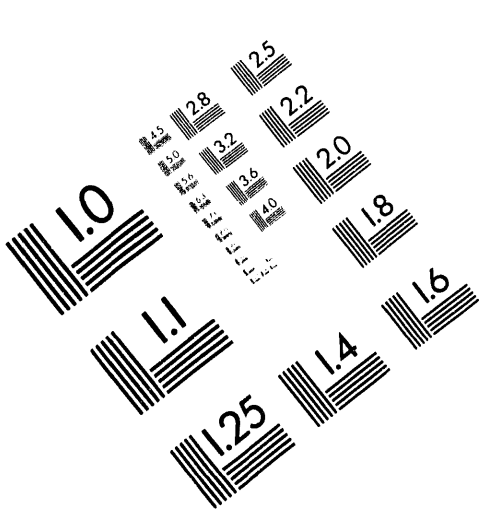
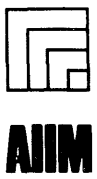
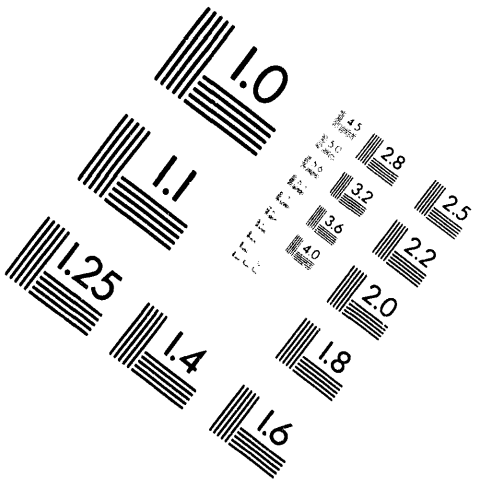


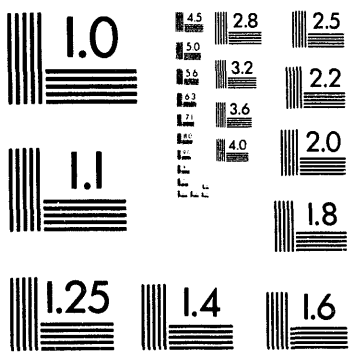


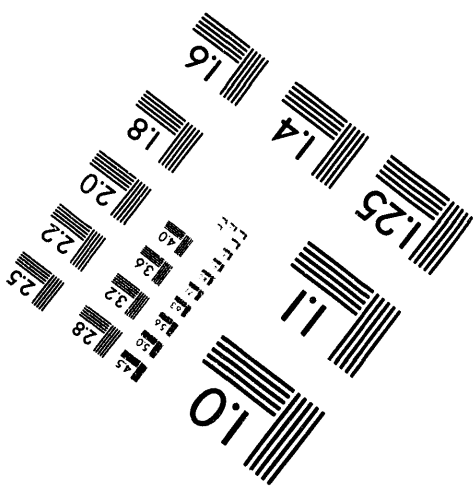
Association for Information and Image Management

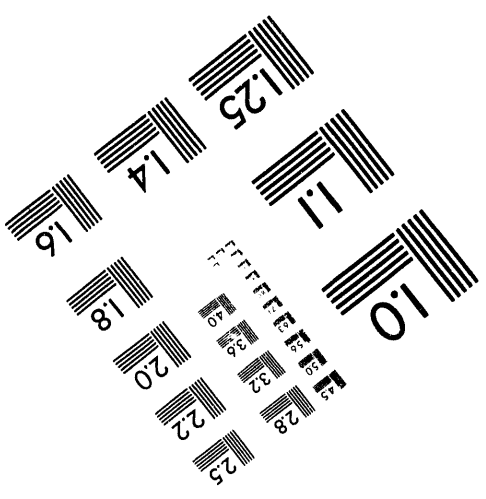
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202









1 of 2

RARE EVENT SIMULATION IN RADIATION TRANSPORT

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF STATISTICS

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Craig Kollman

October 1993

Abstract

This dissertation studies methods for estimating extremely small probabilities by Monte Carlo simulation. Problems in radiation transport typically involve estimating very rare events or the expected value of a random variable which is with overwhelming probability equal to zero. These problems often have high dimensional state spaces and irregular geometries so that analytic solutions are not possible. Monte Carlo simulation must be used to estimate the radiation dosage being transported to a particular location. If the area is well shielded the probability of any one particular particle getting through is very small. Because of the large number of particles involved, even a tiny fraction penetrating the shield may represent an unacceptable level of radiation. It therefore becomes critical to be able to accurately estimate this extremely small probability.

Importance sampling is a well known technique for improving the efficiency of rare event calculations. Here, a new set of probabilities is used in the simulation runs. The results are multiplied by the likelihood ratio between the true and simulated probabilities so as to keep our estimator unbiased. The variance of the resulting estimator is very sensitive to which new set of transition probabilities are chosen. It is shown that a zero variance estimator does exist, but that its computation requires exact knowledge of the solution.

A simple random walk with an associated killing model for the scatter of neutrons is introduced. Large deviation results for optimal importance sampling in random

walks are extended to the case where killing is present. An adaptive "learning" algorithm for implementing importance sampling is given for more general Markov chain models of neutron scatter. For finite state spaces this algorithm is shown to give, with probability one, a sequence of estimates converging exponentially fast to the true solution.

In the final chapter, an attempt to generalize this algorithm to a continuous state space is made. This involves partitioning the space into a finite number of cells. There is a tradeoff between additional computation per iteration and variance reduction per iteration that arises in determining the optimal grid size, which require more work, but achieve a greater variance reduction. All versions of this algorithm can be thought of as a compromise between deterministic and Monte Carlo methods, capturing advantages of both techniques.

Acknowledgements

I am grateful to everyone in the Department of Statistics who has made my time here at Stanford a very rewarding educational experience. The friendly atmosphere among students and professors has helped make this a place very conducive to learning. I have enjoyed the support and companionship of my fellow students. I thank Derek Boothroyd, Josée DuPuis, Chin Shan Chuang, Todd Graves, Eric Kolaczyk, Stephanie Land, Nicole Lazar, Charles Roosen, Steve Stern, Michael Wolf, and Linfeng You for many interesting conversations and good times. The professors all did an excellent job teaching classes. I am especially grateful to Ani Adhikari and David Siegmund for their outstanding courses in probability theory. I would also like to thank Judi Davis for her never ending assistance in many different areas.

I am indebted to my advisor, Peter Glynn, for his encouragement and guidance throughout my thesis research. His enthusiasm for applied probability has lured many students, including myself, into that field. I thank him for his many good ideas and suggestions concerning this dissertation.

I would like to thank Tom Booth and Art Forster for introducing me to this topic and showing interest in my research. I appreciate the financial support of my work by the Radiation Transport Group at Los Alamos National Laboratory. I am also very grateful to the Statistics Group at Los Alamos. My summers there have given me valuable experience working on real world problems. I have learned a great deal from the many excellent statisticians in that group.

Finally, and most importantly, I thank my parents whose love and support have helped me make it through graduate school. They were there for me when things got tough to tell me not to give up.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Problems in radiation transport often involve estimating extremely small probabilities or the expected value of a random variable which is with overwhelming probability equal to zero. High dimensional state spaces and irregular geometries make the problems complex enough to require Monte Carlo simulation for their solution. Ordinary simulation for such rare events is typically insufficient to produce reliable estimates in a reasonable amount of computer time.

Importance sampling is one technique which can dramatically improve the efficiency of our calculations (see Glynn & Iglehart [9]). Here the transition probabilities for the path of the particle are changed in the simulation. The result is multiplied by the likelihood ratio between the true and simulated probabilities in order to keep our estimator unbiased. The variance of the resulting estimator depends heavily on how we choose our new set of transition probabilities. A good choice can give us a substantial variance reduction while a poor choice can increase the variance of our estimator. This dissertation studies methods for making that choice so that the variance of our estimator is made as small as possible.

1.1 Radiation Transport

Suppose we have a neutron emitter at a known site and wish to predict the radiation dosage transported to a particular location. A neutron leaving the emitter will travel a random distance before colliding with a molecule. Upon collision, it may be absorbed with a certain probability, or it may scatter in a random direction with a random change in its energy level. This continues until the neutron is absorbed.

We assume that interactions between neutrons are negligible so that we may simulate their paths one at a time. The radiation dosage at our target location will be the rate at which the source is emitting neutrons multiplied by the expected energy delivered to the target per neutron leaving the source.

The probabilistic mechanism for the scatter of the neutron is assumed to be known from the theory of physics (see Lux & Koblinger [13] and Lewis & Miller [12]). Its path may be modelled as a Markov chain provided the state space contains enough information on time, location, velocity, and energy level.

Figure 1.1 shows a two dimensional version of a simple transport problem. Suppose we have a radioactive source inside a nuclear reactor. We want to shield off a control room so that operators will not be exposed to an unacceptable level of radiation. The arrows show a potential path a particle may take to introduce radiation into our "protected" area. Realistic problems are complicated by the irregular geometries within the reactors: machinery, bending pipes, air ducts, etc.

The probability of reaching the target location is very small for any one particular neutron leaving the source. In order for our particle to *score*, or hit the target, it must penetrate the lead shield, leave the shield at an angle in the general direction of the control room, avoid being absorbed or deflected away from the control room by the pipe, penetrate the cement wall, and scatter off the wall in the direction necessary to reach the target. Because of the large number of neutrons being emitted by the

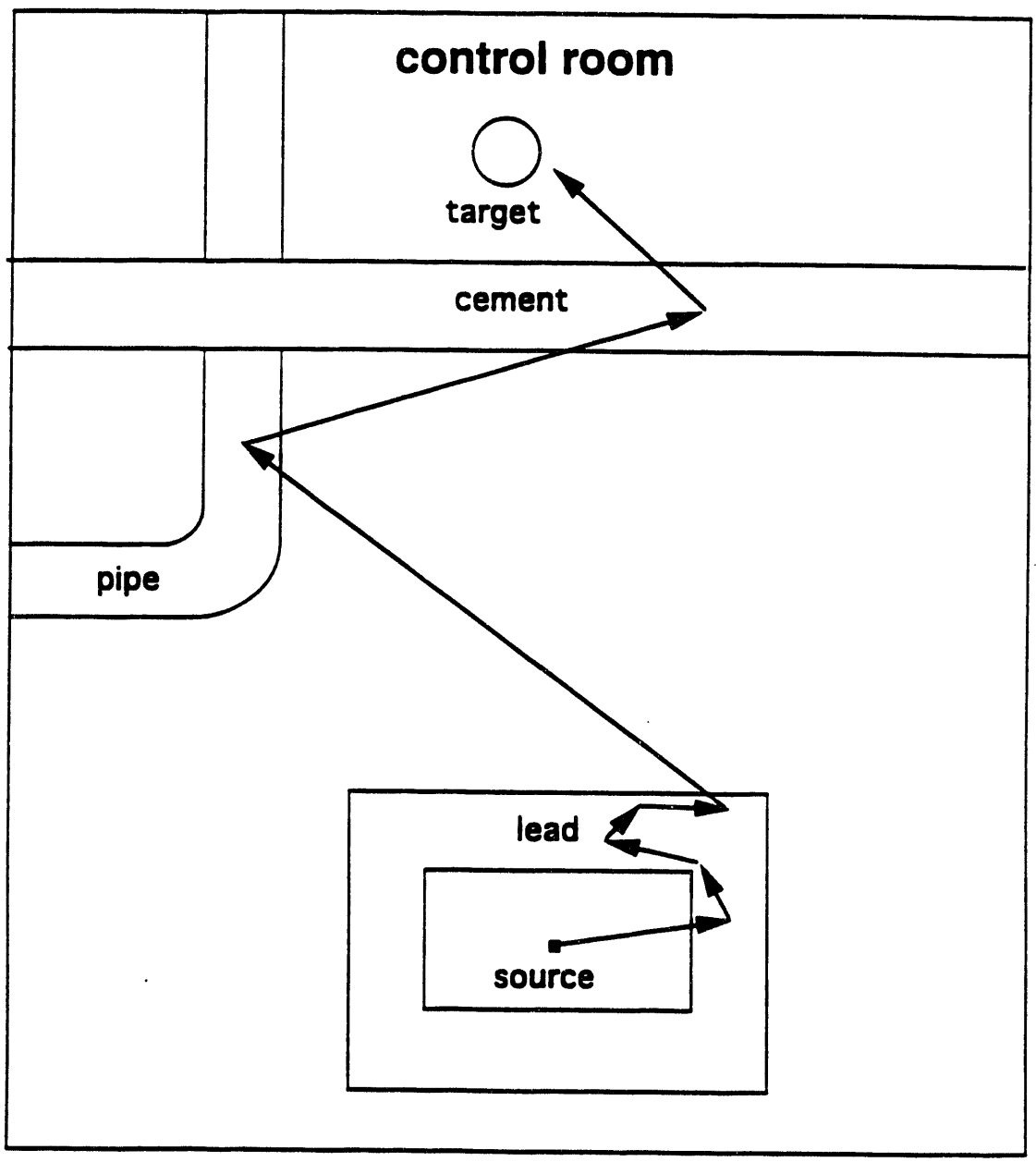


Figure 1.1: A sample transport problem

source, some will reach the target. To predict the amount of radiation at the target we must be able to accurately estimate the expected energy delivered to the target by a single neutron leaving the source.

These types of transport problems are of great interest in reactor design. Applications include reactor safety, shielding problems, criticality safety, and nuclear safeguards. Transport problems also arise in oil well logging where a nuclear source is placed in a borehole along with a detector. By observing how the radiation is transported from the source to the detector, inferences can be made about the geology surrounding the borehole (see Ullo [24]).

Deterministic methods of solution are difficult because of the high dimension of the state space. The problem may require as many as seven dimensions (three for position, three for velocity, and one for time). To solve deterministically, it is necessary to approximate the continuous state space with a finite number of cells. The number of cells required, and hence the amount of work needed for solution, grows rapidly as the partition becomes finer and finer.

The problem is further complicated by a very erratic probability transistion kernel. Figure 1.2 shows the cross section of iron as a function of energy. The cross section represents the instantaneous hazard rate for a collision within an iron medium. We can see that a very slight change in the particle's energy level can change the collision probability by orders of magnitude. Thus, for the discrete version of the problem to closely resemble the continuous problem, the partition must be very fine. In seven dimensions, this requires an excessive number of cells.

Monte Carlo simulation is an alternate method of solution that does not suffer from the curse of dimensionality. The simulation can be done directly in the continuous space avoiding the need for a discrete approximation. By the central limit theorem, we know that the accuracy of our estimator grows as the square root of the amount of work done. This rate does not depend on the dimension of the problem so that

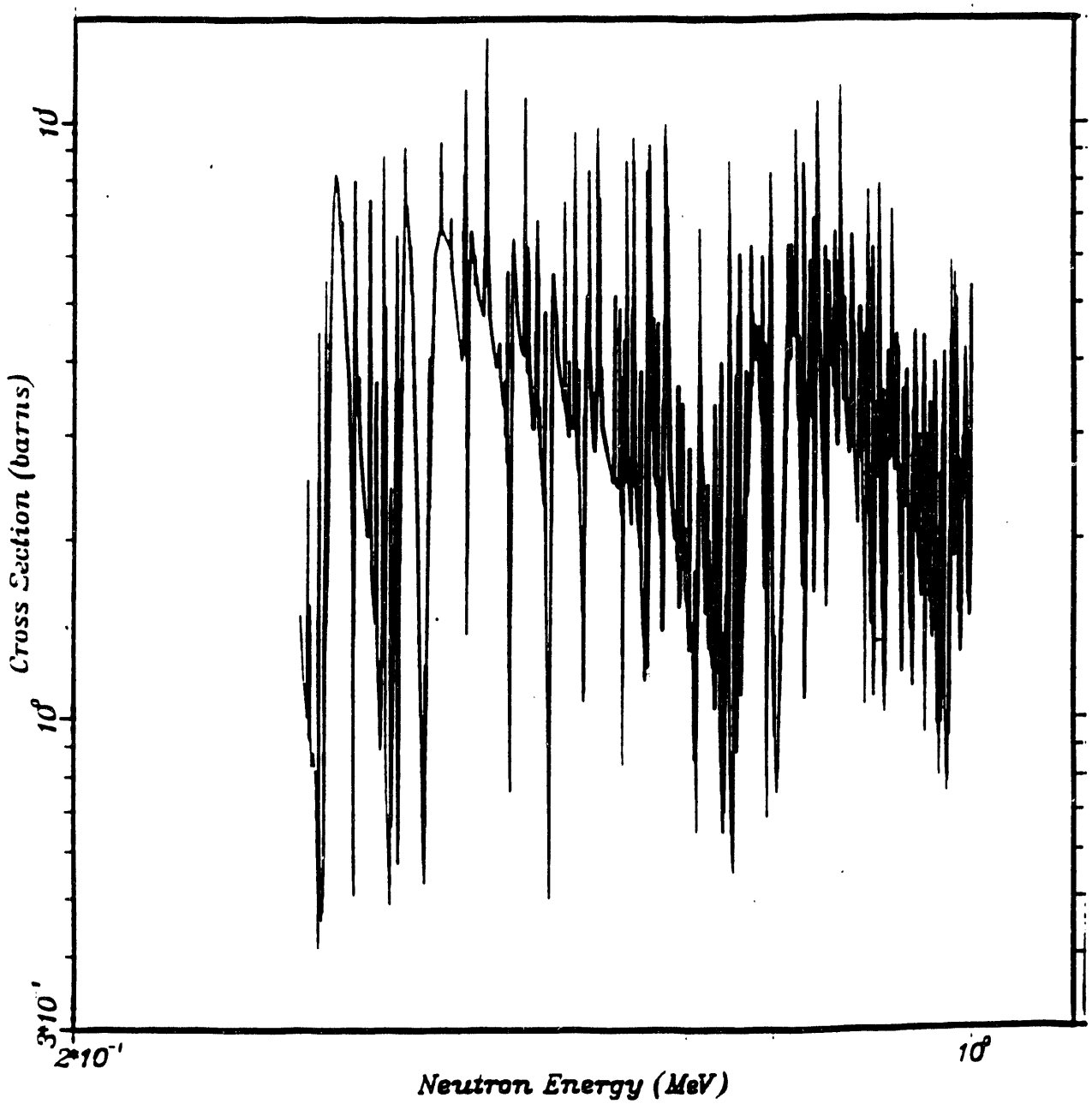


Figure 1.2: Nuclear cross sections of iron at various energy levels.

Monte Carlo may be superior in high dimensional state spaces. However, without methods of variance reduction, simulation would be impractical to estimate such small probabilities.

A great deal of effort has been made to develop methods of efficient simulation of transport problems. The problem is discussed frequently in the nuclear engineering literature. Booth [4], [5], & [6], Troubetzkoy [23], and Cramer et al [8] are good examples. Much progress has been made towards creating programs that can give accurate solutions to realistic problems (see [14]), but these problems are difficult enough so that more efficient methods are still being sought.

1.2 Importance Sampling

Suppose we have a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X defined on that space. We wish to use simulation to estimate the expected value of that random variable, $E^P X$. If we choose a probability measure, Q, on that space such that $Q \gg P$ (ie Q does not assign probability zero to an event having positive probability under P), we may wish to perform the simulation under Q instead of the "true" probability measure P. By multiplying the random variable by the likelihood ratio between P and Q, we obtain an unbiased estimator for the expected value of X under P. That is.

$$E^P X = E^Q \left(\frac{dP}{dQ} \cdot X \right) \tag{1.1}$$

which is the result of the well known Radon-Nikodym theorem (see Theorem 32.2 of Billingsley [1].

In transport problems, Ω is the set of all possible paths a particle can take leaving the source. The random variable X is the amount of energy delivered to the target (which is zero when the particle fails to reach the target). The measure P represents the true probabilities for the scatter of the particle which is assumed to be known

from the theory of physics, and the measure Q represents the probabilities for the scatter of a *simulated* particle in our computer runs.

As an example, suppose we perform importance sampling on the problem in Figure 1.1, and our simulation run has produced the path indicated. Let us say a particle following that path delivered 2 units of energy to the target. Suppose that in choosing our simulation measure Q, we made that particular path 10 times as likely as it should have been under the true probabilities P. We then count that simulation run as having delivered $2 \cdot \frac{1}{10} = 0.2$ units of energy to the target.

Under P, the particle has an extremely small chance of scoring (reaching our target location). The idea of importance sampling is to choose Q in such a way so that the particle is more likely to score. This way, we do not need an extreme number of computer runs before our simulated particle hits the target. We can do this by making collisions less likely to occur under Q, and making absorption less likely when the particle does collide. When choosing the scattering angle, we can give more weight to directions towards the target rather than away from it.

To simply say that we want to change the probabilities so that the particle is more likely to score is very vague. There are many ways this can be done and not all of them will result in a variance reduction. We may decide, for example, to choose Q so that with very high probability the particle goes straight from source to target without a single collision. While this would lead to our simulations producing many scores in a reasonable amount of time, it will not necessarily give us a good estimator. We need to be very careful about just exactly how we make the particle more likely to score.

Ideally, we would like to choose Q to make the variance of our estimator as small as possible. From (1.1) if $Q \gg P$ then

$$Var_Q\left(\frac{dP}{dQ} \cdot X\right) = E^Q\left(\frac{dP}{dQ} \cdot X\right)^2 - \left[E^Q\left(\frac{dP}{dQ} \cdot X\right)\right]^2 = E^Q\left(\frac{dP}{dQ} \cdot X\right)^2 - \left[E^P X\right]^2.$$

If we choose Q by

$$\frac{dQ}{dP} = X \cdot \left[E^P X\right]^{-1} \tag{1.2}$$

then

$$E^Q\left(\frac{dP}{dQ} \cdot X\right)^2 = E^Q\left(X^{-2} \cdot \left[E^P X\right]^2 \cdot X^2\right) = \left[E^P X\right]^2$$

and hence

$$Var_Q\left(\frac{dP}{dQ} \cdot X\right) = 0.$$

It should be pointed out that (1.2) is cheating a little bit since it does not satisfy $Q \gg P$. Note that $Q\{X = 0\} = 0$ while $P\{X = 0\}$ may be positive. In fact, $\{X = 0\}$ represents the neutron not reaching the target which has probability very close to one under P. Nevertheless, it can be shown that (1.1) still holds so long as

$$\frac{dQ}{dP} > 0 \quad \text{on the event} \quad \{X > 0\}.$$

This is clearly satisfied by (1.2).

This choice gives us a perfect estimator. Clearly this is not practical since the formula for Q depends on the unknown quantity $E^P X$. However, this can give us some insight into how we want to go about choosing Q. As an heuristic, we can think of $E^P X$ as a "sum" over all possible paths of the true probability of the path multiplied by the energy that a particle following the path will deliver to the target.

$$\text{expected energy delivered} \approx \sum_{\text{all paths}} P\{path\} \cdot \text{Energy}\{path\}. \tag{1.3}$$

Formula (1.2) essentially says to take dQ proportional to $X \cdot dP$ or equivalently, take

$$Q\{path\} \propto P\{path\} \cdot \text{Energy}\{path\}.$$

We should sample each path in proportion to its contribution to the sum in (1.3), or its **importance**.

