$$\inf_{\phi \in A} I_{\bar{s}}(x, \phi) = I_s(x, \phi^*)$$

then S^* solves the following equation

$$I_{s^*}(x, \phi^*) = 0$$

Then, for any arbitrary small $\delta \rangle 0$,

$P_{s^*}^{\epsilon}(\phi_\delta^*)$ is asymtotically equal to 1. as $\epsilon \rightarrow 0$, where $\phi_\delta^* = \{\phi \in C_T : \| \phi - \phi^* \| < \delta \}$. ($\| \bullet \|$ is a uniform norm on C_T.)

In words, ϕ^* is the most likely path of the event A and S^* is the system in which ϕ^* is the most likely path among all the sample paths.

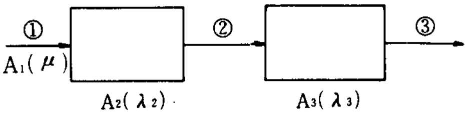
When we consider the Markov chain of queueing system $\{X_n\}$, the distribution function of the probability measure P, governing $\{X_n\}$, violates the assumed condition (b). Hence, we can not blindiy use these two observation to the case of queueing process in finding the optimal modification for importance sampling.

The discontinuity in the probabillity structure of a queueing system (violation of condition (b)), is mainly caused by the nonnegativity condition on the queue length. If there is no customer in a station, the station is in an idle state and no service is offered. This can be thought of a nonenegativity constraint on the queue length process, which imposess a change in the probabilistic behavior of the server when its queue becomes empty. This is the main obstacle in finding the LDP of the queueing system through the above results.

Ⅳ. Potential Process and its LDP

In order to avoid the nonnegativity constraint, we present potential process and Skorohod problem regrading networks of tandem queues, due to Park[9].

Let's Consider a network of tandem queues.



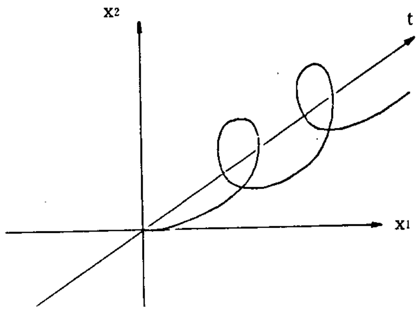
〈Figure 1〉 Network of Tandem Queues ((μ, λ_1, λ_2)−System)

$A_1(t)$: Arrival Process~Poisson Process with μ.

$A_i(t)$: Serivce time~Exponential dist. with λ_i and $\mu \langle \lambda_i$, for i=1, 2(stable system).

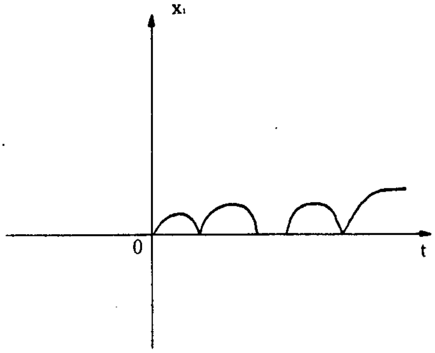
$X(t) = [X_1(t) \ X_2(t)]$,

where $X_i(t)$ is the number of customers at the station i at time t. X(t) is called the queue length process of tandem queues.

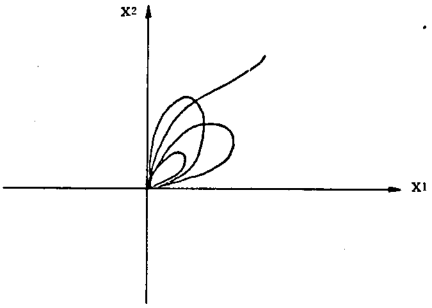


〈Figure 2〉 A Sample Trajectory on(t, x1, x2)

- Busy clycle is a movement from the empty state to the next empty state.
- Each cycle on the figure 4 represents a busy cycle. (Empty state part of trajectory is hidden at the origin.)



〈Figure 3〉 A Sample Trajectory in Station i.



〈Figure 4〉 A Sample Trajectory on (x_1, x_2)

We consider the potential process described as follows;

Let's consider a system in which servers provide units of services and they are not concerned whether there are customers in their stations or not. The time needed for the server in the station i generating a unit of service follows the exponential distribution with parameter λ_i. If they provide a service unit at a time when their queues are empty, then we consider the "queue length" to change from 0 to −1. Similarly if the queue length is already negative we simlpy decrement it by 1 again at the service time. The rest of the topological structure and probabilistic assumptions are defind same as the one for the original system. The variable we observe at time t at each station is the number of arrivals up to time t to the station minus the number of the service units generated in the station up to time t. Let $Y_i(t)$ be the observation in ith element of Y(t). We call Y "potential process".

