REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 12, 2001	Final; Feb 15, 1997 to Sep 30, 2000

4. TITLE AND SUBTITLE

Some Problems in Probability, Statistics and Reliability

5. FUNDING NUMBERS

35904-MA
DAAG 55 – 97 – 1 – 0024

6. AUTHOR(S)

Jayaram Sethuraman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Florida State University, Tallahassee, FL 32306-4330

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARO 35904.1-MA

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Research was carried out in the areas of Image Analysis modeling the phenomena of conformation and coupling, Markov Chain Monte Carlo methods – their limitations and range of applicability, Bayesian Nonparametric Computations – approximations and convergence, predicting a random variable based on a discretized covariate, Flow Models – flows in queues with finite capacity buffers and flows in networks with losses abd capacity controls, and survey articles on modern nonparametric methods.

14. SUBJECT TERMS

Image Analysis, Markov Chain Monte Carlo, Bayesian Nonparametric Methods, Flow Models

15. NUMBER OF PAGES

9

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

Final Report

1. **ARO Proposal Number**	35904–MA
2. **Period Covered by Report**	February 15, 1997 through September 30, 2000
3. **Title of Proposal**	Problems in Probability, Statistics and Reliability
4. **Contract or Grant Number**	DAAG55-97-1-10024
5. **Name of Institution**	The Florida State University
6. **Author of Report**	Jayaram Sethuraman

1 MAIN TOPICS OF RESEARCH CARRIED OUT UNDER THE GRANT

Research was carried out in the areas of Image Analysis modeling the phenomena of conformation and coupling, Markov Chain Monte Carlo methods - their limitations and range of applicability, Bayesian Nonparametric Computations - approximations and convergence, predicting a random variable based on a discretized covariate, Flow Models - flows in queues with finite buffers and flows in networks with losses and capacity controls, and survey articles on modern nonparametric methods.

The later sections give more detailed non-technical summaries on the actual research that ensued from this grant.

2 Publications and Technical Reports under the Grant

1 "Conformation in Metric Pattern Theory: Strong Coupling" jointly with Ulf Grenander

2 "Conformation in Metric Pattern Theory: Weak Coupling" jointly with Ulf Grenander

2 "Joint distributions, conditional distributions and the Gibbs sampler" jointly with K. Athreya

4 "Ergodicity and Tail σ-fields of Markov Chains" jointly with Jim Lynch

5 "Strong Approximations to Dirichlet Distributions with Applications"

6 "Optimal discretization of the independent variable X for predicting the dependent variable Y

7 "Asymptotic Bounds on the Overflow Probability in Markov-Modulated Fluid Models" jointly with Shau-Ming Wu

8 "Loss of Power in Transmission Networks with a Large Number of Nodes"

9 "Flows, Capacities of Channels and Utilization of Networks"

10 "Nonparametric Statistics: Rank-Based Methods" jointly with Myles Hollander

11 "Nonparametric Statistics: Advanced Computational Approaches" jointly with Myles Hollander

3 Nontechnical summary of research carried out under the grant

Papers nos. 1 and 2 deal with problems in Image Analysis. We have known for some time how to generate random images that are realistic, by describing all realistic images as transformations of a template, and placing a distribution on the space of transformations, which generally forms a group. In papers nos. 1 and 2, we model the boundaries of the two random cells as straight lines connecting a large number of sites which are randomly perturbed. The random boundaries of the two cells, attract one another so that they fuse on a particular section of their boundaries. This is modeled by expressing the energy function, describing the randomness of the boundaries, as the sum of three factors - two to describe the randomness of the cells as if they were behaving independently, and the third, parameterized by a constant, called the force of coupling, forcing the the cells to conform on a certain section of the boundary. The final random boundaries of the cells are influenced by these three factors. The influence of the first two factors has been studied before in many papers, including ours. In paper no. 1 we allow the number of sites on the boundary to go to ∞ and the force of coupling to go to ∞. We call this the case of strong coupling. We show that the limit is still random but shows the strong tendency to conform.

In paper no. 2, we allow the number of sites to go to ∞ but keep the force of coupling to remain constant. We call this the case of weak coupling. For this case, we obtain the limiting distribution. The mean of this distribution is shown to be the solution of a second order differential equation which exhibits the effects of conformation. We also present several generalizations of the energy functions, and present similar results.

Paper 3 clarifies the situation in Gibbs sampling, where information pertaining conditional distributions alone is used to generate an observation from the joint distribution. Suppose that P and Q are transition functions on $S_1 \times \mathcal{S}_2$ and $S_2 \times \mathcal{S}_1$, respectively; these are like two conditional distributions. We explore conditions for

the existence and uniqueness of a joint distribution π with conditional distributions P and Q as well as the convergence of the associated Gibbs sampler to this π. Roughly speaking, what is needed is a multiplicative condition on P and Q with appropriate integrability and an irreducibility condition on $R = PQ$. Examples are given to illustrate the consequences of the violation of some of these conditions and to demonstrate that the mere convergence of the Gibbs sampler does not insure the uniqueness of the joint distribution. It is also shown that Markov chains arising in Gibbs sampling are necessarily aperiodic. Similar results are obtained when we studying more than two variables.