Let us consider the path mentioned before where the particle goes straight from source to target without a single collision. A particle following this path does not lose energy in collisions and may deliver a high dose to the detector, but its probability would typically be so small that its term in (1.3) would be negligible. This path would have very low importance and any choice Q which gave it high probability would likely lead to an estimator with very high variance. A particularly poor choice of Q can actually give our estimator infinite variance.

To sucessfully perform importance sampling we need to identify which paths have the high importance and adapt our simulations to favor those paths. For the high dimensional transport problems this is not an easy task. We need to find the optimal tradeoff between paths with few collisions (and hence high energy) and low probability, and paths with more collisions (less energy) but higher probability.

1.3 Summary

In Chapter 2, neutron scatter is modelled as a random walk with killing. While these models are much too simple to cover realistic transport problems, they do provide exellent intuition into the trade-off between paths with many and few collisions. A probability measure Q with the interpretation of "stretching" the distances between collisions is considered. Unlike the measure given by (1.2) this one can be obtained without prior knowledge of $E^P X$. The theory of large deviations is used to prove this choice of Q to be asymptotically optimal among a large class of potential probability measures.

In Chapter 3, more general Markov chain models are considered for neutron scatter. The problem is simplified by considering only finite state spaces. An algorithm is presented which adaptively "learns" which paths are important and iteratively modifies the simulation probabilities in an attempt to converge to the ideal choice of (1.2).

Because, the probabilities are continuously being changed, the algorithm avoids the $n^{-\frac{1}{2}}$ rate of convergence given by iid simulations. The convergence rate is shown to be exponential, paralleling deterministic methods of solution.

Various methods of extending the algorithm to continuous state spaces are discussed in Chapter 4. The algorithm needs to divide the state space into a finite number cells giving it the flavor of deterministic methods. Indeed, the algorithm can be viewed as a compromise between deterministic and Monte Carlo methods. If we try a very fine division of the state space, our algorithm will suffer the same curse of dimensionality of the deterministic methods. However, even a coarse division of the continuous space, which would not suffice for deterministic solutions, may provide a large variance reduction so that simulation can provide good estimates in a reasonable amount of time.

Chapter 2

A Random Walk Model

Instead of a general transport problem, let us consider a simple shielding problem where the particle is penetrating a homogenous solid. For example, suppose a lead shield is placed directly in front of a neutron source. We are interested in the probability that a given particle leaving the source passes completely through the lead before it is absorbed.

A particle penetrates a random distance within the shield before it collides with a molecule. If it is not absorbed, it travels another distance until the next collision. This continues until the particle passes through the shield or is absorbed. Because the shield is homogenous, we would expect the probability law for the inter-collision distance to be independent of the previous collision site. It also seems reasonable that the absorption probability would be independent of collision site. Thus, we could model the path of the particle as an iid random walk with killing.

Let us assume that the shield's height and width are large compared to its depth. We model these two dimensions as being essentially infinite so that we need only keep track of the particle's depth. Note that the particle's depth is itself a one dimensional iid random walk with killing.

In general, suppose we have a one dimensional random walker taking iid steps with

finite mean. After each step he survives with probability $p > 0$, and is killed with probability $1 - p$. We wish to use importance sampling to estimate the probability that the walker crosses some large level $M > 0$, corresponding to the edge of the shield, before he is killed.

To perform importance sampling we need to choose a new probability measure for the step sizes used in the simulation. As we shall see in Section 2.2, there is only one choice that can give us reasonable results for very "large" values of M, corresponding to "rare" events.

Siegmund [21] studies level crossing probabilities for random walks in the context of sequential testing. He introduces the idea of using *exponential twists* of the original distribution in importance sampling. Lehtonen and Nyrhinen [11] study the problem mentioned here for a negative mean random walk without killing ($p = 1$). They arrive at the same optimal distribution as does Siegmund [21], but use a different, asymptotic, notion of optimality and consider a wider class of potential distributions. This chapter extends their results to the case of general $0 < p \leq 1$. Note that due to the killing, this is an interesting problem even when the step sizes have a non-negative mean.

The general setup for a random walk with killing and the use of importance sampling is given in the next section. In Section 2.2 a new probability measure is introduced and is shown to give minimal variance asymptotically as $M \to \infty$. The theory of large deviations is used to obtain a lower bound on the variance over the entire class of eligible probability measures on the real line. The proposed probability measure is shown to be unique in achieving this bound. The method of proof is analogous to that given by Lehtonen and Nyrhinen [11]. An example using a simple model of neutron scatter is provided in Section 2.3 to demonstrate the variance reduction achieved by using the optimal importance measure.

2.1 Preliminaries

Let $\{X_i\}_{i=1}^{\infty}$ be iid random variables with common cdf F having finite mean. They represent the distances travelled by the particle between collisions. Take $S_n = X_1 + \cdots + X_n$. Let $\{Z_n\}_{n=1}^{\infty}$ be iid Bernoulli(p) random variables independent of $\{X_i\}_{i=1}^{\infty}$. Define $T_M = \inf\{n > 0 : S_n \geq M\}$ and $U = \inf\{n > 0 : Z_n = 0\}$. If P is the measure induced on path space, (i.e. the set of sequences $\{X_i, Z_i\}_{i=1}^{\infty}$), then we are interested in $P\{T_M < U\}$. That is, the probability that the random walk crosses M before killing occurs.

To avoid trivial cases assume that F is not a unit mass so that $P\{X_1 = x\} < 1$ for all $x \in \mathcal{R}$. Let $E^P(\cdot)$ denote expectation under P. Let $\mathcal{G}$ denote the collection of all cdf's G on the real line having positive mean such that $G \gg F$. That is, G does not assign probability zero to a set that has positive probability under F. Let G^{∞} be the measure induced on path space by taking $\{X_i\}_{i=1}^{\infty}$ iid with cdf G instead of F. Note that since G has a positive mean $T_M < \infty$ with probability one under G^{∞}. By Wald's likelihood ratio identity (see Siegmund [21] Eq. (33))

$$P\{T_M < U\} = E^P P\{T_M < U \mid T_M\} = E^P p^{T_M} \cdot 1\{T_M < \infty\} =$$

$$E^{G^{\infty}} p^{T_M} \cdot \prod_{i=1}^{T_M} \frac{dF}{dG}(X_i). \tag{2.1}$$

We can perform importance sampling on this problem by choosing $G \in \mathcal{G}$ and generating $X_1, X_2, \ldots$ iid under G until $S_n \geq M$ for some value n. Taking T_M to be this value n, we can estimate $P\{T_M < U\}$ by the random variable

$$Y_{M,G} \stackrel{\text{def}}{=} p^{T_M} \cdot \prod_{i=1}^{T_M} \frac{dF}{dG}(X_i).$$

Equation (2.1) shows that $Y_{M,G}$ is an unbiased estimator for $P\{T_M < U\}$ under G^{∞}. Note that by conditioning on T_M the actual realization of $\{Z_i\}_{i=1}^{\infty}$ becomes irrelevant. This has the effect of removing the killing from the problem, multiplying the "weight" of the neutron by p at each step instead.

We want to select G to minimize $Var(Y_{M,G})$ under G^∞ in some asymptotic sense as $M \to \infty$. Since the mean of $Y_{M,G}$ under G^∞ is the same for all $G \in \mathcal{G}$, this is equivalent to minimizing

$$\eta(M,G) \stackrel{\text{def}}{=} E^{G^\infty}(Y_{M,G}^2).$$

2.2 Optimal Importance Sampling

Let $c(\theta) = log\{E^P e^{\theta X_1}\}$ be the cumulant generating function of F. Define

$$\mathcal{D} = \{\theta : c(\theta) < \infty\}$$

and let $\mathcal{D}^0$ denote its interior. For $\theta \in \mathcal{D}$, let F_θ be the *twisted* cdf given by

$$\frac{dF_\theta}{dF}(x) = e^{\theta x - c(\theta)}.$$

Assume there exists a $w \in \mathcal{D}^0 \cap (0,\infty)$ such that

$$c(w) = log\,\frac{1}{p} \quad \text{and} \quad c'(w) > 0. \tag{2.2}$$

Note that

$$\frac{dF_w}{dF}(x) = pe^{wx}. \tag{2.3}$$

Take Q to be the measure on path space induced by taking $\{X_i\}_{i=1}^\infty$ iid with cdf F_w instead of F. This will turn out to be the optimal measure to use in importance sampling.

The cdf F has mean $c'(0)$ while F_w has mean $c'(w)$. For $\theta \in \mathcal{D}^0$, $c''(\theta)$ is the variance of the cdf F_θ. Because F is not a unit mass neither is F_θ, and so $c''(\theta)$ is strictly positive. It follows that $c'(w) > c'(0)$. Thus, under Q the random walk $\{S_n\}_{n=0}^\infty$ has a larger (positive) drift and should cross the level M in a smaller number of steps. These shorter paths have higher importance due to the killing. The more steps a path takes to cross the level M, the more times the random walk is vulnerable

to killing. If we condition on the random walk under P having crossed a large level M before killing occured, we would therefore expect it to have happened in a relatively small number of large steps. This is exactly what happens under Q.

It will be shown that F_w given by (2.3) represents just the right amount of "stretching" of the inter-collision distances. Any less, and Q would be giving too much weight to paths with many collisions. More stretching would give too much weight to paths with to few collisions. Note by (2.2) the higher the killing, the higher the value of w and hence the more the drift of the random walk gets stretched under Q.

Now $E^Q X_1 = c'(w) > 0$ and by (2.3) $F_w \gg F$. Thus $F_w \in \mathcal{G}$ and we may choose $G = F_w$. In this case $G^\infty = Q$ so (2.1) becomes

$$P\{T_M < U\} = E^Q p^{T_M} \cdot \prod_{i=1}^{T_M} \left(\frac{1}{p} \cdot e^{-wX_i}\right) = E^Q e^{-wS_{T_M}}. \tag{2.4}$$

Theorem 2.1 $\lim_{M\to\infty} \frac{1}{M} \log P\{T_M < U\} = -w.$

Proof: Since $w > 0$ and $S_{T_M} \geq M$, (2.4) implies:

$$\limsup_{M\to\infty} \frac{1}{M} \log P\{T_M < U\} \leq -w. \tag{2.5}$$

Let $c^*(y) = \sup_\theta \theta y - c(\theta)$ be the convex dual of $c(\cdot)$ (see Bucklew [7] or Rockafellar [19]). Let $y > \max(0, E^P X_1)$ and take $N = \left[\frac{M}{y}\right] + 1$ where $[a]$ is the greatest integer $\leq a$.

Then

$$P\{T_M < U\} \geq p^N \cdot P\{S_N \geq M\} \geq p^N \cdot P\{\frac{S_N}{N} \geq y\}$$

and so

$$\frac{1}{M} \log P\{T_M < U\} \geq \frac{N}{M} \log p + \frac{1}{M} \log P\{\frac{S_N}{N} \geq y\} =$$

$$\left(\frac{N}{M}\right) \cdot \log p + \left(\frac{N}{M}\right) \cdot \frac{1}{N} \log P\{\frac{S_N}{N} \geq y\}.$$

By Cramér's theorem (see Bucklew [7], pp. 9-10) we have

$$\liminf_{M\to\infty} \frac{1}{M} P\{T_M < U\} \geq \frac{1}{y} \cdot \log p - \frac{1}{y} \cdot c^*(y). \tag{2.6}$$

Set $y = c'(w)$. Then $\sup_\theta \theta y - c(\theta)$ is achieved at $\theta = w$ so that

$$c^*(y) = wy - c(w) = wy - \log \frac{1}{p} = wy + \log p.$$

Recall $c'(w) > c'(0) = E^P X_1$. By assumption $c'(w) > 0$ so we can plug $y = c'(w)$ into (2.6) to get

$$\liminf_{M\to\infty} \frac{1}{M} \log P\{T_M < U\} \geq \frac{1}{y} \cdot \log p - \frac{1}{y} \cdot [wy + \log p] = -w$$

which together with (2.5) establishes the theorem. ∎

The result of Theorem 2.1 is rather weak. It provides only the exponential rate at which $P\{T_M < U\}$ tends to zero as $M \to \infty$. Indeed, Theorem 2.3 given later in this section is a stronger result. Nevertheless, Theorem 2.1 is good enough to show that F_w has the best asymptotic variance over the class $\mathcal{G}$. The main result of this chapter is given in the next theorem.

Theorem 2.2 *For all cdf's $G \in \mathcal{G}$, $\liminf_{M\to\infty} \frac{1}{M} \log \eta(M, G) \geq -2w$ with equality iff $G = F_w$. Furthermore, $\lim_{M\to\infty} \frac{1}{M} \log \eta(M, F_w) = -2w$.*

Proof: Since $Var(Y_{M,G}) \geq 0$ we have $\eta(M, G) \geq [P\{T_M < U\}]^2$ and hence by Theorem 2.1:

$$\liminf_{M\to\infty} \frac{1}{M} \log \eta(M, G) \geq -2w.$$

Note that when $G = F_w$, $G^\infty = Q$ so that $\eta(M, F_w) = E^Q e^{-2wS_{T_M}} \leq e^{-2wM}$. Thus

$$\lim_{M\to\infty} \frac{1}{M} \log \eta(M, F_w) = -2w.$$

To show uniqueness, suppose we have $G \in \mathcal{G}$ with

$$\liminf_{M\to\infty} \frac{1}{M} \log \eta(M, G) = -2w.$$

Let K be a measure on the real line defined by

$$dK(x) = p^2 \cdot \frac{dF}{dG}(x)\, dF(x) = \left(p \cdot \frac{dF}{dG}(x)\right)^2 dG(x)$$

and note that

$$\eta(M, G) = E^{G^\infty} \prod_{i=1}^{T_M} \left(p \cdot \frac{dF}{dG}(X_i)\right)^2 = \int_{T_M<\infty} dK(x_1) \cdot dK(x_2) \cdots dK(x_{T_M}).$$

Let

$$c_K(\theta) \stackrel{\text{def}}{\equiv} \log \int e^{\theta x} dK(x)$$

and $c_K^*(y) = \sup_\theta \theta y - c_K(\theta)$ be its convex dual. (If $c_K(\cdot)$ is identically $+\infty$, take $c_K^*(\cdot)$ to be identically $-\infty$.)

Since c_K is not necesarily finite anywhere construct a truncated measure as follows. Let $B > 0$ and define L by

$$\frac{dL}{dK}(x) = 1_A(x) \quad \text{where} \quad A = \left\{x : |x| \le B \quad \text{and} \quad \frac{dF}{dG}(x) \le B\right\}.$$

Then we have $c_L(\theta) \stackrel{\text{def}}{\equiv} \log \int_A e^{\theta x} dK(x) \le c_K(\theta)$. Note that $c_L(\cdot)$ is finite everywhere. Take $c_L^*(y) = \sup_\theta \theta y - c_L(\theta)$ and let ζ be the minimal closed interval such that $P\{X_1 \in \zeta\} = 1$. Since F is not a unit mass ζ is not a singleton set and so its interior, ζ^0, is not empty. Let $x > 0$ and take

$$\eta_x(M, G) \stackrel{\text{def}}{\equiv} \int_{\{T_M \le Mx\}} dK(x_1) \cdots dK(x_{T_M}).$$

Note that $\eta_x(M, G) \le \eta(M, G)$. Lehtonen and Nyrhinen [11] prove the following two results (Lemmas 3 & 4, Section 3):

$$\text{If} \quad y \in \zeta^0 \cap \left[\frac{1}{x}, \infty\right) \quad \text{then} \quad \liminf_{M\to\infty} \frac{1}{M} \log \eta_x(M, G) \ge -\frac{1}{y} \cdot c_L^*(y).$$

and

$$\lim_{B\to-\infty} c_B^*(\theta) = c_K^*(\theta).$$

Now for any $y > 0$ we can find an $x > 0$ large enough so that $y \in (\frac{1}{x}, \infty)$. Thus, the two results can be combined to give:

$$\text{If} \quad y \in \zeta^0 \cap (0,\infty) \quad \text{then} \quad \liminf_{M\to-\infty} \frac{1}{M} \log \eta(M,G) \geq -\frac{1}{y} \cdot c_K^*(y). \tag{2.7}$$

By Jensen's inequality

$$c_K(2\theta) = \log \int \left(pe^{-x}\frac{dF}{dG}(x)\right)^2 dG(x) \geq \log\left(\int pe^{-x}\frac{dF}{dG}(x)\, dG(x)\right)^2 = 2c(\theta) + 2\log p \tag{2.8}$$

and hence for all y:

$$c_K^*(y) = \sup_\theta 2\theta y - c_K(2\theta) \leq \sup_\theta 2\theta y - 2c(\theta) - 2\log p = 2c^*(y) - 2\log p. \tag{2.9}$$

By assumption $c'(w) > 0$. Now $c'(w) = \int x\, dF_w(x) = \int x\, pe^{-wx}\, dF(x) \in \zeta^0$. So we can plug $y = c'(w)$ into (2.7) which together with (2.9) gives:

$$\liminf_{M\to-\infty} \frac{1}{M} \log \eta(M,G) \geq -\frac{1}{y} \cdot c_K^*(y) \geq -\frac{1}{y} \cdot (2c^*(y) - 2\log p) = -2w.$$

(recall that $y = c'(w)$ implies $c^*(y) = wy + \log p$).

By supposition $\liminf_{M\to-\infty} \frac{1}{M} \log \eta(M,G) = -2w$ so (2.9) must be an equality when $y = c'(w)$. For $y = c'(w)$, we know that $\sup_\theta \theta y - c(\theta)$ is achieved at $\theta = w$, and hence (2.8) must be an equality when $\theta = w$. But in order for Jensen's inequality in (2.8) to be an equality for $\theta = w$, we must have

$$e^{wx} \cdot \frac{dF}{dG}(x) = \text{constant} \quad G\,a.e. \text{ and hence } F\,a.e. \ (G \gg F).$$

That is, we must have $\frac{dG}{dF}(x) = pe^{-wx}$ $F\,a.e.$ so that $G = F_w$. This establishes uniqueness which completes the proof. ∎

This result clearly establishes F_w as the optimal measure to use in importance sampling for "large" values of M. If G is any other eligible probability measure on the real line then

$$\liminf_{M\to\infty} \frac{1}{M} \log \frac{Var_{G^\infty}(Y_{M,G})}{Var_Q(Y_{M,F_w})} \geq \liminf_{M\to\infty} \frac{1}{M} \log \frac{\eta(M,G)}{\eta(M,F_w)} > 0.$$

The penalty for using anything other than F_w in importance sampling grows exponentially in M.

With a more careful analysis we can describe the relative error when using F_w. If we add the condition that F is non-arithmetic we get a result which is stronger than Theorem 2.1.

Theorem 2.3 *Suppose F is a non-arithmetic distribution. Then there exist finite positive constants C_1 and C_2 such that $P\{T_M < U\} \sim C_1 \cdot e^{-wM}$ and $\eta(M,F_w) \sim C_2 \cdot e^{-2wM}$.*

Proof: From (2.4) we see that

$$P\{T_M < U\} = e^{-wM} \cdot E^Q e^{-w(S_{T_M}-M)}$$

and an analogous calculation shows

$$\eta(M,F_w) = e^{-2wM} \cdot E^Q e^{-2w(S_{T_M}-M)}.$$

Fix $y > 0$ and let $\tau_+ = \inf\{n > 0 : S_n > 0\}$ denote the time of the first ascending ladder. Recall that $E^Q X_1 = c'(w) > 0$ so that $Q\{\tau_+ < \infty\} = 1$ and $E^Q S_{\tau_+} < \infty$. By Corollary 8.33 of Siegmund [22]

$$\lim_{M\to\infty} Q\{S_{T_M} - M > y\} = \frac{\int_y^\infty Q\{S_{\tau_+} > m\}\, dm}{E^Q S_{\tau_+}}.$$

Note that $\lim_{M\to\infty} Q\{S_{T_M} - M > y\}$ is a proper survival function in y. That is, under Q the

random variables $S_{T_M} - M$ converge in distribution as $M \to \infty$. By the bounded convergence theorem there exist positive constants (less than one) C_1 and C_2 such that

$$\lim_{M\to\infty} E^Q e^{-w(S_{T_M}-M)} = C_1 \quad \text{and} \quad \lim_{M\to\infty} E^Q e^{-2w(S_{T_M}-M)} = C_2.$$

The theorem follows. ∎

If we take the relative error of our estimator to be the ratio of the standard deviation to the probability being estimated then Theorem 2.3 shows us that

$$\lim_{M\to\infty} \frac{[Var_Q(Y_{M,F_w})]^{\frac{1}{2}}}{P\{T_M < U\}} = \frac{[C_2 - C_1^2]^{\frac{1}{2}}}{C_1}.$$

By Theorem 2.2 the relative error grows exponentially in M for any other measure G. Thus, F_w is the only choice where the standard deviation of our estimator tends to zero as quickly as $P\{T_M < U\}$ does when $M \to \infty$. No other cdf $G \in \mathcal{G}$ can give reasonable estimators for "large" values of M. Similar results follow when F is arithmetic.

Note we can interpret F_w as the asymptotic distribution of the step sizes X_i conditional on $\{T_M < U\}$. Fix values $x_1, \ldots, x_n$. Then for "large" values of M we have the following heuristic:

$$P\{X_1 \in dx_1, \ldots, X_n \in dx_n \quad \text{and} \quad T_M < U\} \approx$$

$$P\{X_1 \in dx_1, \ldots, X_n \in dx_n, U \geq n\} \cdot P\{T_M < U \mid X_1 \in dx_1, \ldots, X_n \in dx_n, U \geq n\} \approx$$

$$\prod_{i=1}^{n} p \cdot P\{X_i \in dx_i\} \cdot C_1 \cdot e^{-w(M-\sum x_i)}.$$

Dividing both sides by $P\{T_M < U\} \approx C_1 \cdot e^{-wM}$ gives:

$$P\{X_1 \in dx_1, \ldots, X_n \in dx_n \mid T_M < U\} \approx$$

$$\prod_{i=1}^{n} pe^{-wx_i} \cdot P\{X_i \in dx_i\} = Q\{X_1 \in dx_1, \ldots, X_n \in dx_n\}.$$

Conditional on $\{T_M < U\}$, the path of $\{X_i\}$ looks like an iid random walk under Q. We are essentially generating the sample paths under their distribution conditional on the event we wish to estimate.

2.3 An Example

Consider the following two dimensional model of neutron scatter. Figure 2.1 shows a semi-infinite solid occupying the region $\{(x, y) : y \leq M\}$ in the x-y plane. A neutron collides with a molecule at the origin. Upon collision it is absorbed with probability $\frac{1}{2}$. If the particle is not absorbed, it scatters in a random uniform $(0, 2\pi)$ direction and travels a random exponential (1) distance until the next collision. This continues until the neutron is either absorbed or passes through the solid.

We want to use simulation to estimate the probability of a sucessful penetration. That is, the probability the neutron crosses the line $y = M$ before it is absorbed. This is similar to the model problem discussed by Murthy and Indira [15]. In their problem there are only two possible scattering directions, "forward" and "backward". Here the scattering is taken uniformly over all possible directions in the x-y plane.

Note that the distance travelled in the y direction between collisions is given by the random variable

$$X = D \cdot \sin o$$

where D is an exponential (1) random distance between collisions, and o is a uniform $(0, 2\pi)$ angle independent of D.

Thus, the total penetration in the y direction is itself a one dimensional random walk with iid increments. We want to know the probability that the random walk will cross the level M before killing (absorption) occurs.