Based on the results in the section 3, LDP of the Markov chain of the potential process, $\{Y_n\}$, can be easily established and its rate function on [O, T] is

$$I(x,\phi)=\begin{cases}\int_0^T h_\phi(\phi'(t))dt, & \text{if } \phi(0)=x\\ \infty & \text{, o. w.}\end{cases}$$

$$\text{where } h_x(U)=\inf_{\substack{U_1=S_1-S_2\\ U_2=S_2-S_3}}\{s_1k_\mu(\frac{1}{s_1})+s_2k_{\lambda_1}(\frac{1}{s_2})$$
$$+s_3k_{\lambda_2}(\frac{1}{s_3})\}$$

$$\text{and } k_y(s)=\begin{cases}ys-\log ys-1, & s\rangle 0\\ \infty & \text{, o. w..}\end{cases}$$

When we calculte I(x, ϕ), we assume $\mu+\lambda_1+\lambda_2=1$, without loss of generaity.

It is noticed that $h_x(\bullet)$ does not depend on x, Therefore, we denote $hx(\bullet)$ with $h(\bullet)$.

Therefore, LDP of $\{Y_n\}$ implies that

$$P_x^\epsilon(A)\sim \mathrm{Exp}[-\frac{1}{\epsilon}\inf_{\phi\epsilon A}\int_0^{T_\phi} h(\phi'(t))dt],$$

$\epsilon\rightarrow 0$, where T_ϕ is the firtst time when Max ϕ_i reaches K.

V. Skorohod Problem

Our objective in this section is to exhibit the queue length process of a network of tandem queues as a continous mapping Z of the potential process Y, i.e, $Z=\theta \cdot Y$. Then Z should have the same distribution as the queue length process of the network of the tandem queues.

The following theorem is a key to achieve the goal.

Let D^2_T be a set of 2 dimensional right continous functions with left limits for all t on [0, T].

$$\text{Let } M=\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

Theorerm 2.

For a given $\Psi(\cdot)$ with $\Psi(\cdot) \varepsilon R^2(t)$, there is a unique pair of function(ϕ, u), satisfying the following :

(1) $\phi(t)=\Psi(t)-Mu(t)$

(2) $\phi(t)\rangle 0$.

(3) $u_i(\cdot)$ is non-decreasing with $u_i(0)=0$ and $u_i(\cdot)$ increases only at those times t where $\phi_i(t)=0$

For a proof, see Park[9].

The problem of showing the existence of ϕ, u, is called "Skorohod problem" and the pair(ϕ, u) is called the solution of the Skorohod problem.

Let θ and σ be a mapping from C_T to C_T, defined by $\theta(\Psi)=\phi$ and $\sigma(\Psi)=u$.

We name θ "Skorohod map". Based on this map.

Let $Z(\omega)(t)=\theta(Y(\omega))(t)$, $U(\omega)=\sigma(Y(\omega))(t)$. In Lemma 2, we investigate the probability structure of the process Z.

Lemma 2

(1) The processes Z and U satisfy the following almost surely;

1. $Z(t)\langle\rangle 0)\varepsilon R^2$.
2. $U_i(t)$ is nodecreasing with $U_i(0)=0$.
3. $U_i(t)$ increases only at those time t where $Z_i(t)=0$,

(2) Z is a time homogeneous Markov process.

It implies that the process Z and the queue length process X have the same small time increment conditional probability and, therefore, they have the same infinitesimal generator. It follows that they have the same distribution. (cf. p.111 and p. 161 in Ethier and Kurts[7].)

Ⅵ. LDP of Network of Tandem queues

Since the process Z and the queue length process X have the same distribution, the Markov chain $\{Z_n\}$ of Z and $\{X_n\}$ have the same probability structure. It follows that for any closed set F,

$P^\epsilon{}_{\{Z_n\}}(F)=P^\epsilon{}_{\{Y_n\}}(\theta^{-1}(F))$.

and $\theta^{-1}(F)$ is a closed set.

Therefore, by the property of the LDP of the potential process,

ⅰ) for any closed set F,

$\lim_{\epsilon\to 0} \sup -\epsilon \log P^\epsilon{}_{\{Z_n\}}(F \mid x_0=x) \langle \inf_{\phi \varepsilon F} \inf_{\Psi \varepsilon \theta^{-1}(\phi)} I(x, \Psi)$.

By the same reson.