Paper 4 deals with an interesting relationship between the ergodicity of Markov chains and the triviality of their tail σ-fields. Essentially, the main result states that if the Markov chain has a stationary distribution and the tail σ-field is trivial, then the Markov chain is ergodic.

Dirichlet process priors have proved very useful in Bayesian nonparametric analysis. There has been a lot of progress with computational Bayesian methods and it is but natural that results concerning approximations to Dirichlet priors will be very useful. In paper no. 5, we show two types of approximations for Dirichlet process clarify the nature of convergence and prove the convergence. Some applications to hierarchical Bayes problems are also presented.

Paper no. 6 arose from listening to scientists at the DOD Polygraph Institute. They want to predict a certain random variable Y on the basis of another variable X, which will be used only in a discretized form. That is they divide the range of X into 7 intervals and consider the discretized random variable X' taking values $1, 2, \ldots, 7$ on these intervals. The intervals used for this purpose looked strange, but the scientists said that it was based on recommendations of earlier scientists. I suggested that they use equiprobable intervals of X for this discretization, based on an analogy in goodness-of-fit problems. The scientists worked on the new discretization and reported good results; they also published a paper in a journal on this method of discretization. It is only later it struck me that I should go and theoretically prove that this is the optimal thing to do. This is the genesis of this paper. To state it a little more formally, consider a pair of random variables (X, Y) and suppose that we want to predict the dependent variable Y from the independent variable X. Break up the range of X into k intervals and define the discretized variable X' as equal to $1, 2, ..., k$ on those intervals. Can X' predict Y as well as X can predict Y? What is the loss in predictive ability? For a fixed value of k, what is the best way to divide the range of X into k intervals? We describe the optimum way to divide the range of X; this optimum possesses some surprising properties for which we do not see an obvious intuitive explanation. The intervals of this optimal dicretization are not

equiprobable intervals, as I had suggested earlier. We give a complete solution for the case when (X, Y) is bivariate normal. Even though the discretization based on equiprobable intervals is not optimal, its performance is very close to the optimal as seen from numerical calculations. This may explain why the DODPI scientists were getting good results based on my suggestion.

The topic of flows appears in many contexts. Flows could refer to the flow of electric current or telecommunication signals through a network. It could refer to the flow of water through several dams, or to customers through a queue. Paper no. 7 deals with the latter kind of flows while papers nos. 8 and 9 deal with the former.

In paper no. 7 we are examining strategies to keep a dam or a storage device as full as possible and not let it overflow, because there is a severe penalty for such an overflow. We use the the terminology of queuing theory to state this problem more precisely. Consider a buffer of finite size fed by input sources and emptied by demand sources, all modeled by Markov processes. Such models are relevant in telecommunication networks, computer networks and inventory systems. In such systems, an overflow of the buffer corresponds to a catastrophic failure. Engineers would like to find ways of designing the system and the buffer size so that the probability of such an event is greatly minimized. In this paper we establish a large deviation principle which allows us to construct asymptotic bounds on the overflow probability in Markov-modulated fluid flow models as the buffer size goes to infinity. These asymptotic bounds are useful in the optimum design of physical systems governed by such models, namely to reduce costs and maximize performance.

Papers nos. 8 and 9 arose from conversations with scientists interested in efficient networks for transmission and in efficient utilization of networks. Suppose that a certain number units of power (strength of signal) are present at an originating node and are transmitted through a large number of intermediate nodes. Between nodes, there will be naturally occuring dissipation as well as some boosting of power obtainable at some cost. We explore a general probabilistic model to find the distribution of the total loss/boost in transmission, which is the same as knowing the distribution of the final remaining units of power (strength of signal) after transmission has traveled though the large number of nodes.

To make the problem more mathematical, we assume that X units are present at the originating node 0 and it is transmitted through nodes $i = 1, 2, \ldots, n$. The number of units available at node i, after losses/boosts at previous sites, will be denoted by X_i, $i = 1, 2, \ldots, n$. The sites themselves do not have to be on a straight line. We are interested in the final remaining units of power (strength of signal) after traveling through nodes $1, 2, \ldots, n$, that is in the distribution of X_n, where n is large.

There is both dissipation and boosting of power (signal) en route, and so the

loss/boost factors, namely the non-negative random variables $p_i = \frac{X_i}{X_{i-1}}, i = 1, 2, \ldots, n$ will play an important role. If the p_i's are all strictly less than 1, there is only dissipation, and thus X_n, the available power at the end, will be close to zero. To avoid this straightforward case, one must assume that the p_i's hover around 1, which means that it can take values less than and greater than 1; this corresponds to the practical consideration that there must be boosters available en-route.

We exhibit a class of distributions for the loss/boost factors and give a range of parameters for these distributions under which the final signal strength has a limiting distribution. We discover further conditions under which the average of the final signal strength will be equal to the initial strength. Under other alternative conditions there is a loss or a boost. These results can be used to help in the design of transmission networks.