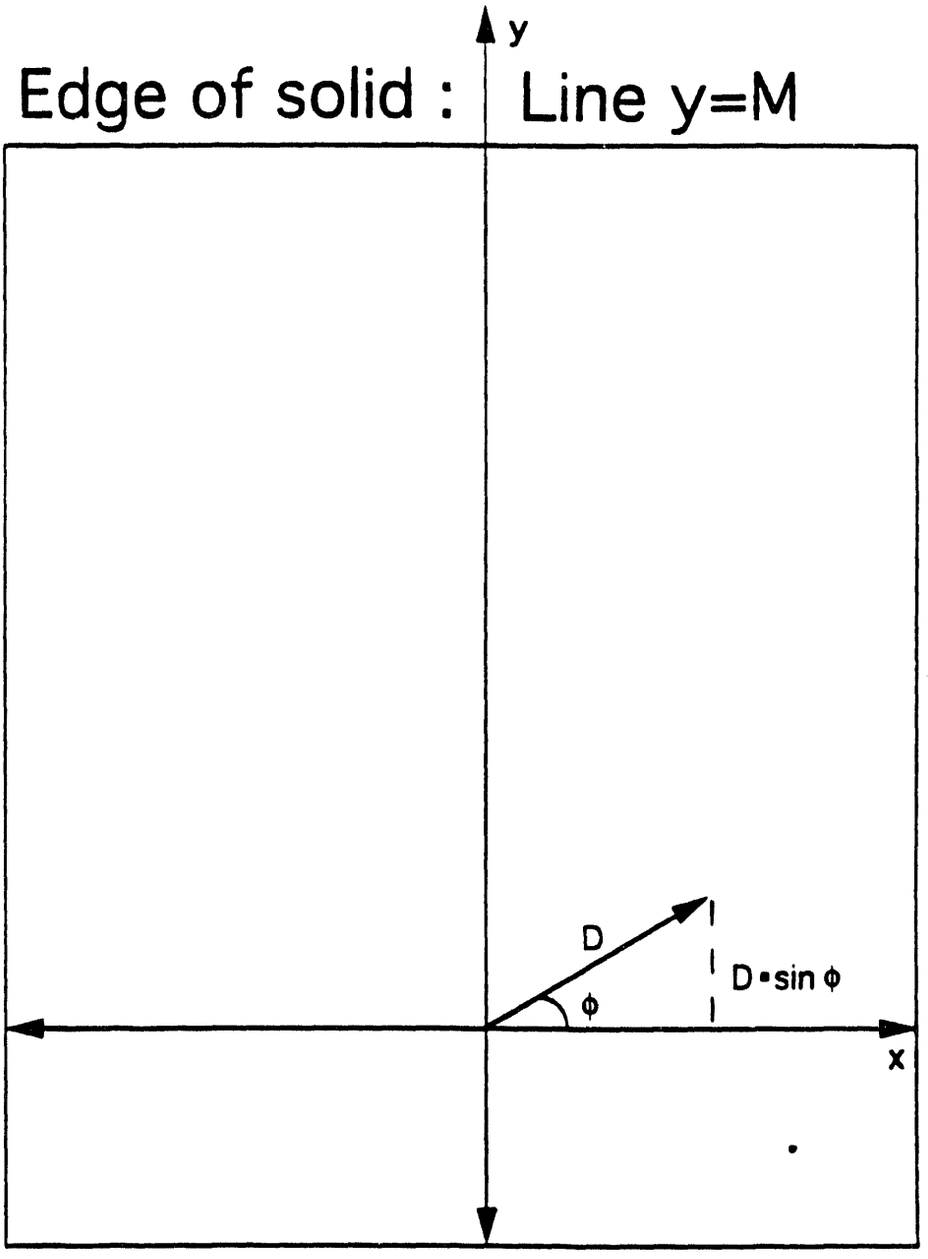


Figure 2.1: A two dimensional random walk model of neutron scatter. The distance, D, between collisions is exponential (1), and the scattering angle, ϕ, is uniform $(0, 2\pi)$. Upon collision, the particle is absorbed with probability 0.5. The distance traveled in the y direction between collisions is given by $D \cdot \sin\phi$.

Now for $x > 0$

$$P\{X > x\} = \frac{1}{2\pi}\int_0^{\pi} \exp\{-\frac{x}{\sin\phi}\}\, d\phi = \frac{1}{\pi}\int_0^{\pi/2} \exp\{-\frac{x}{\sin\phi}\}\, d\phi.$$

Since the probability law of X is symmetric about 0, we see that X has density

$$\frac{1}{\pi}\int_0^{\pi/2} \frac{1}{\sin\phi}\cdot\exp\{-\frac{|x|}{\sin\phi}\}\, d\phi \qquad \text{for} \quad x \in \mathcal{R}. \tag{2.10}$$

Taking $p = \frac{1}{2}$, equation (2.2) becomes

$$\log\left(\frac{1}{\pi}\int_{-\infty}^{\infty} e^{wx}\cdot\int_0^{\pi/2} \frac{1}{\sin\phi}\cdot\exp\{-\frac{|x|}{\sin\phi}\}\, d\phi\, dx\right) = \log 2.$$

Exponentiating both sides, switching the order of integration, and evaluating the integral with respect to x gives the equation

$$\frac{1}{\pi}\int_0^{\pi/2} \frac{1}{1+w\sin\phi} - \frac{1}{1-w\sin\phi}\, d\phi = 2$$

which may be solved numerically to give $w \approx 0.86602540378$.

By (2.3) X has the *twisted* density

$$\frac{e^{wx}}{2\pi}\cdot\int_0^{\pi/2} \frac{1}{\sin\phi}\cdot\exp\{-\frac{|x|}{\sin\phi}\}\, d\phi \qquad \text{for} \quad x \in \mathcal{R} \tag{2.11}$$

under the Q measure.

To perform the importance sampling, we generate $X_1, X_2, \ldots$ iid with the *twisted* density (2.11) until $S_n = X_1 + X_2 + \cdots + X_n \geq M$ for some n. Taking T_M to be this value n, we estimate our probability of complete penetration, $P\{T_M < \infty\}$, by

$$Y_{M,F_w} = e^{-wS_{T_M}}.$$

Note that $E^Q X_1 = c'(w) > 0$ for this choice of w and so with probability one there will eventually be an n such that $S_n \geq M$.

Random variates with the density given in (2.11) were obtained using the acceptance-rejection method (see Algorithm 3.4 of Ripley [18]) with a mixture of the true density

M	estimate for $P\{T_M < t\}$	relative error	error reduction
20	5.72 E-09	0.045	9.28 E−03
40	1.58 E-16	0.046	5.44 E−07
60	4.99 E-24	0.046	3.06 E−11
80	1.67 E-31	0.043	1.82 E−15

Table 2.1: Simulation results. 1000 runs using importance sampling with F_w.

2.10) and an exponential with mean $(.13)^{-1}$. One thousand runs were made at various levels of M: the results are given in Table 2.1. Relative error is the ratio of the standard error of the estimate to the estimate. The final column compares the standard errors of ordinary simulation (binomial) and our importance sampling. Note that as Theorem 2.2 would predict, the error reduction appears to be growing exponentially in M. As Theorem 2.3 would predict, the relative error does not seem to vary with M.

It should be pointed out that the acceptance-rejection algorithm needs to evaluate the integral (2.11) to produce the random variates. Simpson's rule was used with 10^3 intervals. So generating a random variate with density (2.11) is on the order of 10^3 times as much work as generating a random variate with the true density (2.10). From Table 1 we can see that for $M \geq 20$ the error reduction more than makes up for the extra work.

2.4 Adding Energy to the Model

While the random walk model of this chapter is admittedly simplistic, we can add some realism by considering the particle's energy level. That is, instead of merely calculating the probability of penetration, we can calculate the expected energy of a particle after having penetrated the shield. The energy of a particle that gets

absorbed (killed) is counted as zero.

We now interpret the quantity p as the expected percentage of the particle's energy it retains after a collision. Thus, p is a measure of elasticity for the collisions between the neutron and the shielding medium. Here, instead of being a Bernoulli random variable, Z_n is the percentage of energy retained in the n^{th} collision. Note that $\{Z_n = 0\}$ still represents absorption after the n^{th} collision. So instead of $P\{T_M < U\}$, we are now interested in

$$E^P \prod_{i=1}^{T_M} Z_i$$

which represents the energy of a particle after having penetrated the shield. If we still assume that $\{Z_i\}_{i=1}^{\infty}$ is independent of $\{X_i\}_{i=1}^{\infty}$ then equation (2.1) holds with $P\{T_M < U\}$ replaced by $E^P \prod_{i=1}^{T_M} Z_i$. The same optimality results go through when we choose w by (2.2).

The choice of the stretching parameter, w, can now be interpreted as a trade-off between the high energy low probability paths with few collisions, and the lower energy higher probability paths with many collisions. For the case considered in this chapter, iid inter-collision distances, and iid energy losses, the optimal importance scheme may be calculated exactly. This does not achieve the theoretical ideal of (1.2), but is the best that can be done if we restrict $\{X_i\}_{i=1}^{\infty}$ to be iid under Q. More general models are considered in the next chapter.

Chapter 3

Markov Chain Models

In realistic transport problems, there are many different types of solid barriers, the probability of absorption and the distribution for the scattering angle depend on the position and energy of the particle, and we need to keep track of more than just the particle's position in one dimension. The path of the particle, recorded at its collision sites, is modelled as a discrete time Markov chain. As mentioned in Chapter 1, the state space must contain sufficient information on the position, velocity, and energy of the particle. We perform importance sampling on this problem by generating sample paths according to a new set of transition probabilities and multiplying the result by the likelihood ratio between the true and simulated probabilities.

If we choose a good set of new transition probabilities, or importance scheme, we can get a substantial reduction in variance. In fact, it is well known that zero variance importance schemes exist for linear Monte Carlo transport problems (see Booth [6]). Unfortunately, the zero variance scheme depends on the expected score from each state, which if known would make Monte Carlo simulation unnecessary.

We can take a "guess" at the expected scores to compute the importance scheme. If our guess is close to the true expectation we hope that our importance scheme has close to zero variance. Simulation output can be used to update our "guess" at the

expected scores and hopefully produce a better importance scheme. This suggests a learning technique where the importance scheme is continually being updated by the results of the simulation. Such learning techniques have been studied in the past (see Booth [2], [4], & [5], and Troubetzkoy [23]). Because the importance scheme is continually being updated, (hopefully improved) we may expect to obtain better than the $n^{-\frac{1}{2}}$ rate of convergence given by iid simulation.

The behavior of such an adaptive learning algorithm on a finite state problem is analyzed in this chapter. Booth [2] gives empirical results for a simple two state problem which suggest that the convergence is exponential. Here a proof is given that the convergence is indeed exponential for a finite state problem.

Halton [10] describes two similar algorithms. In one algorithm, the importance scheme is chosen by the user and remains constant throughout the algorithm. He proves that this algorithm coverges exponentially fast, but the particular rate of exponential convergence depends on the importance scheme selected. The second algorithm can be made to converge as fast as an arbitrary polynomial in the number of iterations, but the amount of work per iteration is growing as a polynomial. For the algorithm considered here, the user supplies an initial "guess" at the expected scores from which the first importance scheme is computed. The amount of work required per iteration is constant.

In Section 3.2 a general finite state problem is described. Section 3.3 describes importance sampling for these problems and derives the variance associated with using an importance scheme. The algorithm is presented in Section 3.4 and a proof of its exponential rate of convergence is given in Section 3.5.

3.1 The Model

Consider a particle on a finite state Markov chain $\{X_n\}_{n=0}^{\infty}$ with transition matrix $\mathbf{P}$. At each state the particle may be subject to absorption. We assume that eventual absorption is certain (i.e. $\lim_{n\to\infty} \mathbf{P}^n = \mathbf{0}$). When the particle changes state from i to j a nonnegative score s_{ij} is incurred. Typically these scores are zero unless state j corresponds to our target location. In this case, s_{ij} is intended to denote the energy delivered to that location by the particle. We allow an arbitrary nonnegative scoring rule for generality. Let Δ denote the "cemetary", or absorption state. When the particle is absorbed from state i a nonnegative score $s_{i\Delta}$ is incurred. If we take τ to be the time of absorption then a particle's total score is given by

$$Y = \sum_{n=1}^{\tau} s_{X_{n-1},X_n}.$$

So for example, if a particle's path is: $1 \to 2 \to 1 \to \Delta$ the total score would be $s_{12} + s_{21} + s_{1\Delta}$.

We are interested in estimating $\mu_i \stackrel{\text{def}}{=} E[Y \mid X_0 = i]$, that is, the expected total score for a particle starting in state i. Let d be the number of states and take $p_{i\Delta} = 1 - \sum_{j=1}^{d} p_{ij}$ to be the probability of absorption directly out of state i. If we condition on X_1, the first transition, we see that

$$\mu_i = p_{i\Delta} \cdot s_{i\Delta} + \sum_{j=1}^{d} p_{ij} \cdot (s_{ij} + \mu_j). \tag{3.1}$$

In matrix form, this can be written as $\boldsymbol{\mu} = \mathbf{a} + \mathbf{P}\boldsymbol{\mu}$ where $a_i = p_{i\Delta} \cdot s_{i\Delta} + \sum_{j=1}^{d} p_{ij} \cdot s_{ij}$.

Since $\lim_{n\to\infty} \mathbf{P}^n = \mathbf{0}$, $\mathbf{I} - \mathbf{P}$ is invertible and the system of d equations in d unknowns may be solved exactly. Alternatively, the deterministic recursion

$$\boldsymbol{\mu}^{(m+1)} = \mathbf{a} + \mathbf{P}\boldsymbol{\mu}^{(m)}$$

will converge to the true value $\boldsymbol{\mu}$ exponentially fast for any starting point $\boldsymbol{\mu}^{(0)}$. This

suggests the existence of an analogous Monte Carlo procedure which also converges exponentially fast.

3.2 Importance Sampling on Markov Chains

Suppose that instead of simulating the particle under the true probabilites $\{p_{ij}\}$ we choose new probabilities $\{q_{ij}\}$. Each score is weighted by the likelihood ratio between **P** and **Q**. On the event $\{\tau \geq n\}$ set

$$L_n = \prod_{j=1}^{n} \frac{p_{X_{j-1},X_j}}{q_{X_{j-1},X_j}}.$$

We must choose $\mathbf{Q} \gg \mathbf{P}$ (ie $q_{ij} > 0$ whenever $p_{ij} > 0$, and $q_{i\Delta} > 0$ whenever $p_{i\Delta} > 0$) so that L_n is well defined. Then our estimator becomes

$$Y_Q = \sum_{n=1}^{\tau} \left[s_{X_{n-1},X_n} \cdot L_n \right].$$

It is easily shown that Y_Q has expectation μ_i, where $i = X_0$ is the starting point of the simulation.

Let $v_i = Var[Y_Q \mid X_0 = i]$. We can derive v_i by conditioning on X_1, the first transistion. Note that on the event $\{X_1 = j\}$, we have $Y_Q = \frac{p_{ij}}{q_{ij}} \cdot (s_{ij} + Y_Q^*)$, where Y_Q^* is the total weighted score of a particle starting in state j. By the Markov property

$$E(Y_Q \mid X_1 = j, X_0 = i) = \frac{p_{ij}}{q_{ij}} \cdot (s_{ij} + \mu_j) \quad \text{and} \quad Var(Y_Q \mid X_1 = j, X_0 = i) = \left(\frac{p_{ij}}{q_{ij}}\right)^2 \cdot v_j.$$

The variance decomposition $Var(Y_Q|X_0) = Var[E(Y_Q|X_1, X_0)] + E[Var(Y_Q|X_1, X_0)]$ gives us:

$$v_i = q_{i\Delta} \cdot \left(\frac{p_{i\Delta} \cdot s_{i\Delta}}{q_{i\Delta}}\right)^2 + \sum_j \left[q_{ij} \cdot \left(\frac{p_{ij}}{q_{ij}}\right)^2 \cdot (s_{ij} + \mu_j)^2 \right] - \mu_i^2 + \sum_j \left[q_{ij} \cdot \left(\frac{p_{ij}}{q_{ij}}\right)^2 \cdot v_j \right] =$$

$$\frac{p_{i\Delta}^2 \cdot s_{i\Delta}^2}{q_{i\Delta}} + \sum_j \left[\frac{p_{ij}^2 \cdot (s_{ij} + \mu_j)^2}{q_{ij}}\right] - \mu_i^2 + \sum_j \frac{p_{ij}^2}{q_{ij}} \cdot \nu_{jj}$$

Here we take $0/0 = 0$.

Let

$$f_i = \frac{p_{i\Delta}^2 \cdot s_{i\Delta}^2}{q_{i\Delta}} + \sum_j \left[\frac{p_{ij}^2 \cdot (s_{ij} + \mu_j)^2}{q_{ij}}\right] - \mu_i^2$$

take 0/0=0). Define $\mathbf{R}$ to be the matrix whose $(i,j)^{th}$ element is given by

$$r_{ij} = \begin{cases} \frac{p_{ij}^2}{q_{ij}} & \text{if} \quad q_{ij} > 0; \\ \\ 0 & \text{if} \quad q_{ij} = 0. \end{cases}$$

Then in matrix form our equation becomes:

$$\mathbf{v} = \mathbf{f} + \mathbf{R}\mathbf{v}. \tag{3.2}$$

We now show that $\mathbf{v}$ is the minimal solution to (3.2):

Theorem 3.1 $\mathbf{v} = \sum_{n=0}^{\infty} \mathbf{R}^n \mathbf{f}$.

Proof: Note that any $\mathbf{v}$ satisfying (3.2) also satisfies

$$\mathbf{v} = \sum_{j=0}^{n-1} \mathbf{R}^j \mathbf{f} + \mathbf{R}^n \mathbf{v}$$

so that $\hat{\mathbf{v}} = \sum_{n=0}^{\infty} \mathbf{R}^n \mathbf{f}$ is the minimal solution to (3.2).

Take

$$Y_Q^{(n)} = \sum_{j=1}^{\tau \wedge n} s_{X_{j-1},X_j} \cdot L_j + \left[\mu_{X_n} \cdot L_n \cdot 1\{\tau > n\}\right].$$

Note that $Y_Q^{(n)}$ has expectation μ_i under $\mathbf{Q}$, but is unusable in practice since it depends on the unknown vector μ. Let $v_i^{(n)} = Var(Y_Q^{(n)} | X_0 = i)$. Then using an analogous argument as before we see that

$$\mathbf{v}^{(n)} = \mathbf{f} + \mathbf{R}\mathbf{v}^{(n-1)} \qquad n = 1, 2, \ldots.$$

By induction

$$\mathbf{v}^{(n)} = \sum_{j=0}^{n-1} \mathbf{R}^j \mathbf{f}.$$

Now $Y_Q = \lim_{n\to\infty} Y_Q^{(n)}$ so by Fatou's Lemma $\mathbf{v} \leq \sum_{n=0}^{\infty} \mathbf{R}^n \mathbf{f}$. Since $\hat{\mathbf{v}} = \sum_{n=0}^{\infty} \mathbf{R}^n \mathbf{f}$ is the minimal solution to (3.2), the proof is complete. ∎

Note that when we choose

$$q_{ij} = \frac{p_{ij} \cdot (s_{ij} - \mu_j)}{p_{i\Delta} \cdot s_{i\Delta} - \sum_{l=1}^{d} p_{il} \cdot (s_{il} - \mu_l)} \quad \text{and} \quad q_{i\Delta} = \frac{p_{i\Delta} \cdot s_{i\Delta}}{p_{i\Delta} \cdot s_{i\Delta} + \sum_{l=1}^{d} p_{il} \cdot (s_{il} + \mu_l)} \tag{3.3}$$

we get

$$f_i = \left(\frac{p_{i\Delta}^2 \cdot s_{i\Delta}^2}{p_{i\Delta} \cdot s_{i\Delta}} + \sum_{j=1}^{d} \frac{p_{ij}^2 \cdot (s_{ij} - \mu_j)^2}{p_{ij} \cdot (s_{ij} + \mu_j)}\right) \cdot \left(p_{i\Delta} \cdot s_{i\Delta} + \sum_{l=1}^{d} p_{il} \cdot (s_{il} - \mu_l)\right) - \mu_i^2 =$$

$$\left(p_{i\Delta} \cdot s_{i\Delta} + \sum_{j=1}^{d} p_{ij} \cdot (s_{ij} + \mu_j)\right) \cdot \left(p_{i\Delta} \cdot s_{i\Delta} + \sum_{l=1}^{d} p_{il} \cdot (s_{il} + \mu_l)\right) - \mu_i^2 = 0 \quad \text{by (3.1).}$$

This choice of $\mathbf{Q}$ gives $\mathbf{f} = \mathbf{0}$ and by Theorem 3.1 $\mathbf{v} = \mathbf{0}$. We have a zero variance importance scheme. However, this choice of $\mathbf{Q}$ depends on the unknown vector $\boldsymbol{\mu}$.

3.3 A Learning Algorithm

For a vector $\hat{\boldsymbol{\mu}}$ in $\mathcal{R}^d$ we parallel (3.3) by taking

$$q_{ij}(\hat{\boldsymbol{\mu}}) \stackrel{\text{def}}{=} \frac{p_{ij} \cdot (s_{ij} + \hat{\mu}_j)}{p_{i\Delta} \cdot s_{i\Delta} + \sum_{l=1}^{d} p_{il} \cdot (s_{il} + \hat{\mu}_l)} \quad \text{and} \quad q_{i\Delta}(\hat{\boldsymbol{\mu}}) \stackrel{\text{def}}{=} \frac{p_{i\Delta} \cdot s_{i\Delta}}{p_{i\Delta} \cdot s_{i\Delta} + \sum_{l=1}^{d} p_{il} \cdot (s_{il} + \hat{\mu}_l)}.$$

Take $\mathbf{Q}(\hat{\boldsymbol{\mu}})$ to be the matrix whose $(i,j)^{th}$ element is $q_{ij}(\hat{\boldsymbol{\mu}})$.

The idea of the algorithm is to simulate under $\mathbf{Q}(\hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{\mu}}$ represents our best "guess" at $\boldsymbol{\mu}$. We will take $\hat{\boldsymbol{\mu}}$ to be the estimate obtained from previous simulations, but we must be careful to ensure that $\mathbf{Q} \gg \mathbf{P}$. If $s_{ij} = 0$ and we have $\hat{\mu}_j = 0$ then $q_{ij}(\hat{\boldsymbol{\mu}}) = 0$ even though p_{ij} may be positive.

Note that if there exists a $\delta > 0$ such that $s_i \geq \delta$ for all i then $\mu_i \geq \delta$ for all i. The particle is assured of scoring at least δ since it will eventually be absorbed. Since we know that $\mu_i \geq \delta$ we can take the maximum of δ and our estimate without increasing the error. This ensures that $\hat{\mu}_i \geq \delta$ and hence $\mathbf{Q}(\boldsymbol{\mu}) \gg \mathbf{P}$ throughout the algorithm.

If no such δ exists we can easily alter the problem by adding a $\delta > 0$ to each s_i. Since every particle gets absorbed exactly once, this just adds δ to each μ_i. We can subtract δ from each $\hat{\mu}_i$ to shift back to the original problem. Without loss of generality then, there exists a known $\delta > 0$ such that $\mu_i \geq \delta$ for all i.