ⅱ) for any open set G,

$\liminf_{\epsilon\to 0} -\epsilon \log P^\epsilon{}_{\{Z_n\}}(G \mid x_0=x,) \langle \inf_{\phi \varepsilon G} \inf_{\Psi \varepsilon \theta^{-1}(\phi)} I(x, \Psi)$.

Therefore, LDP of $\{X_n\}$ is well-defined with its rate function;

$$I_{\{X_n\}}(x, \phi)=\begin{cases} \inf_{\Psi \varepsilon \theta^{-1}(\phi)} \int_0^T h(\Psi'(t))dt, & \text{if } \phi(0)=x \\ \infty, & \text{,o.w..} \end{cases}$$

Ⅶ. The Most Efficient Change of Measure for Importance Sampling

The observation 1 and 2 present the method to find the optimal change of measure.

First, we find the most likely path of the event A.

$$\inf_{\phi \in A} I_{\{X_n\}}(x, \phi)$$

$$= \inf_{\phi \in A_0} \inf_{\Psi \in \theta^{-1}(\phi)} I_{\{Y_n\}}(\Psi),$$

where $A_0 = \{\phi \in A : \phi(0)=0\}$

$$= \inf_{\phi \in A_0} \inf_{\Psi \in \theta^{-1}(\phi)} \int_0^{T_\phi} h(\Psi'(t))dt$$

$$= \inf_{\substack{S_1 > 0,\ S_2,\ S_3 > 0 \\ S_1 - S_2 > 0 \text{ or } S_2 - S_3 > 0 \\ S_1 + S_2 + S_3 = 1}} T\phi \left[s_1 h_\lambda\left(\frac{1}{S_1}\right) + s_2 h_{\mu_1}\left(\frac{1}{S_2}\right) + s_3 h_{\mu_2}\left(\frac{1}{S_3}\right) \right],$$

$$\text{where } T_\phi = \begin{cases} \dfrac{1}{S_1 - S_2}, & \text{if } S_1 < S_2 \text{ and } S_2 > S_3 \\ \dfrac{1}{S_1 - S_2} & \text{o. w.} \end{cases}$$

This nolinear programming problem can be easily solved by Kuhn–Tucker condition and the minimum is achieved at $S_1 = \lambda_1$, $S_2 = \mu$, $S_3 = \lambda_2$, when $\lambda_1 < \lambda_2$ and at $S_1 = \lambda_1$, $S_2 = \lambda_2$, $S_3 = \mu$, when $\lambda_2 < \lambda_1$

That is, the minimum is obtain when μ is exechanged with smaller of λ_1 and λ_2

Ⅷ. Future Research

This result can be easily generalized to the case of general tree–type networks. To study the congestion event in transportation networks more precisely, this research should be generalized to the case that the service rates and arrival rates are functions of state variables. The forcus of continuing effort will be in this direction.

Bibliography

[1] P. Billingsley. Probability and Measure. John Wiley, 1979.

[2] C. Costantini. The Skorohod Oblique Reflection Problem and a Diffusion Approximation for a Class of Transport Processes. Ph. D thesis, University of Wisconsin, Madison, 1987.

[3] M. Cottrell, J. C. Fort, and G. Malgouyres, "Large Deviations and Rare Events in the Study of Stochastic Algorithms." IEEE Trans. on Auto. Control. 28 : 907–920, 1983.

[4] H. Cramer, "On a New Limit Theorem in the Theory of Probability." Colloquium on the Theory of Probability, Herman, Paris, 1987.

[5] P. Dupuis and H. Ishii, "On When the Solution to the Skorohod Problem is Lipschitz Continuous, with Applications." Stochastic and Stochastic Reports. 35 : 31–62, 1991.

[6] P. Dupuis. "Large Deviation Estimates for Queueing System." Conference Proc.. Imperial College workshop, to appear.

[7] S. N. Ethier and T. G. Kurtz. Markov Processes : Charaterization and Convergence, Jonh Wiley, 1986.

[8] M. I. Freidlin and A. D. Wentzell. Random Perturbation of Dynamical Systems. Springer, Berlin, New York, 1984.

[9] Young W. Park. Large Deviation Theory for Queueing System, PhD. Dissertation, VPI & SU, 1991.

[10] H. Tanaka. "Stochastic Differential Equations with Reflecting Boundary Con-

ditions in Convex Regions." Hiroshima Math. J.. 9 : 163－177, 1979.

[11] S. R. S. Varadhan, Large Deviations and Application, SIAM, 1984.

[12] A. Weiss. "A New Technique for Analyzing Large Traffic System." Adv. Appl. Prob.. 18 : 506－532, 1986.