In paper no. 9, we consider a network with X units of power at the source. The power flows through a large number of available channels $C_1, C_2, \ldots, C_n$. Each channel C_i has a **lower capacity** L_i and an **upper capacity** U_i. Channel C_i will fall into disuse if one tries to transport less than L_i units through it and thus one should maintain a flow of at least L_i units all the time. Trying to transport more than U_i units through channel C_i will destroy that channel. We will assume that units to be transported X, the lower capacities L_i and upper capacities U_i are random. We define a measure M of the level of utilization of the network. We will give approximations to the probability that the the available channels are adequate, and obtain the limiting distribution of the measure of utilization M.

Papers nos. 10 and 11 are invited articles for the International Encyclopedia for Social and Behavioral Sciences describing the field of Nonparametrics. Paper no. 10 describes rank based methods and paper no. 11 deals with the recent advanced nonparametric methods based on computers.

The success of Nonparametric methods is based on their wide applicability; these methods typically require only modest assumptions concerning the underlying populations from which the data are drawn. Under these mild assumptions, exact hypothesis tests, exact confidence intervals, exact multiple comparison procedures, and exact confidence bands can be obtained. Nonparametric methods have excellent efficiency properties with respect to their parametric competitors and are also robust in the sense that they are relatively insensitive to outlying observations and departures from the model. In paper no. 10, we present some classical rank-based nonparametric methods in two- and k-sample location problems.

Statistical methods are useful in obtaining information about the unknown state of nature or the "parameter" as it is usually referred to in the literature. A statis-

tician collects suitable data whose distribution depends on the unknown parameter. A statistical inference procedure is then devised to produce information about the unknown parameter or a function of that parameter. The classical methods of inference assume that the probability distributions that govern the data depend only a few unknown quantities (or parameters). Such procedures are called parametric procedures. When one is not able to make such strong assumptions about the probability models that govern the data, and/or when one can be more sure of the rankings among the data than the exact values, one uses robust inference procedures known as nonparametric methods. Tests and estimates based on the simpler nonparametric methods can be obtained from easy computations based on the data. Procedures that have validity under more general models will require more heavy computations. When there is some uncertainty about the probability models and/or when expert information concerning the problem at hand is available, one should use Bayesian methods. These methods can also tend to be computationally intensive. In paper no. 11 we describe briefly the main ingredients of some new Bayesian and computational methods in nonparametric inference. These include bootstrapping and Gibbs sampling.

4 Professional activities during the period covered by the grant

- Participated in the ASA/IMS Joint Annual Meeting at Anaheim, CA – August 1997.
- Participated in the Army Statistician's Conference at George Mason University, Fairfax, VA – October 1997.
- Presented an invited paper "Further properties of Dirichlet measures" at the Meeting of the Bernoulli Society held at the Indian Statistical Institute, Calcutta, India, January 1998.
- Gave an invited talk "Further properties of Dirichlet measures" at the 1998 Luckacs Symposium "Statistics for the 21st Century" at Bowling Green University, Bowling Green, OH – April 24-26, 1998

- Gave an invited talk "Further properties of Dirichlet measures" at the International Conference in Reliability and Survival Analysis at Northern Illinois University, Dekalb, IL – May 21-24, 1998

- External Reviewer of the Applied Statistics Master's Program at New Jersey Institute of Technology, Newark, NJ – June 8, 1998

- Presented an invited paper "Specification of Joint Distributions from Marginal and Conditional Distributions" at the Symposium on Decision Theory at Purdue University, Lafayette, IN – June 18-21, 1998

- Participated in the ASA/IMS Joint Annual Meeting at Dallas, TX – August9-13, 1998

- Presented a paper "Conformation in Metric pattern Theory" at the Army Statistician's Conference at New Mexico State University, Las Cruces, NM – October 21-23, 1998

- Gave a talk "Conformation in Metric pattern Theory" at the Meeting of the Florida Chapter of the American Statistical Association, University of Florida, Gainesville, FL – March 1999

- Participated in the ASA/IMS Joint Annual Meeting at Baltimore, ND – August 1999

- Gave a colloquium talk "Reduction in Predictive Ability Caused by Discretization of the Independent Variable" in the Department of Statistics, Florida State University, October 1999

- Presented a paper "Reduction in Predictive Ability Caused by Discretization of the Independent Variable" at the Army Statistician's Conference at West Point, NY – October 1999

- Presented an invite plenary talk "Limit Theorems for Models in Pattern Analysis" at the International Conference on Stochastic Processes and its Applications at Cochin, India – December 1999

- Presented an invited talk "Approximations to Dirichlet Processes" at the 2000 Annual meeting of the Canadian Statistical Association in Ottawa in June 2000.

5 Ph. D. Degrees Awarded

Shau-Ming Wu "Asymptotic Bounds for Markov Modulated Fluid Models, Based on the Large Deviation Principle".

6 Honors Received

None

7 Service to DOD Institutes

Member of the Scientific Review Panel of the DOD Polygraph Institute

Outside of this committee, I have given research advice, which has improved some of their statistical practices and also led one member of the Institute to publish a paper in a journal.