Algorithm 3.1

Choose an integer k and start with an initial guess $\hat{\boldsymbol{\mu}}^{(0)} > \delta \mathbf{1}$. The algorithm is defined inductively:

1. Suppose after m iterations the algorithm has produced an estimate $\hat{\boldsymbol{\mu}}^{(m)}$.

2. For each state $i = 1, 2, \ldots, d$, run k independent simulations starting the particle in state i, using $\mathbf{Q}(\hat{\boldsymbol{\mu}}^{(m)})$ as the transistion matrix.

3. Let τ_{mzi} be the absorption time for the z^{th} simulation ($z = 1, \ldots, k$) starting from state i. Take

$$Y_{mzi} = \sum_{n=1}^{\tau_{mzi}} s_{X_{n-1}, X_n} \cdot L_n$$

where the $\{X_n\}$ and $\{L_n\}$ are obtained from the z^{th} simulation starting from state i (this dependence on z is suppressed to keep the notation managable). Let

$$\bar{Y}_{mki} = \frac{1}{k} \cdot \sum_{z=1}^{k} Y_{mzi}.$$

Here, the notation is chosen to emphasize that the sample mean is based on k simulation runs.

4. Define $\hat{\mu}^{(m+1)}$ by:

$$\hat{\mu}_i^{(m+1)} = max\left(\bar{Y}_{mki}, c\right) \quad i = 1, 2, \ldots, d. \qquad \blacksquare$$

Remarks: 1. Since the probability law of $\hat{\mu}^{(m+1)}$ depends only on $\hat{\mu}^{(m)}$, the sequence $\{\hat{\mu}^{(m)}\}_{m=0}^{\infty}$ is itself a Markov process on $\mathcal{R}^d$. It clearly has the state μ as an absorbing state. Thus, a unit mass at 0 is an invariant measure for this Markov chain. Transition probabilities typically converge exponentially fast to the invariant measure in Markov chains. If it can be shown this is the unique invariant measure we may, therefore, expect $\{\hat{\mu}^{(m)}\}$ to converge exponentially fast to μ. A proof of this is given in the next section.

2. By construction, $\hat{\mu}^{(m)} \geq \delta \mathbf{1}$ and hence $\mathbf{Q}(\hat{\mu}^{(m)}) \gg \mathbf{P}$ for all m.

3. By taking the maximum of our estimate and δ we introduce a bias. However, since we have set up our problem to ensure $\mu_i \geq \delta$ we know with probability one that $|\hat{\mu}_i^{(m+1)} - \mu_i| \leq |\bar{Y}_{mki} - \mu_i|$.

4. This algorithm wastes information by not using paths starting from state i to help estimate μ_j for other states j. For example, suppose starting from state 1 we get the path $1 \rightarrow 2 \rightarrow \Delta$. Then in addition to providing information about μ_1 we can use the path from the point it hit state 2 (i.e. $2 \rightarrow \Delta$) to help estimate μ_2. The strong Markov property says this has the same probability distribution as a particle started in state 2. The proof given in the next section would also cover this improved version which uses such information. The critical point is that *at least* k independent paths starting from state i are used in the estimator $\hat{\mu}_i^{(m+1)}$.

3.4 Exponential Convergence

We have made two assumptions about the problem:

Assumption 3.1. $\lim_{n\to\infty} \mathbf{P}^n = 0$.

Assumption 3.2 There exists a $\delta > 0$ such that $s_i \geq \delta$ for every i (see discussion in Section 3.3).

The main result of this chapter is stated in the next theorem:

Theorem 3.2 *Under the assumptions stated above there exists deterministic constants $\theta > 0$ and K such that if Algorithm 3.1 is run with $k \geq K$ then with probability one $e^{\theta m} \cdot \|\hat{\mu}^{(m)} - \mu\| \to 0$ as $m \to \infty$.*

Before giving the proof we need to derive some preliminary results. The critical point when studying this algorithm is how quickly the variance tends to zero as we approach the perfect importance scheme. Lemma 3.1 shows that the variance is locally bounded by a quadratic in the distance from the true mean μ.

Notation: In Section 3.2 we defined the quantities **R**, **f**, and **v** depending on our choice of **Q**. Now that our **Q** is a function of $\hat{\mu}$, we can define these quantities as functions of $\hat{\mu}$. Take $\mathbf{R}(\hat{\mu})$ to be the matrix whose $(i,j)^{th}$ element is given by

$$r_{ij}(\hat{\mu}) = \begin{cases} \frac{p_{ij}^2}{q_{ij}(\hat{\mu})} & \text{if} \quad q_{ij}(\hat{\mu}) > 0. \\ \\ 0 & \text{if} \quad q_{ij}(\hat{\mu}) = 0. \end{cases}$$

and $\mathbf{f}(\hat{\mu})$ to be the vector whose i^{th} element is

$$f_i(\hat{\mu}) = \frac{p_{i\Delta}^2 \cdot s_{i\Delta}^2}{q_{i\Delta}(\hat{\mu})} + \sum_j \left[\frac{p_{ij}^2 \cdot (s_{ij} + \mu_j)^2}{q_{ij}(\hat{\mu})} \right] - \mu_i^2$$

(take $0/0 = 0$). Let $v_i(\hat{\mu})$ be the variance of Y_Q starting in state i when simulating under $\mathbf{Q}(\hat{\mu})$. Take $\mathbf{v}(\hat{\mu})$ to be the vector whose i^{th} element is $v_i(\hat{\mu})$. Then by Theorem 3.1 of Section 3.2

$$\mathbf{v}(\hat{\mu}) = \sum_{n=0}^{\infty} \mathbf{R}^n(\hat{\mu})\mathbf{f}(\hat{\mu}).$$

Lemma 3.1 *There exists a matrix* $\mathbf{A}$ *and an* $\epsilon > 0$ *such that* $\mathbf{1}^{\mathrm{T}}\mathbf{v}(\hat{\mu}) < (\hat{\mu} - \mu)^{\mathrm{T}}\mathbf{A}(\hat{\mu} - \mu)$ *whenever* $\|\hat{\mu} - \mu\| < \epsilon$.

Proof: By Assumption 3.2 we know $\mu \geq \delta \mathbf{1}$ and so $\mathbf{Q}(\mu) \gg \mathbf{P}$. If $p_{ij} > 0$ then $q_{ij}(\mu) > 0$ and thus

$$r_{ij}(\mu) = \frac{p_{ij}^2}{q_{ij}(\mu)} = \frac{p_{ij}^2 \cdot [p_{i\Delta} \cdot s_{i\Delta} + \sum_{l=1}^{d} p_{il} \cdot (s_{il} + \mu_l)]}{p_{ij} \cdot (s_{ij} + \mu_j)} = \frac{p_{ij} \cdot \mu_i}{s_{ij} + \mu_j} \leq \frac{p_{ij} \cdot \mu_i}{\mu_j}$$

(when $p_{ij} = 0$ the inequality between the first and last terms still holds since they are both zero). An induction argument shows that for all $n = 1, 2, \ldots$

$$r_{ij}^{(n)}(\mu) \leq \frac{p_{ij}^{(n)} \cdot \mu_i}{\mu_j}$$

where $r_{ij}^{(n)}(\mu)$ and $p_{ij}^{(n)}$ are the $(i,j)^{th}$ elements of $\mathbf{R}^n(\mu)$ and $\mathbf{P}^n$ respectively. From Assumption 3.1 we know

$$\sum_{n=0}^{\infty} \mathbf{P}^n < \infty \quad \text{and hence} \quad \sum_{n=0}^{\infty} \mathbf{R}^n(\mu) < \infty.$$

Thus

$$(\mathbf{I} - \mathbf{R}(\mu))^{-1} = \sum_{n=0}^{\infty} \mathbf{R}^n(\mu).$$

Now on the open region in which $(\mathbf{I} - \mathbf{R})^{-1}$ is defined, each of its elements is a convergent power series in each r_{ij}. By Theorem 8.1 of Rudin [20] the elements of $(\mathbf{I} - \mathbf{R})^{-1}$ are infinitely differentiable functions of r_{ij} for each fixed pair (i,j). When

all the elements of $\hat{\mu}$ are positive $\mathbf{Q}(\hat{\mu}) \gg \mathbf{P}$ and so both $r_{ij}(\hat{\mu})$ and $f_i(\hat{\mu})$ are infinitely differentiable in $\hat{\mu}$. By Theorem 3.1

$$\mathbf{v}(\hat{\mu}) = \sum_{n=0}^{\infty} \mathbf{R}^n(\hat{\mu})\,\mathbf{f}(\hat{\mu}) = (\mathbf{I} - \mathbf{R}(\hat{\mu}))^{-1}\,\mathbf{f}(\hat{\mu})$$

for $\hat{\mu}$ sufficiently close to μ so that $\mathbf{I} - \mathbf{R}(\hat{\mu})$ is invertible. Thus, $\mathbf{v}(\cdot)$ is infinitely differentiable at μ. In Section 3.2 we showed that $\mathbf{f}(\mu) = 0$ and hence $\mathbf{v}(\mu) = \mathbf{0}$. Thus, μ is a global minimum for $\mathbf{v}(\cdot)$, and so

$$\frac{\partial v_i}{\partial \hat{\mu}_l}(\mu) = 0 \quad \text{for all } i \text{ and } l.$$

So, $\mathbf{1}^T\mathbf{v}(\hat{\mu})$ is a function from $\mathcal{R}^4 \rightarrow \mathcal{R}$ that is infinitely differentiable at μ with $\mathbf{1}^T\mathbf{v}(\mu) = 0$, and $\nabla[\mathbf{1}^T\mathbf{v}](\mu) = \mathbf{0}$. The lemma follows by Taylor's theorem. ∎

This bound on the variance is good enough to show that exponential convergence is at least *possible* provided we start the algorithm close enough to the true mean. Recall the initial guess, $\hat{\mu}^{(0)}$, is supplied by the user and is considered deterministic in our analysis.

Lemma 3.2 *There exists a constant $c \in (0,1)$, an integer K_1, an $\epsilon > 0$, and a $\beta > 0$ such that if Algorithm 3.1 is run with $\|\hat{\mu}^{(0)} - \mu\| < \epsilon$ and $k \geq K_1$ then*
$Pr\{\|\hat{\mu}^{(m)} - \mu\| \leq c^m \cdot \|\hat{\mu}^{(0)} - \mu\|$ for all $m\} \geq \beta$.

Proof: By Assumption 3.2 we know $\mu_i \geq \delta$ for all i. Using step 4 of the algorithm we have with probability one

$$\|\hat{\mu}_i^{(m+1)} - \mu_i\|^2 \leq \|\bar{Y}_{mki} - \mu_i\|^2.$$

It follows that

$$E\left(\|\hat{\mu}_i^{(m+1)} - \mu_i\|^2 \mid \hat{\mu}^{(m)}\right) \leq E\left(\|\bar{Y}_{mki} - \mu_i \mid \mu^{(m)}\|^2 = \frac{r_i(\hat{\mu}^{(m)})}{k}\right.$$

Let ϵ and $\mathbf{A}$ be as in Lemma 3.1. Then on the event $\{\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \epsilon\}$ we have

$$E(\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 \mid \hat{\boldsymbol{\mu}}^{(m)}) = \sum_{i=1}^{k} E(\|\hat{\mu}_i^{(m+1)} - \mu_i\|^2 \mid \hat{\boldsymbol{\mu}}^{(m)}) \leq$$

$$\frac{(\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu})^T \mathbf{A} (\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu})}{k}.$$

Let

$$b \stackrel{\text{def}}{=} \frac{1}{k} \cdot \sup_{\|\phi\|=1} \|\mathbf{A}\phi\|.$$

Choose an integer K_1 large enough so that $k \geq K_1$ implies $b < 1$. Then on the event $\{\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \epsilon\}$

$$E(\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 \mid \hat{\boldsymbol{\mu}}^{(m)}) \leq b \cdot \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2. \tag{3.4}$$

We would like (3.4) to hold with probability one, not just on the event $\{\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \epsilon\}$. Construct a *coupled* process as follows: Let $T = \inf\{m : \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| \geq \epsilon\}$ and take

$$\boldsymbol{\lambda}^{(m)} \stackrel{\text{def}}{=} \begin{cases} \hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu} & \text{if} \quad m \leq T. \\ \mathbf{0} & \text{if} \quad m > T. \end{cases}$$

By forcing the process $\{\boldsymbol{\lambda}^{(m)}\}$ to be zero after $\{\hat{\boldsymbol{\mu}}^{(m)}\}$ leaves the ϵ-neighborhood of $\boldsymbol{\mu}$, we impose the analog of (3.4) to hold with probability one for the process $\{\boldsymbol{\lambda}^{(m)}\}$. By construction

$$\{\boldsymbol{\lambda}^{(m+1)} = \mathbf{0}\} \supseteq \{\boldsymbol{\lambda}^{(m)} = \mathbf{0}\} \cup \{\|\boldsymbol{\lambda}^{(m)}\| \geq \epsilon\}$$

and

$$\{0 < \|\boldsymbol{\lambda}^{(m)}\| < \epsilon\} \subseteq \{T \geq m+1\} \subseteq \{\boldsymbol{\lambda}^{(m)} = \hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\} \cap \{\boldsymbol{\lambda}^{(m+1)} = \hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\}.$$

Thus,

$$\boldsymbol{\lambda}^{(m+1)} = (\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}) \cdot 1\{0 < \|\boldsymbol{\lambda}^{(m)}\| < \epsilon\}$$

Together with (3.4) these give:

$$E\left(\|\boldsymbol{\lambda}^{(m+1)}\|^2 \mid \boldsymbol{\lambda}^{(m)}\right) = E\left(\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 \cdot 1\{0 < \|\boldsymbol{\lambda}^{(m)}\| < \epsilon\} \mid \boldsymbol{\lambda}^{(m)}\right) =$$

$$E\left(E\left(\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 \mid \{\hat{\boldsymbol{\mu}}^{(n)}\}_{n=0}^{m}\right) \mid \boldsymbol{\lambda}^{(m)}\right) \cdot 1\{0 < \|\boldsymbol{\lambda}^{(m)}\| < \epsilon\} =$$

$$E\left(E\left(\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 \mid \hat{\boldsymbol{\mu}}^{(m)}\right) \mid \boldsymbol{\lambda}^{(m)}\right) \cdot 1\{0 < \|\boldsymbol{\lambda}^{(m)}\| < \epsilon\} \leq$$

$$E\left(b \cdot \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2 \cdot 1\{0 < \|\boldsymbol{\lambda}^{(m)}\| < \epsilon\} \mid \boldsymbol{\lambda}^{(m)}\right) \leq b \cdot \|\boldsymbol{\lambda}^{(m)}\|^2. \tag{3.5}$$

By induction we have

$$E\|\boldsymbol{\lambda}^{(m)}\|^2 \leq b^m \cdot \|\boldsymbol{\lambda}^{(0)}\|^2. \tag{3.6}$$

(note that $\boldsymbol{\lambda}^{(0)} = \hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}$ is deterministic.)

Choose c so that $b < c^2 < 1$, and note that $c < 1$. Define events

$$F_m = \{\|\boldsymbol{\lambda}^{(m)}\| \leq c^m \cdot \|\boldsymbol{\lambda}^{(0)}\|\} \qquad m = 1, 2, \ldots .$$

We want to show that the values $Pr\{F_m^c \mid F_{m-1}\}$ are small enough so that the event $\bigcap_{m=1}^{\infty} F_m$ has positive probability.

By (3.6) and Markov's inequality

$$Pr\{F_m\} \geq 1 - \frac{b^m \cdot \|\boldsymbol{\lambda}^{(0)}\|^2}{c^{2m} \cdot \|\boldsymbol{\lambda}^{(0)}\|^2} = 1 - \left(\frac{b}{c^2}\right)^m. \tag{3.7}$$

By Markov's inequality, (3.6), and (3.7) for $m > 1$

$$Pr\{F_m^c \mid F_{m-1}\} \leq \frac{E\left(\|\boldsymbol{\lambda}^{(m)}\|^2 \mid F_{m-1}\right)}{c^{2m} \cdot \|\boldsymbol{\lambda}^{(0)}\|^2} = \frac{E\left(\|\boldsymbol{\lambda}^{(m)}\|^2 \cdot 1_{F_{m-1}}\right)}{c^{2m} \cdot \|\boldsymbol{\lambda}^{(0)}\|^2 \cdot \Pr\{F_{m-1}\}} \leq$$

$$\frac{E\|\boldsymbol{\lambda}^{(m)}\|^2}{c^{2m} \cdot \|\boldsymbol{\lambda}^{(0)}\|^2 \cdot Pr\{F_{m-1}\}} \leq \left(\frac{b}{c^2}\right)^m \cdot \left[1 - \left(\frac{b}{c^2}\right)^{m-1}\right]^{-1}. \tag{3.8}$$

Take

$$\gamma_m \stackrel{\text{def}}{=} 2 \cdot \left(\frac{b}{c^2}\right)^m.$$

We can choose M large enough so that for $m > M$

$$1 - \left(\frac{b}{c^2}\right)^{m-1} > \frac{1}{2}$$

and

$$\sum_{m=M+1}^{\infty} \gamma_m < 1.$$

By (3.8)

$$P\{F_m^c \mid F_{m-1}\} \leq \gamma_m \quad \text{for} \quad m > M. \tag{3.9}$$

By (3.5) and the conditional Markov inequality

$$P\{F_m^c \mid F_{m-1}\} \leq \frac{b \cdot \|\lambda^{(m)}\|^2}{c^2 \cdot \|\lambda^{(m)}\|^2} = \frac{b}{c^2} \quad \text{for all} \quad m. \tag{3.10}$$

Take

$$\beta \stackrel{\text{def}}{=} \left(\frac{b}{c^2}\right)^M \cdot \prod_{m=M+1}^{\infty} (1 - \gamma_m).$$

Note β does not depend on our choice of $\hat{\mu}^{(0)}$. By the relation $\prod(1-\gamma_m) \geq 1 - \sum \gamma_m$, and our choices of c and M, we have $\beta > 0$.

The sequence $\{\lambda^{(m)}\}_{m=0}^{\infty}$ is a stopped Markov process and is therefore Markov. By (3.9) and (3.10) and the Markov property

$$Pr\{\bigcap_{m=1}^{\infty} F_m\} = \prod_{m=1}^{\infty} Pr\{F_m \mid \bigcap_{n=1}^{m-1} F_n\} = \prod_{m=1}^{\infty} Pr\{F_m \mid F_{m-1}\} \geq \beta.$$

If $\|\hat{\mu}^{(0)} - \mu\| < \epsilon$ then $\{\|\lambda^{(T)}\| \geq \epsilon\} \subseteq \{T < \infty\}$, and hence

$$\{\bigcap_{m=1}^{\infty} F_m\} \supseteq \{T = \infty\} \supseteq \{\lambda^{(m)} = \hat{\mu}^{(m)} - \mu \quad \text{for all m}\}.$$

That is, these processes never de-couple on this event. So

$$Pr\{\|\hat{\mu}^{(m)} - \mu\| \leq c^m \cdot \|\hat{\mu}^{(0)} - \mu\| \text{ for all } m\} \geq Pr\{\bigcap_{m=1}^{\infty} F_m\} \geq \beta. \quad \blacksquare$$

We must start our initial guess, $\hat{\mu}^{(0)}$, close enough to the true μ for the probability bound of Lemma 3.2 to hold. However, even if we start with $\|\hat{\mu}^{(0)} - \mu\| \geq \epsilon$ we can wait to see if $\|\hat{\mu}^{(m^*)} - \mu\| < \epsilon$ for some m^*. If this happens the strong Markov property tells us there is probability at least β that

$$\|\hat{\mu}^{(m)} - \mu\| < c^{m-m^*} \cdot \|\hat{\mu}^{(m^*)} - \mu\| \quad \text{for all} \quad m \geq m^*.$$

That is, every time the process $\{\hat{\mu}^{(m)}\}_{m=0}^{\infty}$ enters the ϵ-neighborhood of μ, there is probability at least β for exponential convergence. If we can show this must happen infinitely often then exponential convergence would be certain. We are now ready to complete the proof.

Proof of Theorem 3.2: Choose a real number H large enough so that

$$\frac{\max_i \mu_i - \delta}{H} < \frac{1}{2d}$$

By step 4 of the algorithm

$$0 < \hat{\mu}_i^{(m+1)} \leq \bar{Y}_{mki} - \delta.$$

Recall that $\bar{Y}_{mki}$ is an unbiased estimator of μ_i conditional on $\hat{\mu}^{(m)}$ so that

$$E(\hat{\mu}_i^{(m+1)} \mid \hat{\mu}^{(m)}) < \mu_i - \delta.$$

By Markov's inequality

$$Pr\{\hat{\mu}_i^{(m+1)} > H \mid \hat{\mu}^{(m)}\} \leq \frac{1}{2d}.$$

A union bound gives us

$$Pr\{\hat{\mu}^{(m+1)} \leq H\,\mathbf{1} \mid \hat{\mu}^{(m)}\} \geq \frac{1}{2}. \tag{3.11}$$

Let $\mathcal{U}$ denote the set of vectors in $\mathcal{R}^d$ having all positive components. For a vector $\hat{\mu}$ in $\mathcal{U}$ let $\mathcal{L}_i(\cdot \mid \hat{\mu})$ denote the probability law of Y_{mzi} when simulating under $\mathbf{Q}(\hat{\mu})$. That is, for measurable $A \subseteq \mathcal{R}$

$$\mathcal{L}_i(A \mid \hat{\mu}) = Pr\{Y_{mzi} \in A \mid \hat{\mu}^{(m)} = \hat{\mu}\}.$$

Note that the right hand side does not depend on m or z. Recall that the transition probabilities $q_{ij}(\hat{\mu})$ are continuous functions of $\hat{\mu}$. For $\hat{\mu} \in \mathcal{U}$, $\mathbf{Q}(\hat{\mu}) \gg \mathbf{P}$ so that the

likelihood ratios $\{L_n\}$ and estimators $\{Y_{mzi}\}$ are well defined continuous functions of $\hat{\mu}$. It follows that the probability measures $\mathcal{L}_i(\cdot \mid \hat{\mu})$ are continuous in the sense that for any convergent sequence of vectors in $\mathcal{U}$ the corresponding probability measures converge in distribution:

$$\text{If} \quad \hat{\mu}_n \to \hat{\mu}, \quad \text{then} \quad \mathcal{L}_i(\cdot \mid \hat{\mu}_n) \stackrel{\mathcal{D}}{\Rightarrow} \mathcal{L}_i(\cdot \mid \hat{\mu}).$$

Let $\alpha > 0$ and take $E^{\hat{\mu}}(\cdot)$ to denote expectation under $\mathcal{L}_i(\cdot \mid \hat{\mu})$. Suppose we have a sequence of vectors in $\mathcal{U}$ with $\hat{\mu}_n \to \hat{\mu}$. Then the the probability measures for the random variable $Y_{mzi} \cdot 1\{Y_{mzi} \le \alpha\}$ under $\mathbf{Q}(\hat{\mu}_n)$ converge in distribution to the probability measure of $Y_{mzi} \cdot 1\{Y_{mzi} \le \alpha\}$ under $\mathbf{Q}(\hat{\mu})$. By the bounded convergence theorem

$$E^{\hat{\mu}_n}(Y_{mzi} \cdot 1\{Y_{mzi} \le \alpha\}) \to E^{\hat{\mu}}(Y_{mzi} \cdot 1\{Y_{mzi} \le \alpha\}).$$

That is,

$$E^{\hat{\mu}}(Y_{mzi} \cdot 1\{Y_{mzi} \le \alpha\}) \quad \text{is a continuous function of } \hat{\mu} \text{ for each fixed } \alpha.$$

For $\hat{\mu} \in \mathcal{U}$

$$E^{\hat{\mu}}(Y_{mzi}) = \mu_i \tag{3.12}$$

which is a constant and hence continuous function of $\hat{\mu}$. So, on $\mathcal{U}$

$$E^{\hat{\mu}}(Y_{mzi} \cdot 1\{Y_{mzi} > \alpha\}) = E^{\hat{\mu}}(Y_{mzi}) - E^{\hat{\mu}}(Y_{mzi} \cdot 1\{Y_{mzi} \le \alpha\}) \quad \text{is continuous in } \hat{\mu}.$$

Since Y_{mzi} has finite mean

$$\lim_{\alpha \to \infty} E^{\hat{\mu}}(Y_{mzi} \cdot 1\{Y_{mzi} > \alpha\}) = 0.$$

If we restrict $\hat{\mu}$ to $\{\hat{\mu} : \delta \mathbf{1} \le \hat{\mu} \le H \mathbf{1}\}$ then we can think of $E^{\hat{\mu}}(Y_{mzi} \cdot 1\{Y_{mzi} > \alpha\})$ as a family of continuous functions of $\hat{\mu}$ on a compact set indexed by α. These functions tend monotonically to zero pointwise as $\alpha \to \infty$, so by Theorem 7.13 of Rudin [20]

the convergence is uniform. That is.

$$\lim_{\alpha \to \infty} \sup_{\delta \mathbf{1} \le \hat{\mu} \le H\mathbf{1}} E^{\hat{\mu}}(Y_{mi} \cdot 1\{Y_{mi} > \alpha\}) = 0 \tag{3.13}$$

so that the family of probability measures

$$\{\mathcal{L}_i(\cdot \mid \hat{\mu}) : \delta \mathbf{1} \le \hat{\mu} \le H\mathbf{1}\}$$

is uniformly integrable. By (10.2) of Parzen [17] this implies that the weak law of large numbers holds uniformly over $\{\hat{\mu} : \delta \mathbf{1} \le \hat{\mu} \le H\mathbf{1}\}$. Let ϵ be as in Lemma 3.2. Then

$$\lim_{\alpha \to \infty} \sup_{\delta \mathbf{1} \le \hat{\mu} \le H\mathbf{1}} Pr\{|\bar{Y}_{mki} - \mu_i| > \frac{\epsilon}{\sqrt{d}} \mid \hat{\mu}^{(m)} = \hat{\mu}\} = 0. \tag{3.14}$$

Note this is where the assumption $\mu_i > \delta$ for all i is critical. Equation (3.12) does not necessarily hold unless all components of $\hat{\mu}$ are positive. We must have a compact set in order for (3.13) to hold and so we must bound $\hat{\mu}$ away from $\mathbf{0}$.

By (3.14) we can choose K_2 large enough so that $k \ge K_2$ implies

$$\sup_{\delta \mathbf{1} \le \hat{\mu} \le H\mathbf{1}} Pr\{|\bar{Y}_{mki} - \mu_i| > \frac{\epsilon}{\sqrt{d}} \mid \hat{\mu}^{(m)} = \hat{\mu}\} < \frac{1}{2d}.$$

If the algorithm is run with $k \ge K_2$ then a union bound gives us

$$Pr\{\|\hat{\mu}^{(m+1)} - \mu\| < \epsilon \mid \hat{\mu}^{(m)}\} \ge \frac{1}{2} \tag{3.15}$$

on the event $\{\delta \mathbf{1} \le \hat{\mu}^{(m)} \le H\mathbf{1}\}$.

Let K_1, c, and β be as in Lemma 3.2. Let $K = max(K_1, K_2)$. Suppose the algorithm is run with $k \ge K$. Now $\{\hat{\mu}^{(m)}\}_{m=0}^{\infty}$ is Markov so (3.11) and (3.15) imply

$$Pr\{\|\hat{\mu}^{(m+2)} - \mu\| < \epsilon \mid \hat{\mu}^{(m)}\} \ge \frac{1}{4}.$$

Hence by the conditional Borel-Cantelli lemma (see Section 12.15 of Williams [25])

$$Pr\{\|\hat{\mu}^{(m)} - \mu\| < \epsilon \quad i.o.\} = 1. \tag{3.16}$$

Now define two sequences of stopping times $\{U_n\}$ and $\{W_n\}$ inductively as follows

$$W_0 = 0, \quad U_{n+1} = \inf\{m > W_n : \|\hat{\mu}^{(m)} - \mu\| < \epsilon\}, \quad \text{and}$$

$$W_n = \inf\{m > U_n : \|\hat{\mu}^{(m)} - \mu\| > c^{m-U_n} \cdot \|\hat{\mu}^{(U_n)} - \mu\|\}.$$

$\{W_n\}$ just marks times at which exponential convergence fails and $\{U_n\}$ marks the next time after failure that $\{\hat{\mu}^{(m)}\}$ enters the ϵ-neighborhood of μ. By Lemma 3.2 and the strong Markov property

$$Pr\{W_n = \infty \mid U_n < \infty\} \geq \beta. \tag{3.17}$$

By (3.16)

$$Pr\{U_n < \infty \mid W_{n-1} < \infty\} = 1. \tag{3.18}$$

Let

$$G_n = \{W_{n-1} < \infty \quad \text{and} \quad W_n = \infty\}.$$

Note that

$$\bigcap_{i=1}^{n-1} G_i^c = \bigcap_{i=1}^{n-1} \{W_i < \infty\}$$

and so by the Markov property, (3.17), and (3.18)

$$Pr\{G_n \mid \bigcap_{i=1}^{n-1} G_i^c\} = Pr\{G_n \mid W_{n-1} < \infty\} \geq$$

$$Pr\{U_n < \infty \mid W_{n-1} < \infty\} \cdot Pr\{W_n = \infty \mid U_n < \infty\} = 1 \cdot \beta = \beta.$$

It follows that

$$Pr\{\bigcup_{n=1}^{\infty} G_n\} = 1.$$

Now

$$G_n = \{W_{n-1} < \infty \quad \text{and} \quad U_n = \infty\} \cup \{U_n < \infty \quad \text{and} \quad W_n = \infty\}.$$

By (3.18) the event on the left has probability zero and so

$$Pr\Big\{\bigcup_{n=1}^{\infty}\{U_n < \infty \quad \text{and} \quad W_n = \infty\}\Big\} = 1. \tag{3.19}$$

Recall from Lemma 3.2 that $c < 1$ and hence $-log_e(c) > 0$. Choose $0 < \theta < -log_e(c)$ and note

$$\{e^{m\theta} \cdot \|\hat{\mu}^{(m)} - \mu\| \to 0\} \supset \{U_n < \infty \quad \text{and} \quad W_n = \infty\} \quad \text{for all } n.$$

Thus by (3.19)

$$Pr\{e^{m\theta} \cdot \|\hat{\mu}^{(m)} - \mu\| \to 0\} \geq Pr\Big\{\bigcup_{n=1}^{\infty}\{U_n < \infty \quad \text{and} \quad W_n = \infty\}\Big\} = 1$$

which is the statement of the theorem. ∎

3.5 Using Previous Information

The algorithm described in Section 3.3 uses information from previous iterations only to compute the matrix $\mathbf{Q}(\hat{\mu}^{(m)})$. Once this matrix is used to run the simulations, $\hat{\mu}^{(m)}$ is no longer considered in our estimation. The estimator $\hat{\mu}^{(m+1)}$ is based solely on the results of the most recent simulation runs. It seems more efficient to base our estimator on all available information and take $\hat{\mu}^{(m+1)}$ to be some weighted average of the most recent simulation runs and $\hat{\mu}^{(m)}$. That is, suppose we choose some $0 < \alpha \leq 1$ and replace step 4 of the algorithm by

4*. Define $\hat{\mu}^{(m+1)}$ by

$$\hat{\mu}_i^{(m+1)} = \alpha \cdot \max\Big(\frac{1}{k} \cdot \sum_{z=1}^{k} Y_{mzi}, \delta\Big) + (1-\alpha) \cdot \hat{\mu}_i^{(m)} \qquad i = 1, 2, \ldots, d.$$

Note that the sequence $\{\hat{\mu}^{(m)}\}$ is still Markov. In fact, if we look closely at the arguments given for the original algorithm, we see that they all apply to this version of the algorithm except the ones establishing equations (3.4) and (3.16). To derive (3.4), just note that on the event $\{\|\hat{\mu}^{(m)} - \mu\| < \epsilon\}$ we have

$$E\left(\|\hat{\mu}^{(m+1)} - \mu\|^2 \mid \hat{\mu}^{(m)}\right) \leq \alpha^2 \cdot \sum_{i=1}^{d} E\left(\left(\frac{1}{k} \cdot \sum_{z=1}^{k} Y_{mzi} - \mu_i\right)^2 \mid \hat{\mu}^{(m)}\right) + (1-\alpha)^2 \cdot \|\hat{\mu}^{(m)} - \mu\|^2 \leq$$

$$\left(\alpha^2 b + (1-\alpha)^2\right) \cdot \|\hat{\mu}^{(m)} - \mu\|^2.$$

Thus, the analog of (3.4) holds for the new algorithm if we replace b by

$$b^* = \alpha^2 b + (1-\alpha)^2.$$

Note that $b < 1$ and $0 < \alpha \leq 1$ implies that $b^* < 1$.

To establish (3.16) we note that (3.14) implies that we may choose k large enough so that

$$\sup_{1 \leq \hat{\mu} \leq H1} Pr\{|\bar{Y}_{mki} - \mu_i| > \frac{\epsilon}{2\sqrt{d}} \cdot \hat{\mu}^{(m)} = \hat{\mu}\} < \frac{1}{2d}.$$

Now since $\alpha > 0$ we can choose an integer N large enough so that $(1-\alpha)^N < \frac{\epsilon}{2H}$. Let

$$E_m = \{(|\bar{Y}_{mki} - \mu_i| > \frac{\epsilon}{2\sqrt{d}} \quad \text{for all} \quad i = 1, 2, \ldots, d\}.$$

On the event $\{\hat{\mu}^{(m)} \leq H\mathbf{1}\}$ a union bound gives us

$$Pr\{E_{m+1} \mid \hat{\mu}^{(m)}\} \geq \frac{1}{2}. \tag{3.20}$$

Note that by step 4* of the modified algorithm

$$\hat{\mu}_i^{(m+N)} = (1-\alpha)^N \cdot \hat{\mu}_i^{(m)} + \alpha \cdot \sum_{n=1}^{N} (1-\alpha)^{N-n} \cdot \bar{Y}_{m+n,k,i}.$$

By choice of N, if $\hat{\mu}^{(m)} \leq H\mathbf{1}$ then $(1-\alpha)^N \cdot \hat{\mu}_i < \frac{\epsilon}{2}$ for all i. If, in addition, $\bar{Y}_{m+n,k,i} - \mu_i| < \frac{\epsilon}{2\sqrt{d}} \quad n = 1, \ldots, N$ then

$$|\hat{\mu}_i - \mu| < \frac{\epsilon}{\sqrt{d}}.$$

It follows that

$$\{\|\hat{\mu}^{(m+N)} - \mu\| < \epsilon\} \supset \{\hat{\mu}^{(m)} \leq H\mathbf{1}\} \cap \bigcap_{n=1}^{N} E_{m+n}.$$

By the Markov property, (3.11), and (3.20)

$$Pr\{\|\hat{\mu}^{(m+N+1)} - \mu\| < \epsilon \mid \hat{\mu}^{(m)}\} \geq \left(\frac{1}{2}\right)^{N+1}.$$

Equation (3.16) now follows by conditional Borel-Cantelli (Section 12.15 Williams [25]).

Ideally, we would like to choose α to minimize the variance of the resulting estimator. Recall that in the proof of Theorem 3.2 there exists a deterministic constant, $b < 1$, such that with probability one

$$Var\left(\sum_{i=1}^{k} \bar{Y}_{mki} \mid \hat{\mu}^{(m)}\right) \leq b \cdot \|\hat{\mu}^{(m)} - \mu\|^2.$$

This suggests the heuristic:

$$Var(\bar{Y}_{mki}) \approx b \cdot Var(\hat{\mu}_i^{(m)}).$$

If this were the case, then to give each estimator a weight inversely proportional to its variance we should take

$$\alpha = \frac{1}{1+b}.$$

However, b is the norm of a $d \times d$ matrix, and would be too much work to compute for large d. Even if b were known, it is only a theoretical bound. As we shall see in the next section, empirical evidence suggests this variance bound is not tight. Nevertheless, when we perform the algorithm, we may estimate the conditional variance, $Var(\bar{Y}_{mki} \mid \hat{\mu}^{(m)})$, from our k data points at each iteration. If this quantity appears to be decreasing exponentially at a constant rate, as in our heuristic, we may empirically estimate the ratio, $\hat{b}$, of variances on sucessive iterations and choose

$$\alpha = \frac{1}{1+\hat{b}}.$$

Of course no proof has been given that the rate is constant, but for the problems condsidered in the next section that does appear to be the case.

3.6 Sample Problems

Now that the theory for the algorithm has been presented, two examples are given here. Empirical results are given to confirm the predicted exponential rate of

convergence.

One consideration when running Algorithm 3.1 are the choices of k and α. As in the previous section we can get some heuristic ideas for how we should select these quantities, but they require knowledge of the exponential rate at which our estimators are converging.

Suppose we run some trial iterations using the suggestion in Remark 4 of Section 3.3, but without using information from previous iterations as discussed in Section 3.5. For these initial runs we choose k_0 arbitrarily adjusting up if necessary until the process appears to be converging. Based on our empirical standard deviations at each iteration, we can estimate the ratio, $\hat{b}$, of conditional variances on consecutive iterations. Now $\hat{w} = k_0 \cdot \hat{b}$ represents what would be the ratio of variances on succesive iterations if we had chosen $k = 1$. Note that $\hat{w}$ is not necessarily less than one since $k = 1$ may not be large enough to force convergence.

If we choose k simulations per iteration we estimate that we would see the conditional variances decrease by a factor $\frac{\hat{w}}{k}$ at each iteration. Once we make that choice of k, we would employ the heuristic of the previous section and take

$$\alpha = \frac{1}{1 - \left(\frac{\hat{w}}{k}\right)} = \frac{k}{k + \hat{w}} \tag{3.21}$$

This should reduce the ratio of consecutive variances to

$$\left(\frac{k}{k+\hat{w}}\right)^2 \cdot \frac{\hat{w}}{k} - \left(\frac{\hat{w}}{k+\hat{w}}\right)^2 = \frac{\hat{w}}{k+\hat{w}}. \tag{3.22}$$

Suppose we budget T units of computer time for our simulations. Let r denote the ratio of time spent updating the matrix $\mathbf{Q}(\hat{\boldsymbol{\mu}}^{(m)})$ to the time required to run one simulation starting from each state. (This can be estimated empirically and will tend to a constant as $\hat{\boldsymbol{\mu}}^{(m)}$ tends to $\boldsymbol{\mu}$.) Now (3.22) is clearly minimized by taking k arbitrarily large, but the price you pay for large k is that you only have time to run

$$\frac{T}{r+k}$$

number of iterations. Thus, we should choose k to minimize

$$\left(\frac{\dot{w}}{k+\dot{w}}\right)^{\frac{\dot{}}{r+\kappa}} = \left[\left(\frac{\dot{w}}{k+\dot{w}}\right)^{\frac{\dot{}}{r+\kappa}}\right]^{T}.$$

Note the minimizing value does not depend on T. We choose k to minimize

$$\left(\frac{\dot{w}}{k-\dot{w}}\right)^{\frac{\dot{}}{r+\kappa}} \tag{3.23}$$

which can be done numerically.

Of course there is no guarantee that the resulting value of k will actually be large enough to force the algorithm to converge. Recall that there has been no proof that the rate of convergence is constant. Even if this were always true, our trial iterations may have given us a poor estimate of the rate. This is merely a simple heuristic attempt to make reasonable choices of k and α.

In the second example, a continuous space transport problem is approximated by a finite state problem.

3.6.1 A four state problem

Let us start off by looking at a simple four state problem whose solution is apparent. Take the transistion matrix to be

$$\mathbf{P} = \begin{pmatrix} 0.93 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.93 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.93 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.93 \end{pmatrix}$$

and the score matrix

$$\mathbf{S} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1000 \\ 1 & 1 & 1 & 1 & 1000 \\ 1 & 1 & 1 & 1 & 1000 \\ 1 & 1 & 1 & 1 & 1000 \end{pmatrix}$$

where the final column coresponds to the absorption score $s_{i\Delta}$.

Note that the problem is symmetric in the four states. The rows of S are identical so that the transition score does not depend on the state of the particle. There is probability 0.01 of absorption from any of the four states. Thus, the particle will remain in the system for a geometric (.99) number of transitions and then get absorbed. The particle gets a score of 1 for each transition within the system and a score of 1000 when it is absorbed. We expect the particle to make 99 transitions before being absorbed and so the particle has a total expected score of 1099. That is, by inspection we see this problem has solution

$$\mu = (1099, 1099, 1099, 1099)^T.$$

The simplest version of this algorithm, without using the suggestion in Remark 4 of Section 3.3 or the modification in Section 3.5, was run on this problem with an initial guess,

$$\hat{\mu}^{(0)} = (1, 1, 1, 1)^T.$$

Ignoring for the moment optimal choice of sample sizes, $k = 15$ simulations were run per iteration. The results given in Table 3.1.

Estimates are given for all four states by iteration. The value in the standard deviation column is the empirical standard deviation of the 15 estimates for μ_1. This value divided by the square root of 15, therefore, gives the estimated standard error for $\hat{\mu}_1^{(m)}$ conditional on the previous estimate.

Since these k data points are iid conditional on the previous estimate, we may use the central limit theorem to obtain confidence intervals on μ_1. Note that we can not trust our estimate of the error in the first three iterations of the algorithm. The true mean is more than 30 standard errors away from our estimate on the first iteration. Here our initial guess $\hat{\mu}^{(0)}$ gives us a very poor choice $\mathbf{Q}(\hat{\mu}^{(0)})$ for use in importance sampling. As a result, our estimator has a very skewed distribution and $k = 15$ runs

iteration	expected score	standard deviation
1	55.782953	1.182E+02
	20.749360	
	29.518720	
	29.518720	
2	273.641157	2.365E+02
	174.246171	
	116.641568	
	144.616742	
3	748.456441	9.280E+02
	862.519766	
	452.722070	
	410.659111	
	.	
	.	
	.	
50	1098.999991	1.861E-04
	1099.000092	
	1099.000044	
	1098.999968	
	.	
	.	
	.	
80	1099.000000	6.895E-08
	1099.000000	
	1099.000000	
	1099.000000	

Table 3.1: Empirical results from a simple four state problem

is not good enough for the central limit theorem to take effect. As our estimate, $\hat{\mu}^{(m)}$, gets closer to the true μ, the conditional distribution of the subsequent estimator becomes less skewed and the resulting confidence intervals become more reliable. By iteration number 50, our standard error gives a reasonable estimate for the actual sampling error.

Theorem 3.2 asserts that the error should be decreasing exponentially by iteration. If we plot the logarithm of the absolute value of the actual error against iteration number we would expect it to decrease linearly. Figure 3.1 shows $\log_{10}(|\hat{\mu}_1^{(m)} - \mu_1|)$ as a function of m. As expected, the relationship appears linear. Note that Theorem 3.2 gives only an upper bound, there is no guarantee that the rate of decrease will be constant although that appears to be the case here. The dashed line represents the upper 95% confidence limit based on the empirical standard errors and the central limit Theorem. As mentioned before, these limits can not be trusted for the first few iterations, but between the fifth and eightieth iteration the predicted upper limit fails to bound the actual error only twice.

Algorithm 3.1 was replicated 1000 times to check the actual coverage rates of the confidence intervals. The results by iteration are shown in Figure 3.2. Taking our estimate plus or minus twice the estimated standard error should cover the true mean roughly 95% of the time. We see that after the 25^{th} iteration the coverage rates are reasonably close to the desired level.

From the plot, we can estimate the slope of the line by a least squares regression. If we ignore the first few iterations the slope is approximately -0.13. This translates into an estimate of $\hat{b}$ by

$$\hat{b} = 10^{(2 \cdot -0.13)} \approx 0.55.$$

We can compare this to the theoretical value of b as defined in the proof of Theorem 3.2. Recall that Lemma 3.1 stated the existence of a matrix $\mathbf{A}$ such that

$$\mathbf{1}^T \mathbf{v}(\hat{\mu}) < (\hat{\mu} - \mu)^T \mathbf{A} (\hat{\mu} - \mu)$$

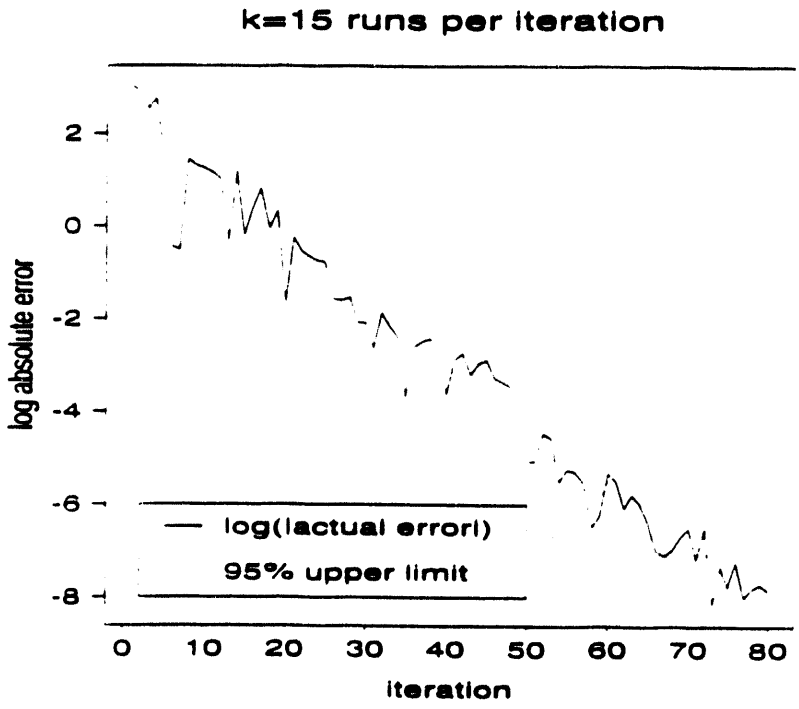


Figure 3.1: Log error decreases linearly with iteration number

for $\hat{\mu}$ sufficiently close to μ. The matrix $\mathbf{A}$ can be taken as any matrix whose elements are strictly less than the elements of the second derivative matrix of $\mathbf{1}^T \mathbf{v}(\hat{\mu})$ evaluated at $\hat{\mu} = \mu$.

For this simple problem that matrix can be estimated by entering various values of $\hat{\mu}$ close to μ and observing empirical values of $\mathbf{v}(\hat{\mu})$. This was done using $k = 10{,}000$ at enough different values $\hat{\mu}$ with $\|\hat{\mu} - \mu\| \leq 10^{-5}$ to determine $\mathbf{A}$. For this problem we have

$$\mathbf{A} \approx \begin{pmatrix} 11.038 & -3.444 & -3.444 & -3.444 \\ -3.444 & 11.038 & -3.444 & -3.444 \\ -3.444 & -3.444 & 11.038 & -3.444 \\ -3.444 & -3.444 & -3.444 & 11.038 \end{pmatrix}$$

Note this could also have been done analytically by differentiating equation (3.2) twice.

We can compute the norm

$$\sup_{\|\phi\|=1} \|\mathbf{A}\phi\| \approx 14.482$$

and hence from the proof of Theorem 3.2

$$b = \frac{1}{k} \cdot \sup_{\|\phi\|=1} \|\mathbf{A}\phi\| \approx \frac{14.482}{15} \approx 0.97.$$

Compared to the empirical value $\hat{b} \approx 0.55$, we see that the algorithm appears to be converging much faster than our theoretical bound. Recall this bound was obtained from Markov's inequality and is not necessarily tight.

Algorithm 3.1 was replicated 1000 times and the above procedure used to estimate the slope of the line in Figure 3.1. We can interpret the negative of this slope as the exponential rate at which the conditional variances are converging to zero. These rates are given in Figure 3.3. Over 90% of the values are in the range [0.1, 0.2]. The rate seems fairly consistent, the variation may be entirely due to error in the sample variance.

Suppose we wish to incorporate the modification in Section 3.5 and use information from previous iterations in our estimate. We need to choose the weight, α, to give the most recent simulation runs. Since we should weight in inverse proportion to the variance, we should choose

$$\alpha = \frac{1}{1+\hat{b}}.$$

Now of course in practice we would not have access to the solid line in Figure 3.1 since it depends on the unknown solution μ. But we can use the dashed line which depends only on the empirical standard deviations. Based on this, we would still estimate $\hat{b} \approx 0.55$ and choose

$$\alpha = \frac{1}{1.55} \approx 0.65.$$

Figure 3.4 shows the improvement obtained by using previous information. Here, $\log_{10}(|\hat{\mu}_1^{(m)} - \mu_1|)$ is plotted against m for both versions of the algorithm. The solid

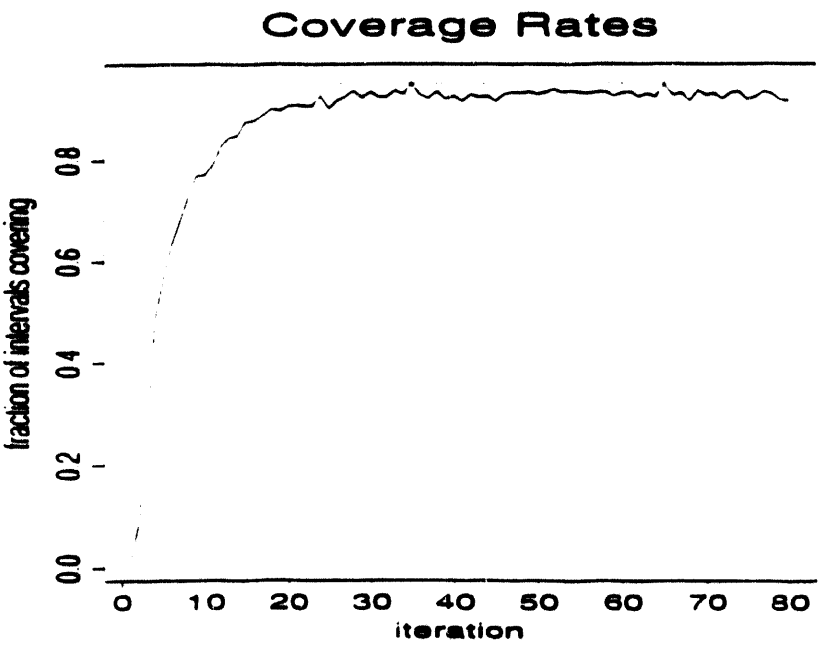


Figure 3.2: Actual coverage rates by iteration for 1000 replications of Algorithm 3.1. Dotted line is target rate of 95%.

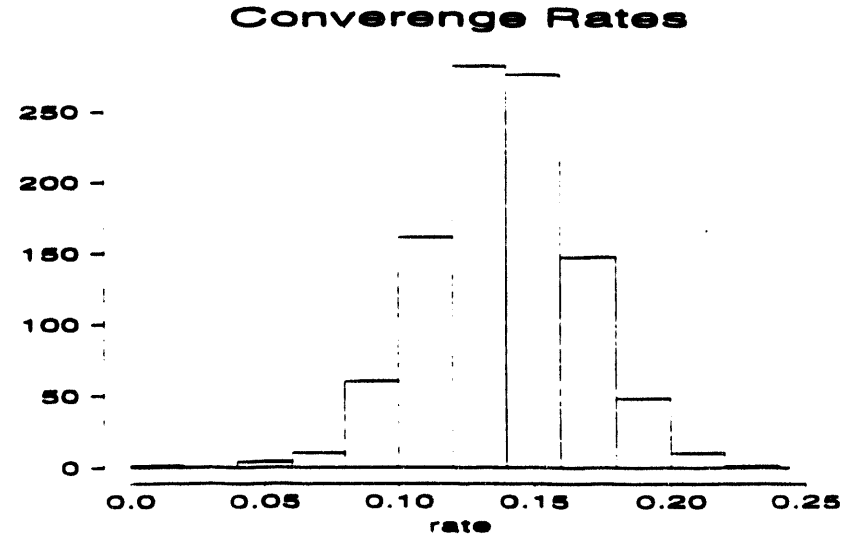


Figure 3.3: Estimated rates of decrease for the conditional variance from 1000 replications of Algorithm 3.1.

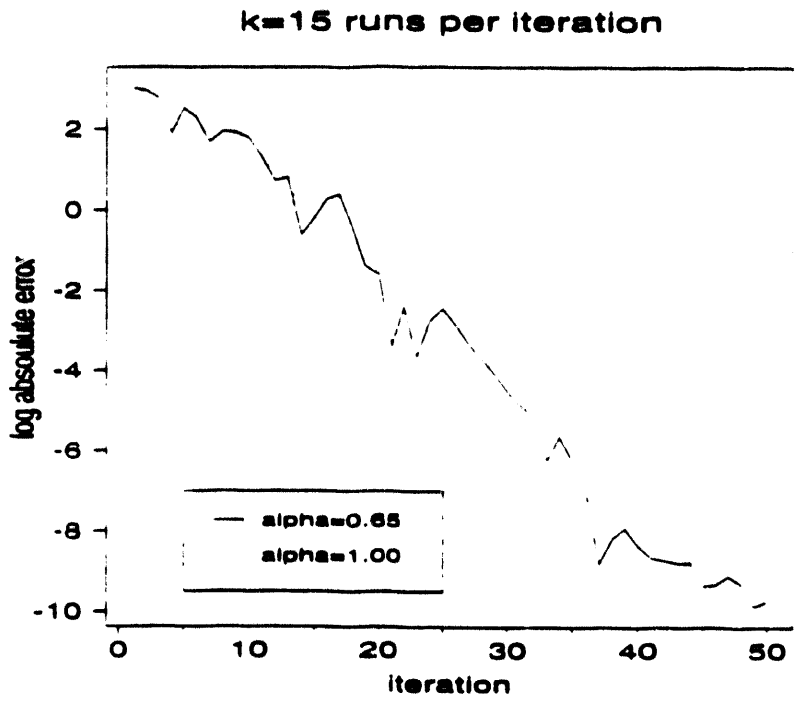


Figure 3.4: Using previous information improves the rate of convergence.

line is decreasing at a faster rate than the dashed line which represents using only the most recent simulations ($\alpha = 1.00$). Note that the improvement does not seem to take effect until the estimates are fairly close to the actual expectation. This indicates that the exponential rates of convergence are only local properties of the process. Recall that the bounds of Lemma 3.1 only hold locally. Our estimate needs to be "close" to the true expectation before they take effect.

Now let us consider Remark 4 of Section 3.3. If we generate a path that involves several states then we not only have information on the expected score from the initial state, we really have information for every state hit by the path. For example, suppose a simulation run generates the path $1 \rightarrow 4 \rightarrow 2 \rightarrow 1 \rightarrow \Delta$. Once the particle makes its initial transition to state 4 we know by the Markov property that the remaining path has the same probability distribution as a particle starting in state 4. Thus, we can use the path from the point it hit state 4, that is, $4 \rightarrow 2 \rightarrow 1 \rightarrow \Delta$ to help

estimate μ_4. Similarly we can use the part of the path $2 \rightarrow 1 \rightarrow \Delta$ to help estimate μ_2.

Here we use the same path to help estimate three different states. By the strong Markov property the scores computed from these partial paths give unbiased estimators. We might be tempted to use the partial path $1 \rightarrow \Delta$ to give a second estimate of μ_1 from this path. However, using the same path to give two different estimates for the same μ_i will produce a bias.

In general, suppose we wish to estimate a paramter, μ, and we have one unbiased estimator X and *possibly* a second unbiased estimator Y. In other words, on some random event E we get a second estimator Y which conditional on E has mean μ. The event E here denotes a path hitting an appropriate state so that Y will be the scores computed on the partial path from that point. We use the estimator

$$X \cdot 1_{E^c} + \left(\frac{X+Y}{2}\right) \cdot 1_E. \tag{3.24}$$

We use X alone if that is all we get, and we use both X and Y if we get access to both.

Under what conditions is this unbiased? We need X to be unbiased for μ *conditional* on E. If X represents an estimate for μ based on one path and Y represents an estimate for μ from a different path then we are ok. Here E represents the second path hitting the appropriate state which is independent of what happened on the first path, and hence independent of X. However, if X and Y represent estimates based on the same path, then the event E is not independent of X and (3.24) will not necessarily be unbiased.

So we can further modify this algorithm by using each path to contruct an estimate for each state hit by the path. As long as only one estimate is computed per state for each path, our estimator will remain unbiased. Note that our estimators for different states are now correlated since some of them may be based on the same paths.

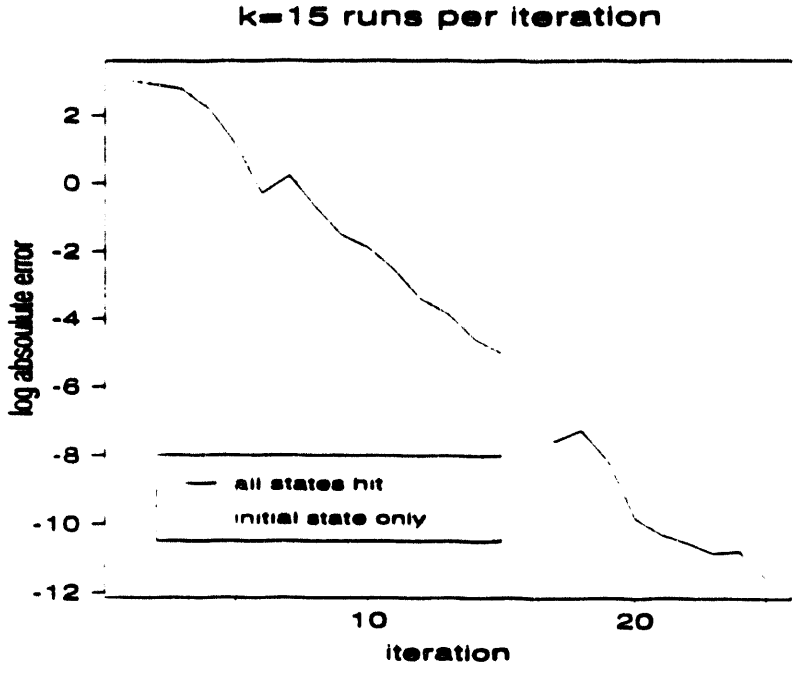


Figure 3.5: Using path information for every state that gets hit increases the effective sample size.

This does not effect the proof of Theorem 3.2. Nowhere was independence between estimators for different states used in the argument. The proof used a variance bound for each state individually. This was based on the fact that k independent unbiased estimators were used to compute $\hat{\mu}_i^{(m)}$. Now a random number of estimators go into computing $\hat{\mu}_i^{(m)}$, but we are guaranteed at least k since we will run that many starting from each state during the iteration. By the strong Markov property conditional on the random number of estimates that go into computing $\hat{\mu}_i^{(m)}$, they are independent and have mean μ_i. Thus, the argument given for the original algorithm hold with this modification.

Figure 3.5 compares empirical results between this modification and the original version of the algorithm. Since $\hat{\mu}_i^{(m)}$ will be based on more than just the k runs we make starting from state i, we are effectively increasing the sample size. Clearly this modification performs much better than the original algorithm. Empirically, the slope

seems to be about -0.70 which translates to $\hat{b} \approx 0.04$.

Admittedly, this modification increases the amount of work we must do per iteration. For each path generated we must keep track of likelihood ratios and scores separately for each state that gets hit. Typically, in transport simulations this work is negligible compared to what is involved in generating the sample path. Here, selecting the next state according to $\mathbf{Q}(\tilde{\mu}^{(m)})$ usually requires work proportional to the number of states in the Markov chain. Keeping track of an extra likelihood ratio and score adds very little extra computation. For each additional state considered, we get the equivalent amount of information as if we had generated an entire new path. So for a relatively small amount of additional computation we get the equivalent of an increased sample size.

Before leaving this example we should consider the choice of k. So far $k = 15$ has been completely arbitrary. For this problem the time spent updating $\mathbf{Q}(\tilde{\mu}^{(m)})$ was negligible compared to the time spent generating sample paths, so we take $r \approx 0$. Note that when $r = 0$, (3.23) is minimized by $k = 1$ for any positive value $\hat{w}$. This suggests trying to choose k as the smallest number that will force the algorithm to converge. For this problem, when Remark 4 of Section 3.3 was used with $k = 1$ the resulting estimators appeared stable suggesting that the algorithm was converging.

3.7 Approximating a continuous space transport problem

Consider the following one dimensional transport problem. A particle within a shield enters a collision at the origin and is absorbed with probability 0.4. If it survives, it travels a random exponential (1) distance. This continues until the particle is absorbed or crosses the level $M = 50$ corresponding to the edge of the shield. We wish to estimate the probability that the particle will penetrate the shield

(cross the level $M = 50$) before it is absorbed.

Again, this problem was chosen because its solution is apparent. Conditional on surviving the initial collision, the particle's total penetration is a geometric sum of exponentials and is hence exponential. Its mean is given by $(0.4)^{-1} = 2.5$. The probability of crossing the point $M = 50$ is therefore given by

$$0.6 \cdot e^{(-0.4 \cdot 50)} \approx 1.24 \cdot 10^{-9}.$$

We can approximate this problem by one of the type described in Section 3.1. Divide up the interval $[0, 50]$ into d intervals as shown in Figure 3.6. Let the d^{th} interval denote crossing the level $M = 50$ so we give a score of one for that state. That is, take

$$s_{ij} = \begin{cases} 1. & \text{if j=d:} \\ 0. & \text{otherwise.} \end{cases}$$

Once the d^{th} state is reached the problem is over so we set

$$p_{dj} = \begin{cases} 1; & \text{if } j = \Delta \\ 0. & \text{otherwise.} \end{cases}$$

Now from a point x, we know there is probability $0.6 \cdot e^{-(50-x)}$ of surviving the collision and crossing the level $M = 50$ on the next step. Since we have partitioned the interval $[0, 50]$ into d cells, each one has length $50 \cdot d^{-1}$. Thus we take

$$p_{id} = 0.6 \cdot \exp\{-50 \cdot d^{-1} \cdot (d - i)\} \quad i = 1, \ldots d - 1.$$

Since we have an absorption probability of 0.4 from every state we take

$$p_{i\Delta} = 0.4 \quad i = 1, \ldots d.$$

If $0 \leq x < y < 50$ then the transition density for the continuous problem is given by

$$p(x, y) = 0.6 \cdot \exp\{-(y - x)\}.$$

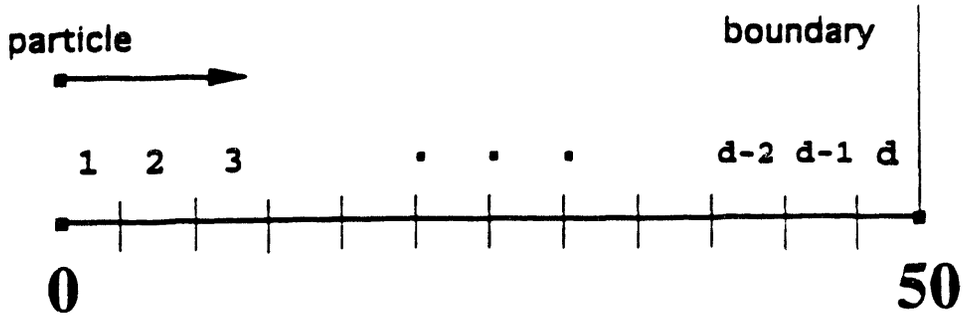


Figure 3.6: Approximate the continuous space with a finite number of cells.

We parallel this by taking

$$p_{ij} \propto \exp\{-50 \cdot d^{-1} \cdot (j-i)\} \quad i,j < d-1.$$

That is, we take

$$p_{ij} = C_i \cdot \exp\{-50 \cdot d^{-1} \cdot (j-i)\} \quad i,j < d-1$$

with

$$C_i = 0.6 \cdot (1 - \exp\{-\frac{50 \cdot (d-i)}{d}\}) \cdot [\sum_{l=1}^{d-1} \exp\{-50 \cdot d^{-1} \cdot (l-i)\}]^{-1}$$

so that

$$\sum_{j=1}^{d+1} p_{ij} = 1.$$

Fifteen trial iterations were run with $k_0 = 5$. a plot of the log of sample standard deviation against iteration is given in Figure 3.7. The process appears to make a relatively steady descent starting with the seventh iteration. A least squares fit estimates the slope at -0.67 which translates into

$$\hat{b} = 10^{(-2 \cdot 0.67)} \approx 0.05.$$

We take $\hat{w} = k_0 \cdot \hat{b} \approx 0.25$. From these runs the ratio of time spent updating $\mathbf{Q}(\hat{\mu}^{(m)})$ to the time running one sample path from each state was estimated at $r = 0.17$. These values are used in (3.24) to obtain the minimizing value of $k = 1$.

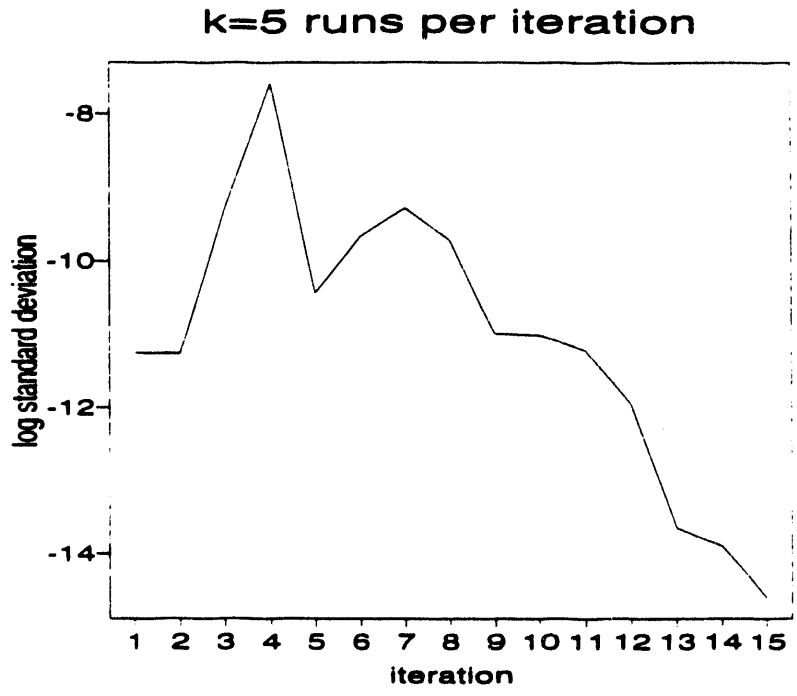


Figure 3.7: Trial iterations to estimate rate of convergence.

Here $\hat{w}$ and r are so small that we would rather get in as many iterations as possible rather than running several simulations per iteration. From (3.22) we take

$$\alpha = \frac{k}{k + \hat{w}} = \frac{1}{1.25} = 0.8.$$

Six more iterations were run with this value of α and $k = 1$. Of course with only one simulation per iteration. there is no way to estimate the variace for the estimator μ_1. Note the particle can not go "backwards" in this model so that there is no hope of getting estimates based on paths initiated from other states. For the last iteration $k = 30$ runs were made so that a standard deviation and a confidence interval based on the central limit theorem could be calculated.

The final result was an approximate 95% confidence interval of

$$9.3404459618 \cdot 10^{-10} \pm 5.479 \cdot 10^{-17}$$

for μ_1. One more iteration was run with $k = 500$ to verify that this interval does

actually cover the true value μ_1. Note from Figure 3.7 that the standard deviation has decreased over the final six iterations from the value of roughly 10^{-14} after the first fifteen trial iterations. This is an indication that $k = 1$ was large enough to induce convergence.

Compare the solution of the finite state problem, $9.34 \cdot 10^{-10}$, to the theoretical value $1.24 \cdot 10^{-9}$. Even with $d = 1000$ states we have almost a 25% error by approximating a continuous problem with a finite one. Better methods for extending this algorithm to a continuous state space are considered in the next chapter.

Chapter 4

Algorithms for Continuous State Spaces

An algorithm for solving problems with finite state Markov chains was presented and studied in the previous chapter. Recall that in transport problems the state space contains information on the particle's position, velocity, and energy level. The possible values for these quantities form a continuum, not a finite set of points. Let us take the state space to be a compact subset of some high dimensional Euclidean space.

In this chapter, methods for extending the algorithm to the continuous space are given and their behavior analyzed. The essential idea of the algorithm is to use previous information on the expected score from each state to compute a low variance importance scheme. Clearly we can not store information on infinitely many states, so we must try to characterize the score function with a finite amount of data.

4.1 Transport Problems as Markov Chains

Suppose we have a transport problem with state space $\mathcal{S}$, a compact subset of a Euclidean space. The position of the particle at its collision sites forms a Markov chain $\{X_n\}_{n=1}^{\infty}$. If we have a particle entering a collision at a point $\mathbf{x} \in \mathcal{S}$, the distribution for the next collision site is given by the density, $p(\mathbf{x}, \cdot)$, with respect to Lebesgue measure on $\mathcal{S}$. That is, for measurable $A \subseteq \mathcal{S}$

$$P\{X_{n+1} \in A \mid X_n = \mathbf{x}\} = \int_A p(\mathbf{x}, \mathbf{y})\, d\mathbf{y}.$$

Take

$$p(\mathbf{x}, \Delta) = 1 - \int_{\mathcal{S}} p(\mathbf{x}, \mathbf{y})\, d\mathbf{y}$$

to be the absorption probability out of state $\mathbf{x}$. As in the the previous chapter, we assume

$$\lim_{n \to \infty} \int_{\mathcal{S}} p^{(n)}(\mathbf{x}, \mathbf{y})\, d\mathbf{y} = 0 \quad \text{for all} \quad \mathbf{x} \in \mathcal{S}$$

where $p^{(n)}(\mathbf{x}, \mathbf{y})$ is given by

$$p^{(1)}(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}, \mathbf{y}) \quad \text{and} \quad p^{(n+1)}(\mathbf{x}, \mathbf{y}) = \int_{\mathcal{S}} p^{(n)}(\mathbf{x}, \mathbf{z}) \cdot p(\mathbf{z}, \mathbf{y})\, d\mathbf{z} \quad n = 1, 2, \ldots .$$

That is, eventual absorption is certain.

For each pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2$ there is the nonnegative score, $s(\mathbf{x}, \mathbf{y})$, incurred by a particle making a transition from $\mathbf{x}$ to $\mathbf{y}$. Again, this score will typically be zero unless state $\mathbf{y}$ corresponds to our target, in which case $s(\mathbf{x}, \mathbf{y})$ denotes the energy being transported there. If

$$\tau = \inf\{n : X_n = \Delta\}$$

denotes the absorption time then we are interested in

$$\mu(\mathbf{x}) = E\left(\sum_{n=1}^{\tau} s(X_{n-1}, X_n) \mid X_0 = \mathbf{x}\right).$$

The system of linear equations (3.1) here becomes a system of integral equations

$$\mu(\mathbf{x}) = p(\mathbf{x}, \Delta) \cdot s(\mathbf{x}, \Delta) + \int_{\mathcal{S}} p(\mathbf{x}, \mathbf{y}) \cdot (s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y}))\, d\mathbf{y} \quad \mathbf{x} \in \mathcal{S}. \tag{4.1}$$

We would like to be able to partition the state space $\mathcal{S}$ into cells, and approximate this system with a finite state problem. If we can approximate this system arbitarily well with finite state problems then we can parallel the theory developed in the previous chapter. Let us impose the following regularity conditions.

Assume there exists a closed set $\mathcal{W} \subset \mathcal{S}$ having Lebesgue measure zero such that:

Assumption 4.1 There exists a $\lambda < 1$ such that

$$\limsup_{n \to \infty} \lambda^{-n} \cdot \sup_{\mathbf{x} \in \mathcal{S}} \int_{\mathcal{S}} p^{(n)}(\mathbf{x}, \mathbf{y})\, d\mathbf{y} < \infty.$$

Assumption 4.2 For each $\mathbf{x} \notin \mathcal{W}$, the functions $p^{(n)}$ and s are differentiable in $\mathbf{x}$, except possibly on set $\mathbf{y}$ having Lebesgue measure zero (the exceptional set may depend on $\mathbf{x}$). Where defined, each derivative is continuous in $(\mathbf{x}, \mathbf{y})$, and

$$\lim_{h \to 0} \int_{\mathcal{S}} \left| \frac{\partial p^{(n)}}{\partial x_i}(\mathbf{x}, \mathbf{y}) - \left(\frac{p^{(n)}(\mathbf{x} + h \cdot \mathbf{e}_i, \mathbf{y}) - p^{(n)}(\mathbf{x}, \mathbf{y})}{h} \right) \right| d\mathbf{y} = 0$$

$$\lim_{h \to 0} \int_{\mathcal{S}} \left| \frac{\partial s}{\partial x_i}(\mathbf{x}, \mathbf{y}) - \left(\frac{s(\mathbf{x} + h \cdot \mathbf{e}_i, \mathbf{y}) - s(\mathbf{x}, \mathbf{y})}{h} \right) \right| d\mathbf{y} = 0.$$

Here, $\mathbf{e}_i$ denotes a vector whose i^{th} component is 1 and whose other components are all 0.

Assumption 4.3 The sum

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}} \left| \frac{\partial p^{(n)}}{\partial x_i}(\mathbf{x}, \mathbf{y}) \right| d\mathbf{y}$$

converges uniformly on $\mathbf{x} \in \mathcal{S} - \mathcal{W}$. Furthermore,

$$\sup_{\mathbf{x} \in \mathcal{S} - \mathcal{W}} \sum_{n=1}^{\infty} \int_{\mathcal{S}} \left| \frac{\partial p^{(n)}}{\partial x_i}(\mathbf{x}, \mathbf{y}) \right| d\mathbf{y} < \infty.$$

Assumption 4.4 The score function s and its derivative are uniformly bounded,

$$\sup_{(\mathbf{x},\mathbf{y})\in\mathcal{S}^2} |s(\mathbf{x},\mathbf{y})| + \left|\frac{\partial s}{\partial x_i}(\mathbf{x},\mathbf{y})\right| < \infty.$$

Note that a sufficient condition for Assupmtion 4.1 to hold is that

$$\inf_{\mathbf{x}\in\mathcal{S}} p(\mathbf{x},\Delta) > 0.$$

We allow for an exceptional set, $\mathcal{W}$, to handle boundaries where the transition probabilities may change abruptly, the edge of a lead shield or a cement wall, etc. As we impose finer and finer partitions on $\mathcal{S}$, the percentage of cells containing points in $\mathcal{W}$ will tend to zero.

Take

$$a(\mathbf{x}) = p(\mathbf{x},\Delta)\cdot s(\mathbf{x},\Delta) + \int_{\mathcal{S}} p(\mathbf{x},\mathbf{y})\cdot s(\mathbf{x},\mathbf{y})\, d\mathbf{y}.$$

Iterating equation (4.1) n times gives

$$\mu(\mathbf{x}) = a(\mathbf{x}) + \sum_{j=1}^{n-1}\int_{\mathcal{S}} p^{(j)}(\mathbf{x},\mathbf{y})\cdot a(\mathbf{y})\, d\mathbf{y} + \int_{\mathcal{S}} p^{(n)}(\mathbf{x},\mathbf{y})\cdot \mu(\mathbf{y})\, d\mathbf{y}.$$

By Assumptions 4.1 - 4.4, the final term tends to zero as $n \to \infty$ so that

$$\mu(\mathbf{x}) = a(\mathbf{x}) + \sum_{n=1}^{\infty}\int_{\mathcal{S}} p^{(n)}(\mathbf{x},\mathbf{y})\cdot a(\mathbf{y})\, d\mathbf{y} \tag{4.2}$$

with $\mu(\cdot)$ being differentiable for $\mathbf{x}\notin\mathcal{W}$. By Assumptions 4.3 and 4.4

$$\sup_{\mathbf{x}\in\mathcal{S}} \|\nabla\mu(\mathbf{x})\| < \infty.$$

To avoid technical problems, let us again assume there exists a known $\delta > 0$ such that

$$\inf_{\mathbf{x}\in\mathcal{S}} \mu(\mathbf{x}) \geq \delta.$$

See the discussuion in Section 3.3. We shall require $\hat{\mu}(\mathbf{x}) \geq \delta$, $\mathbf{x}\in\mathcal{S}$ for our estimates, $\hat{\mu}(\cdot)$, of $\mu(\cdot)$.

4.2 Importance Sampling on Continuous Spaces

Just as in the finite state case, we need to choose a transition kernel $q(\mathbf{x}, \mathbf{y})$ to use in our simulations instead of the true $p(\mathbf{x}, \mathbf{y})$. Our choice must satisfy

$$q(\mathbf{x}, \mathbf{y}) > 0 \quad \text{whenever} \quad p(\mathbf{x}, \mathbf{y}) > 0 \qquad \text{and} \quad q(\mathbf{x}, \Delta) > 0 \quad \text{whenever} \quad p(\mathbf{x}, \Delta) > 0.$$

We must also make sure our choice satisfies

$$\lim_{n \to \infty} q^{(n)}(\mathbf{x}, \Delta) = 1 \quad \mathbf{x} \in \mathcal{S}.$$

Note that this would always hold in the finite state case, but not necessarily in a continuous space.

Let

$$L_n = \prod_{j=1}^{n} \frac{q(\mathbf{X}_{j-1}, \mathbf{X}_j)}{p(\mathbf{X}_{j-1}, \mathbf{X}_j)}$$

and take

$$Y_q = \sum_{n=1}^{\tau} s(\mathbf{X}_{n-1}, \mathbf{X}_n) \cdot L_n.$$

The Radon-Nikodym theorem tells us that on the event $\{\mathbf{X_0} = \mathbf{x}\}$, Y_q has expectation $\mu(\mathbf{x})$.

Just as in the finite state problem, there exists a zero variance importance scheme. We just parallel (3.3) taking

$$q(\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y})]}{\mu(\mathbf{x})}.$$

The problem here is that even if we knew the function $\mu(\cdot)$ exactly, we would need an infinite amount of computer memory to store it. In practice, of course, we must approximate $\mu(\cdot)$ with a finite amount of data. We may partition $\mathcal{S}$ into a finite number of disjoint cells $\{C_j\}_{j=1}^{d}$ and obtain an estimate $\hat{\mu}(\cdot)$ which is constant on each C_j. This approximation will add some variance to our estimator.

Note that even with a crude partition of $\mathcal{S}$ we have eliminated the problem of bias encountered in the example in Section 3.6.2. Since we are doing the simulations directly in the continuous space, we obtain unbiased estimators for the solution to (4.1). In the finite state example, we obtained unbiased estimators for the solution to the approximating system of d equations in d unknowns.

Clearly the better our approximation, the better our choice of $q(\cdot,\cdot)$ will resemble the zero variance scheme, and the smaller our variance. However, a very close approximation, will require a fine partition of the state space which will require a very large number of cells. We must consider the tradeoff between low variance estimators requiring very fine approximations and the work needed to obtain them. To do this, we must calculate how the variance grows as a function of our approximation error $\hat{\mu} - \mu$.

Suppose we have an estimate, $\hat{\mu}(\cdot)$, for the function $\mu(\cdot)$. We use our best guess at the zero variance importance scheme

$$q(\mathbf{x},\mathbf{y},\hat{\mu}) = \frac{p(\mathbf{x},\mathbf{y})\cdot[\,s(\mathbf{x},\mathbf{y})+\hat{\mu}(\mathbf{y})\,]}{p(\mathbf{x},\Delta)\cdot s(\mathbf{x},\Delta)+\int_{\mathcal{S}} p(\mathbf{x},\mathbf{z})\cdot[\,s(\mathbf{x},\mathbf{z})+\hat{\mu}(\mathbf{z})\,]\,d\mathbf{z}}.$$

Following the notation of the previous chapter take

$$f(\mathbf{x},\hat{\mu}) = \frac{[\,p(\mathbf{x},\Delta)\cdot s(\mathbf{x},\Delta)\,]^2}{q(\mathbf{x},\Delta,\hat{\mu})} + \int_{\mathcal{S}} \frac{\left(p(\mathbf{x},\mathbf{y})\cdot[\,s(\mathbf{x},\mathbf{y})+\hat{\mu}(\mathbf{y})\,]\right)^2}{q(\mathbf{x},\mathbf{y},\hat{\mu})}\,d\mathbf{y} - [\mu(\mathbf{x})]^2$$

and let

$$r(\mathbf{x},\mathbf{y},\hat{\mu}) = \frac{[\,p(\mathbf{x},\mathbf{y})\,]^2}{q(\mathbf{x},\mathbf{y},\hat{\mu})}.$$

Here we interpret 0/0 to be 0. Take $v(\mathbf{x},\hat{\mu})$ to be the variance of our estimator Y_q when simulating under $q(\cdot,\cdot,\hat{\mu})$ and starting $\mathbf{X}_0 = \mathbf{x}$. Then (3.2) in this case becomes

$$v(\mathbf{x},\hat{\mu}) = f(\mathbf{x}) + \int_{\mathcal{S}} r(\mathbf{x},\mathbf{y},\hat{\mu})\cdot v(\mathbf{y},\hat{\mu})\,d\mathbf{y}.$$

The argument given in Theorem 3.1 applies here so that we have the solution

$$v(\mathbf{x},\hat{\mu}) = f(\mathbf{x},\hat{\mu}) + \sum_{n=1}^{\infty} \int_{\mathcal{S}} r^{(n)}(\mathbf{x},\mathbf{y},\hat{\mu})\cdot f(\mathbf{y},\hat{\mu})\,d\mathbf{y}. \tag{4.3}$$

An analogous argument to that given in Section 3.2 shows that $f(\mathbf{x}, \mu) = 0$ and hence $v(\mathbf{x}, \mu) = 0$ for all $\mathbf{x} \in \mathcal{S}$.

Now for fixed $\mathbf{x} \in \mathcal{S}$, the functional $f(\mathbf{x}, \cdot)$ is differentiable with respect to the function $\hat{\mu}$ in the sense of Definition A.1 in Appendix A. That is, there exists a function $\nabla f(\mathbf{x}, \hat{\mu}, \mathbf{y})$ such that for a function $h : \mathcal{S} \to \mathcal{R}$

$$f(\mathbf{x}, \hat{\mu} + h) = f(\mathbf{x}, \hat{\mu}) + \int_{\mathcal{S}} \nabla f(\mathbf{x}, \hat{\mu}, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} + o(\|h\|) \tag{4.4}$$

where

$$\|h\| = \text{ess sup}\, h = \inf\{M : L\{\mathbf{y} : h(\mathbf{y}) \geq M\} > 0\}.$$

Here, L denotes Lebesgue measure on $\mathcal{S}$. The functional $o(\cdot)$ is taken to have the property that for any sequence of functionals, $\{h_n\}$ with $\|h_n\| \to 0$

$$\lim_{n \to \infty} \frac{o(h_n)}{\|h_n\|} = 0.$$

Note that for fixed $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2$ the functional $q(\mathbf{x}, \mathbf{y}, \hat{\mu})$ is not differentiable in the sense of (4.4) because it depends too heavily on the value $\hat{\mu}(y)$. Any functional satisfying (4.4) would not change simply by altering the value of $\hat{\mu}$ at a single point. The functional $f(\mathbf{x}, \hat{\mu})$ becomes differentiable because it is the integral over $\mathcal{S}$ of functions involving q and $\hat{\mu}$. See the calculation following Theorem A.3 in Appendix A for details.

For the remainder of this chapter we deal with the class of estimates,
$\mathcal{L} = \left\{\hat{\mu} : \inf_{\mathbf{x} \in \mathcal{S}} \hat{\mu}(\mathbf{x}) \geq \delta\right\}$ with topology induced by the norm $\|\hat{\mu}\| = \text{ess sup}\, \hat{\mu}$.

Theorem 4.1 *There exists an open collection, $\mathcal{M}$, of functions, $\hat{\mu} : \mathcal{S} \to \mathcal{R}$, such that $\mu \in \mathcal{M}$, and for each fixed $\mathbf{x} \in \mathcal{S}$, $v(\mathbf{x}, \cdot)$ is differentiable on $\mathcal{M}$. Furthermore,*

$$\nabla v(\mathbf{x}, \mu, \mathbf{y}) = 0 \quad \textit{for all} \quad (\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2.$$

Proof: Let

$$g_n(\mathbf{x}, \hat{\mu}) = \int_S r^{(n)}(\mathbf{x}, \mathbf{y}, \hat{\mu}) \cdot f(\mathbf{y}, \hat{\mu})\, d\mathbf{y} \qquad n = 1, 2 \ldots \tag{4.5}$$

and

$$g_0(\mathbf{x}, \hat{\mu}) = f(\mathbf{x}, \hat{\mu}).$$

Note

$$v(\mathbf{x}, \hat{\mu}) = \sum_{n=0}^{\infty} g_n(\mathbf{x}, \hat{\mu}). \tag{4.6}$$

When we take $\hat{\mu} = \mu$, we can parallel the argument given in the proof of Lemma 3.1

$$r(\mathbf{x}, \mathbf{y}, \mu) = \frac{\mu(\mathbf{x}) \cdot p(\mathbf{x}, \mathbf{y})}{s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y})} \leq \frac{\mu(\mathbf{x}) \cdot p(\mathbf{x}, \mathbf{y})}{\mu(\mathbf{y})}.$$

For $\hat{\mu} \in \mathcal{L}$ with $\|\hat{\mu} - \mu\|$ sufficiently small

$$r(\mathbf{x}, \mathbf{y}, \hat{\mu}) \leq \frac{2\,\mu(\mathbf{x}) \cdot p(\mathbf{x}, \mathbf{y})}{\mu(\mathbf{y})}.$$

and by induction

$$r^{(n)}(\mathbf{x}, \mathbf{y}, \hat{\mu}) \leq \frac{2\,\mu(\mathbf{x}) \cdot p^{(n)}(\mathbf{x}, \mathbf{y})}{\mu(\mathbf{y})}$$

Let

$$\mathcal{M} = \{\hat{\mu} \in \mathcal{L} : \|\hat{\mu} - \mu\| < \eta\}$$

where η is chosen small enough so that the above inequality holds.

By Assumption 4.2 this implies

$$\lim_{n \to \infty} \lambda^{-n} \cdot \sup_{\hat{\mu} \in \mathcal{M}, \mathbf{x} \in S} \int_S r^{(n)}(\mathbf{x}, \mathbf{y}, \hat{\mu})\, d\mathbf{y} < \infty. \tag{4.7}$$

For $\hat{\mu} \in \mathcal{M}$, $f(\mathbf{x}, \hat{\mu})$ is uniformly bounded on $\mathbf{x} \in \mathcal{S}$, and thus by (4.5)

$$\lim_{n \to \infty} \lambda^{-n} \cdot \sup_{\hat{\mu} \in \mathcal{M}, \mathbf{x} \in S} g_n(\mathbf{x}, \hat{\mu})\, d\mathbf{y} < \infty. \tag{4.8}$$

Let

$$c(\mathbf{x}, \hat{\mu}) = p(\mathbf{x}, \Delta) \cdot s(\mathbf{x}, \Delta) + \int_{\mathcal{S}} p(\mathbf{x}, \mathbf{z}) \cdot [s(\mathbf{x}, \mathbf{z}) + \hat{\mu}(\mathbf{z})] \, d\mathbf{z}$$

be the denominator term in $q(\mathbf{x}, \mathbf{y}, \hat{\mu})$. For fixed $\mathbf{x}$, the functional $c(\mathbf{x}, \cdot)$ is clearly differentiable in $\hat{\mu}$ with

$$\nabla c(\mathbf{x}, \hat{\mu}, \mathbf{y}) = p(\mathbf{x}, \mathbf{y}).$$

Now

$$g_{n+1}(\mathbf{x}, \hat{\mu}) = \int_{\mathcal{S}} r(\mathbf{x}, \mathbf{y}, \hat{\mu}) \cdot g_n(\mathbf{y}, \hat{\mu}) \, d\mathbf{y} = c(\mathbf{x}, \hat{\mu}) \cdot \int_{\mathcal{S}} \frac{p(\mathbf{x}, \mathbf{y}) \cdot g_n(\mathbf{x}, \hat{\mu})}{s(\mathbf{x}, \mathbf{y}) - \hat{\mu}(\mathbf{y})} \, d\mathbf{y}. \tag{4.9}$$

Note that for fixed $\mathbf{x} \in \mathcal{S}$, $g_0(\mathbf{x}, \cdot) = f(\mathbf{x}, \cdot)$ is differentiable in $\hat{\mu}$ (see Theorem A.3). Using the differentiation rule given in Theorem A.3, and an induction argument, we see that for fixed $\mathbf{x} \in \mathcal{S}$, $g_n(\mathbf{x}, \cdot)$ is differentiable in $\hat{\mu}$ and

$$\nabla g_{n+1}(\mathbf{x}, \hat{\mu}, \mathbf{y}) = -c(\mathbf{x}, \hat{\mu}) \cdot \frac{p(\mathbf{x}, \mathbf{y}) \cdot g_n(\mathbf{x}, \hat{\mu})}{[s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})]^2} + \tag{4.10}$$

$$p(\mathbf{x}, \mathbf{y}) \cdot \int_{\mathcal{S}} \frac{p(\mathbf{x}, \mathbf{z}) \cdot g_n(\mathbf{z}, \hat{\mu})}{s(\mathbf{x}, \mathbf{z}) + \hat{\mu}(\mathbf{z})} \, d\mathbf{z} + c(\mathbf{x}, \hat{\mu}) \cdot \int_{\mathcal{S}} r(\mathbf{x}, \mathbf{z}, \hat{\mu}) \cdot \nabla g_n(\mathbf{z}, \hat{\mu}, \mathbf{y}) \, d\mathbf{z}.$$

Sucessively iterating this equation gives

$$\nabla g_n(\mathbf{x}, \hat{\mu}, \mathbf{y}) = -c(\mathbf{x}, \hat{\mu}) \cdot \frac{p(\mathbf{x}, \mathbf{y}) \cdot g_n(\mathbf{x}, \hat{\mu})}{[s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})]^2} + p(\mathbf{x}, \mathbf{y}) \cdot \int_{\mathcal{S}} \frac{p(\mathbf{x}, \mathbf{z}) \cdot g_n(\mathbf{z}, \hat{\mu})}{s(\mathbf{x}, \mathbf{z}) + \hat{\mu}(\mathbf{z})} d\mathbf{z} +$$

$$\sum_{j=1}^{n} \int_{\mathcal{S}} r^{(j)}(\mathbf{x}, \mathbf{z}, \hat{\mu}) \cdot \left[c(\mathbf{z}, \hat{\mu}) \cdot \frac{p(\mathbf{z}, \mathbf{y}) \cdot g_{n-j}(\mathbf{z}, \hat{\mu})}{[s(\mathbf{z}, \mathbf{y}) + \hat{\mu}(\mathbf{y})]^2} + p(\mathbf{z}, \mathbf{y}) \cdot \int_{\mathcal{S}} \frac{p(\mathbf{z}, \mathbf{w}) \cdot g_{n-j}(\mathbf{w}, \hat{\mu})}{s(\mathbf{z}, \mathbf{w}) + \hat{\mu}(\mathbf{w})} \, d\mathbf{w} \right] d\mathbf{z} +$$

$$\int_{\mathcal{S}} r^{(n+1)}(\mathbf{x}, \mathbf{z}, \hat{\mu}) \cdot \nabla f(\mathbf{z}, \hat{\mu}, \mathbf{y}) \, d\mathbf{z}.$$

If we choose $\tilde{\lambda} > \lambda$ then by (4.7), and (4.8) we have

$$\lim_{n \to \infty} \tilde{\lambda}^{-n} \cdot \sup_{\hat{\mu} \in \mathcal{M}, \mathbf{x} \in \mathcal{S}} \int_{\mathcal{S}} \nabla g_n(\mathbf{x}, \hat{\mu}, \mathbf{y}) \, d\mathbf{y} < \infty. \tag{4.11}$$

Choosing $\lambda < \tilde{\lambda} < 1$ we see that the sum

$$\sum_{n=1}^{\infty} \int_{\mathcal{S}} \nabla g_n(\mathbf{x}, \hat{\mu}, \mathbf{y})$$

converges uniformly over $\mathbf{x} \in \mathcal{S}$, and $\hat{\mu} \in \mathcal{M}$.

From Theorem A.4, and (4.6), for each fixed $\mathbf{x} \in \mathcal{S}$, the function $v(\mathbf{x}, \cdot)$ is differentiable on $\mathcal{M}$ and

$$\nabla v(\mathbf{x}, \hat{\mu}, \mathbf{y}) = \sum_{n=0}^{\infty} \nabla g_n(\mathbf{x}, \hat{\mu}, \mathbf{y}).$$

Since $v(\cdot, \mu)$ is identically zero, it has a local minimum at μ for each $\mathbf{x} \in \mathcal{S}$, and so

$$\nabla v(\mathbf{x}, \mu, \mathbf{y}) = 0 \quad \text{for all} \quad (\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2. \quad \blacksquare$$

We now need to develop a notion of a second derivative. Unfortunately, we can not simply define it as the derivative of the derivative since this would rule out too many functionals. For example, in (4.10) the first term of $\nabla g_{n+1}(\mathbf{x}, \hat{\mu}, \mathbf{y})$ depends on $\hat{\mu}$ only through the value $\hat{\mu}(\mathbf{y})$. For fixed $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2$, therefore, $\nabla g_{n+1}(\mathbf{x}, \cdot, \mathbf{y})$ can not satisfy (4.4). To include such functionals, g, a $\nabla_D^2 g$ term is added to handle the dependence of $\nabla g(\hat{\mu}, \mathbf{y})$ on the value $\hat{\mu}(\mathbf{y})$. See the discussion preceeding Definition A.2 in Appendix A.

We consider a functional g to be twice differentiable if in addition to being differentiable in the sense of (4.4), there exist mappings $\nabla^2 g : \mathcal{M} \times \mathcal{S}^2 \to \mathcal{R}$, and $\nabla_D^2 g : \mathcal{M} \times \mathcal{S} \to \mathcal{R}$ such that

$$g(\hat{\mu} + h) = g(\hat{\mu}) + \int_{\mathcal{S}} \nabla g(\hat{\mu}, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} + \tag{4.12}$$

$$\frac{1}{2} \cdot \left[\iint_{\mathcal{S}^2} h(\mathbf{y}) \cdot \nabla^2 g(\hat{\mu}, \mathbf{y}, \mathbf{z}) \cdot h(\mathbf{z})\, d\mathbf{y}\, d\mathbf{z} + \int_{\mathcal{S}} \nabla_D^2 g(\hat{\mu}, \mathbf{y}) \cdot [h(\mathbf{y})]^2\, d\mathbf{y} \right] + o(\|h\|^2).$$

Theorem 4.2 *Let $\mathcal{M}$ be as in Theorem 4.1. For each fixed $\mathbf{x} \in \mathcal{S}$, the function $v(\mathbf{x}, \cdot)$ is twice differentiable in the sense of (4.12).*

Proof: The argument is analogous to the proof of Theorem 4.1. Recall

$$f(\mathbf{x}, \hat{\mu}) = \frac{[p(\mathbf{x}, \Delta) \cdot s(\mathbf{x}, \Delta)]^2}{q(\mathbf{x}, \Delta, \hat{\mu})} + \int_{\mathcal{S}} \frac{\left(p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})]\right)^2}{q(\mathbf{x}, \mathbf{y}, \hat{\mu})}\, d\mathbf{y} - [\mu(\mathbf{x})]^2 =$$

$$c(\mathbf{x},\hat{\mu})\cdot\left[p(\mathbf{x},\Delta)\cdot s(\mathbf{x},\Delta)+\int_S \frac{p(\mathbf{x},\mathbf{y})\cdot[s(\mathbf{x},\mathbf{y})+\mu(\mathbf{y})]^2}{s(\mathbf{x},\mathbf{y})+\hat{\mu}(\mathbf{y})}\,d\mathbf{y}-[\mu(\mathbf{x})]^2\right].$$

Fixing $\mathbf{x}\in\mathcal{S}$, we consider $f(\mathbf{x},\cdot)$ a functional on $\mathcal{M}$. Now

$$c(\mathbf{x},\hat{\mu})=p(\mathbf{x},\Delta)\cdot s(\mathbf{x},\Delta)+\int_S p(\mathbf{x},\mathbf{z})\cdot[s(\mathbf{x},\mathbf{z})+\hat{\mu}(\mathbf{z})]\,d\mathbf{z}$$

is linear in $\hat{\mu}$ and so for $\mathbf{x}$ fixed

$$\nabla^2 c(\mathbf{x},\hat{\mu},\mathbf{y},\mathbf{z})=\nabla_D^2 c(\mathbf{x},\hat{\mu},\mathbf{y})=0$$

for all $(\mathbf{y},\mathbf{z})\in\mathcal{S}^2$. Now apply Theorem A.7 with $g(\mathbf{y},\hat{\mu})=c(\mathbf{x},\hat{\mu})$ and

$$\theta(\mathbf{y},t)=\frac{p(\mathbf{x},\mathbf{y})\cdot[s(\mathbf{x},\mathbf{y})+\mu(\mathbf{y})]}{s(\mathbf{x},\mathbf{y})+t}\quad\text{together with}\quad \nu(\mathbf{y})=p(\mathbf{x},\mathbf{y})+1$$

to conclude that $f(\mathbf{x},\cdot)$ is twice differentiable with

$$\nabla^2 f(\mathbf{x},\hat{\mu},\mathbf{y},\mathbf{z})=p(\mathbf{x},\mathbf{z})\cdot p(\mathbf{x},\mathbf{y})\cdot\left(\frac{s(\mathbf{x},\mathbf{y})+\mu(\mathbf{y})}{s(\mathbf{x},\mathbf{y})+\hat{\mu}(\mathbf{y})}\right)^2$$

and

$$\nabla_D^2 f(\mathbf{x},\hat{\mu},\mathbf{y})=2\,c(\mathbf{x},\hat{\mu})\cdot p(\mathbf{x},\mathbf{y})\cdot\frac{[s(\mathbf{x},\mathbf{y})+\mu(\mathbf{y})]^2}{[s(\mathbf{x},\mathbf{y})+\hat{\mu}(\mathbf{y})]^3}.$$

Note here that $g(\mathbf{y},\hat{\mu})=c(\mathbf{x},\hat{\mu})$ does not depend on $\mathbf{y}$.

To obtain the second derivative analogs of (4.10), we use an induction argument applying Theorem A.5 and Theorem A.7 to the equation (4.9) to conclude that $g_n(\mathbf{x},\cdot)$ is twice differentiable with

$$\nabla^2 g_{n+1}(\mathbf{x},\hat{\mu},\mathbf{y},\mathbf{z})=$$

$$c(\mathbf{x},\hat{\mu})\cdot\left[\frac{-2\,p(\mathbf{x},\mathbf{y})\cdot\nabla g_n(\mathbf{y},\hat{\mu},\mathbf{z})}{[s(\mathbf{x},\mathbf{y})+\hat{\mu}(\mathbf{y})]^2}+\int_S\frac{p(\mathbf{x},\mathbf{w})\cdot\nabla^2 g_n(\mathbf{w},\hat{\mu},\mathbf{z},\mathbf{y})}{s(\mathbf{x})+\hat{\mu}(\mathbf{y})}\,d\mathbf{w}\right]+ \tag{4.13}$$

$$2\,p(\mathbf{x},\mathbf{y})\cdot\nabla g_n(\mathbf{x},\hat{\mu},\mathbf{z}).$$

and

$$\nabla_D^2 g_{n+1}(\mathbf{x}, \hat{\mu}, \mathbf{y}) =$$

$$c(\mathbf{x}, \hat{\mu}) \cdot \left[\frac{2\, p(\mathbf{x}, \mathbf{y}) \cdot g_n(\mathbf{y}, \hat{\mu})}{[\, s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})\,]^3} + \int_S \frac{p(\mathbf{x}, \mathbf{z}) \cdot \nabla_D^2 g_n(\mathbf{z}, \hat{\mu}, \mathbf{y})}{s(\mathbf{x}, \mathbf{z}) + \hat{\mu}(\mathbf{z})}\, d\mathbf{z} \right]. \tag{4.14}$$

We now use the same argument that was applied to equation (4.10) in the proof of Theorem 4.1. Successively iterating equations (4.13) and (4.14), and using (4.7), (4.8), and (4.11) we see that

$$\lim_{n \to \infty} \tilde{\lambda}^{-n} \cdot \sup_{\hat{\mu} \in \mathcal{M}, \mathbf{x} \in S} \iint_S \nabla^2 g_n(\mathbf{x}, \hat{\mu}, \mathbf{y}, \mathbf{z})\, d\mathbf{y}\, d\mathbf{z} < \infty$$

and

$$\lim_{n \to \infty} \tilde{\lambda}^{-n} \cdot \sup_{\hat{\mu} \in \mathcal{M}, \mathbf{x} \in S} \int_S \nabla_D^2 g_n(\mathbf{x}, \hat{\mu}, \mathbf{y})\, d\mathbf{y} < \infty$$

where $\tilde{\lambda}$ is as in (4.11).

Now apply Theorem A.8 to equation (4.6). ∎

We can combine Theorems 4.1 and 4.2 to obtain the following analog to Lemma 3.1

$$v(\mathbf{x}, \hat{\mu}) = \frac{1}{2} \cdot \left[\iint_{S^2} (\hat{\mu}(\mathbf{y}) - \mu(\mathbf{y})) \cdot \nabla^2 v(\mathbf{x}, \mu, \mathbf{y}, \mathbf{z}) \cdot (\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z}))\, d\mathbf{y}\, d\mathbf{z} + \right. \tag{4.15}$$
$$\left. \int_S \nabla_D^2 v(\mathbf{x}, \mu, \mathbf{y}) \cdot [(\hat{\mu}(\mathbf{y}) - \mu(\mathbf{y}))]^2\, d\mathbf{y} \right] + o(\|\hat{\mu} - \mu\|^2).$$

4.3 Partitioning the State Space

As mentioned in the previous section, we might partition the space S into a finite collection of disjoint cells $\{C_j\}_{j=1}^d$ with

$$\omega = \sup_j \sup_{(\mathbf{x}, \mathbf{y}) \in C_j^2} \|\mathbf{x} - \mathbf{y}\|.$$

Suppose we can impose such a partition so that each point $\mathbf{x} \in \mathcal{W}$ lies on the boundary of some cell C_j. For example, if $\mathcal{W}$ is the edge of a cement wall we can allow that edge to define cell boundaries. No one cell would contain points corresponding to both air

and cement. Suppose for each cell C_j, we select a point $\mathbf{y}_j$ in the interior of C_j and approximate the function μ by

$$\hat{\mu}(\mathbf{y}) = \sum_{j=1}^{d} \mu(\mathbf{y}_j) \cdot 1\{\mathbf{y} \in C_j\}. \tag{4.16}$$

By the mean value theorem, for each $\mathbf{y}$ in the interior of C_j, there exists an $\mathbf{y}^* \in C_j$ so that

$$\mu(\mathbf{y}) - \mu(\mathbf{y}_j) = \nabla\mu(\mathbf{y}^*) \cdot (\mathbf{y} - \mathbf{y}_j).$$

Taking $h(\mathbf{y}) = \hat{\mu}(\mathbf{y}) - \mu(\mathbf{y})$, we have

$$\int_{C_j} \nabla_D^2 v(\mathbf{x}, \mu, \mathbf{y}) \cdot [h(\mathbf{y})]^2 d\mathbf{y} \leq \sup_{\mathbf{y} \in \mathcal{S}} \|\nabla\mu(\mathbf{y})\| \cdot \omega^2 \cdot \int_{C_j} \nabla_D^2 v(\mathbf{x}, \mu, \mathbf{y})\, d\mathbf{y}$$

and thus

$$\int_{\mathcal{S}} \nabla_D^2 v(\mathbf{x}, \hat{\mu}, \mathbf{y}) \cdot [h(\mathbf{y})]^2 d\mathbf{y} \leq \sup_{\mathbf{y} \in \mathcal{S}} \|\nabla\mu(\mathbf{y})\| \cdot \int_{\mathcal{S}} \nabla_D^2 v(\mathbf{x}, \mu, \mathbf{y})\, d\mathbf{y} \cdot \omega^2.$$

Similarly

$$\iint_{\mathcal{S}^2} (\hat{\mu}(\mathbf{y}) - \mu(\mathbf{y})) \cdot \nabla^2 v(\mathbf{x}, \mu, \mathbf{y}, \mathbf{z}) \cdot (\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z}))\, d\mathbf{y}\, d\mathbf{z} \leq$$

$$\sup_{\mathbf{y} \in \mathcal{S}} \|\nabla\mu(\mathbf{y})\| \cdot \iint_{\mathcal{S}^2} \nabla^2 v(\mathbf{x}, \mu, \mathbf{y}, \mathbf{z}) d\mathbf{y}\, d\mathbf{z} \cdot \omega^2.$$

By (4.15)

$$v(\mathbf{x}, \hat{\mu}) \leq \left[\frac{1}{2} \cdot \iint_{\mathcal{S}^2} \nabla^2 v(\mathbf{x}, \mu, \mathbf{y}, \mathbf{z})\, d\mathbf{y}\, d\mathbf{z} + \int_{\mathcal{S}} \nabla_D^2 v(\mathbf{x}, \mu, \mathbf{y})\, d\mathbf{y} \cdot \sup_{\mathbf{y} \in \mathcal{S}} \|\nabla\mu(\mathbf{y})\|\right] \cdot \omega^2 +$$

$$o(\omega^2). \tag{4.17}$$

Our variance is proportional to the square of the cell diameters in the partition we choose. Of course in practice we would not know the values $\mu(\mathbf{y}_j)$ so that we would have to replace (4.16) by

$$\hat{\mu}(\mathbf{y}) = \sum_{j=1}^{d} \hat{\mu}_j \cdot 1\{\mathbf{y} \in C_j\} \tag{4.18}$$

where $\hat{\mu}_j$ represents an estimate for some notion of the expected score starting from cell C_j.

We may be tempted to impose finer and finer partitions on $\mathcal{S}$ in an attempt to converge to the zero variance importance sampling scheme. However, if we do this, we will encounter the same curse of dimensionality that hinder deterministic methods. In high dimensional spaces the number of cells is growing very rapidly as we try to make the cell diameters, ω, small. For example, in Section 1.1 we saw that $\mathcal{S}$ may contain information on the particle's position, velocity, and time. This would require seven dimensions so that the number of cells in our partition is growing as

$$d = O(\omega^{-7}).$$

The amount of work required to obtain estimates $\hat{\mu}_j$ for each cell is growing too rapidly to make very fine partitions worth our while. We are much better off imposing a relatively crude partition on our state space with d being a manageable number of cells. Once the estimates $\hat{\mu}_j$ are obtained, we can perform iid importance sampling to obtain the $O(n^{-\frac{1}{2}})$ rate of convergence given by the central limit theorem.

In high dimensional spaces this rate is often superior to those given by deterministic methods. The initial work to obtain the estimates $\hat{\mu}_j$ gives us a variance reduction so that we need not run an excessive amount of simulations before our particle hits the target. At the same time, we retain the superior rate given by Monte Carlo methods.

To obtain estimates for $\hat{\mu}_j$, we can run simulations starting from each state analogous to Algorithm 3.1 for finite state problems.

4.4 Computational Issues

We have already noted that the finite approximation (4.16) to the function μ can not be used in practice as it requires knowledge of the solution. Even the approximation (4.18) will be difficult or impossible to implement in practice. Recall

$$q(\mathbf{x}, \mathbf{y}, \hat{\mu}) = \frac{p(\mathbf{x}, \mathbf{y}) \cdot [\, s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y}) \,]}{p(\mathbf{x}, \Delta) \cdot s(\mathbf{x}, \Delta) + \int_{\mathcal{S}} p(\mathbf{x}, \mathbf{z}) \cdot [\, s(\mathbf{x}, \mathbf{z}) + \hat{\mu}(\mathbf{z}) \,]\, d\mathbf{z}}. \tag{4.19}$$

Now we can generate random variates using the acceptance-rejection method (see Algorithm 3.4 of Ripley [18]) without computing the integral

$$\int_S p(\mathbf{x},\mathbf{z})\cdot[\,s(\mathbf{x},\mathbf{z})+\hat{\mu}(\mathbf{z})\,]\,d\mathbf{z}$$

explicitly. However, when we calculate the likelihood ratio we will need to evaluate the exact value of $q(\mathbf{x},\mathbf{y},\hat{\mu})$. The function $\hat{\mu}$ is constant on each cell C_j, but we will not typically be able to compute

$$\int_{C_j} p(\mathbf{x},\mathbf{z})\cdot s(\mathbf{x},\mathbf{z})\,d\mathbf{z} \quad \text{or} \quad \int_{C_j} p(\mathbf{x},\mathbf{z})\,d\mathbf{z}$$

without error. If we replace it with a numerical approximation we introduce a bias to our estimator.

Instead, we can take

$$\tilde{q}(\mathbf{x},\mathbf{y},\hat{\mu}) = \frac{p(\mathbf{x}_i,\mathbf{x}_j)\cdot[\,s(\mathbf{x}_i,\mathbf{x}_j)+\hat{\mu}(\mathbf{x}_j)\,]}{p(\mathbf{x}_i,\Delta)\cdot s(\mathbf{x}_i,\Delta)+\sum_l p(\mathbf{x}_i,\mathbf{x}_l)\cdot[\,s(\mathbf{x}_i,\mathbf{x}_l)+\hat{\mu}(\mathbf{x}_l)\,]} \tag{4.20}$$

where i and j are chosen so that $\mathbf{x}\in C_i$ and $\mathbf{y}\in C_j$. Note that $\tilde{q}(\mathbf{x},\mathbf{y},\hat{\mu})$ can be computed explicitly. We might also consider the better approximation

$$\tilde{q}(\mathbf{x},\mathbf{y},\hat{\mu}) = \frac{p(\mathbf{x},\mathbf{x}_j)\cdot[\,s(\mathbf{x},\mathbf{x}_j)+\hat{\mu}(\mathbf{x}_j)\,]}{p(\mathbf{x},\Delta)\cdot s(\mathbf{x},\Delta)+\sum_l p(\mathbf{x},\mathbf{x}_l)\cdot[\,s(\mathbf{x},\mathbf{x}_l)+\hat{\mu}(\mathbf{x}_l)\,]}. \tag{4.21}$$

The importance sampling scheme (4.20) has the advantage that the denominator terms may be computed and stored in advance of the simulation runs. This will require d units of storage. The denominator in (4.21) depends on the exact value of $\mathbf{x}$ and will need to be computed at each transition of each simulation run. We might hope that using (4.20) or (4.21) in place of (4.19) will still achieve the rate in (4.17), but as we shall see in the next section this need not always be the case.

In continuous problems we can not store the transition kernel $p(\mathbf{x},\mathbf{y})$ explicitly as it contains an infinite amount of information. We need to define it implicitly. For

example, the transistion mechanism for the scatter of the particle may be given by implicitly defining the probability laws for the particle's scatter angle, velocity, and distance until the next collision as a function of the current state $\mathbf{x}$. For a given pair $(\mathbf{x}, \mathbf{y})$ the transition density $p(\mathbf{x}, \mathbf{y})$ can be calculated from these quantities rather than storing a matrix $\mathbf{P}$ as was done in the finite state case.

Since we do not store the quantities $p(\mathbf{x}, \mathbf{y})$ explicitly, we might hope to also avoid using $O(d^2)$ amount of storage for the kernel $q(\mathbf{x}, \mathbf{y}, \hat{\mu})$. We can reduce this to $O(d)$ if we use (4.20) and store only the denominators. The numerator terms can be computed as they are needed. Note this could have also been done in the finite state case, but we were already storing the matrix $\mathbf{P}$ so this would not have reduced our overall storage requirements to $O(d)$.

We can speed up the generation of sample paths by using the acceptance-rejection technique given in Algorithm 3.4 of Ripley [18]. Here we select the next state, $\mathbf{y}$, by some probability measure, $\nu(\mathbf{x}, \mathbf{y})$, that we can generate implicitly. For example, we may use the true measure $p(\mathbf{x}, \mathbf{y})$, or a uniform measure on $\mathcal{S}$, or perhaps a mixture of the two. Then we accept the choice, $\mathbf{y}$, with a probability proportional to the ratio of $q(\mathbf{x}, \mathbf{y}, \hat{\mu})$ to $\nu(\mathbf{x}, \mathbf{y})$. If the choice is rejected, we repeat the process until the value $\mathbf{y}$ is accepted.

Generating $\mathbf{y}$ by a probability measure implicitly defined such as $p(\mathbf{x}, \mathbf{y})$, or uniform on $\mathcal{S}$ will typically be much faster than generating $\mathbf{y}$ according to $q(\mathbf{x}, \mathbf{y}, \hat{\mu})$ directly. To generate $\mathbf{y}$ under the true $p(\mathbf{x}, \mathbf{y})$, we may only have to generate a random angle, velocity, and inter-collision distance. To generate $\mathbf{y}$ directly under $q(\mathbf{x}, \mathbf{y}, \hat{\mu})$, which will not generally have any such structure, we will have to use a "brute force" method of generating a uniform $(0, 1)$ variate, U, and searching through the cells one at a time until

$$\sum_{l=1}^{j} q(\mathbf{x}, \mathbf{x}_j, \hat{\mu}) \cdot \int_{C_j} d\mathbf{z} \geq U$$

and then choosing $\mathbf{y}$ uniformly on C_j. This will take $O(d)$ amount of work for each new state selected.

The expected amount of work required for acceptance-rejection method is equal to

$$\sup_{\mathbf{y}\in\mathcal{S}} \frac{q(\mathbf{x},\mathbf{y},\hat{\mu})}{\nu(\mathbf{x},\mathbf{y})}.$$

We must try to find a measure, $\nu(\mathbf{x},\mathbf{y})$, with the above quantity reasonably small, from which we can easily generate random variates. As we impose finer and finer partitions on the state space $\mathcal{S}$, $q(\mathbf{x},\mathbf{y},\hat{\mu})$ will approach the perfect importance scheme $q(\mathbf{x},\mathbf{y},\mu)$. The acceptance-rejection method will require essentially the same amount of work as if we were generating variates from $q(\mathbf{x},\mathbf{y},\mu)$ by acceptance-rejection. The amount of work required to select new states, therefore, should not be sensitive to the number of states, d.

4.5 Sample Problems

In this section model problems with continuous state spaces are given. In Section 4.5.1 we take another look at the problem introdued in Section 3.6.2, attacking it directly in the continuous space rather than approximating it with a finite state problem. This removes the bias, giving us a reliable estimator for the expected score of the particle in continuous space. In Section 4.5.2 a two dimensional problem with two different shielding media is introduced. Unlike the previous problem, no analytic solution is readily apparent so that our algorithm will have to "learn" which are the important states much like Algorithm 3.1.

4.5.1 A One Dimensional Problem Revisited

Let us consider the shielding problem in Section 3.6.2. Here, a particle initially colliding at $x = 0$ in one dimensional space survives the collision with probability 0.6.

If it survives, it travels an exponential (1) distance to the next collision. The particle successfully penetrates the shield if it crosses the level $x = 50$ before it is absorbed.

Here we have $\mathcal{S} = [0, 50]$ with

$$p(x, y) = \begin{cases} 0.6 \cdot \exp\{-(y - x)\} & \text{if} \quad x \leq y < 50; \\ 0 & \text{otherwise.} \end{cases}$$

We consider the particle to have penetrated the shield when $y > 50$. To keep $\mathcal{S}$ compact we can alias all values greater than 50 into the state 50 itself. Thus, for a particle at state x, there will be an atom of probability mass

$$0.6 \cdot \exp\{-(50 - x)\}$$

at the value $y = 50$. We then take

$$s(x, y) = \begin{cases} 1 & \text{if} \quad y = 50; \\ 0 & \text{otherwise.} \end{cases}$$

and

$$p(50, \Delta) = 0$$

to terminate the process once the particle has penetrated.

Note that this does not quite conform to the setup in Section 4.1 since we do not allow for atoms. Since we reqired the transition kernel to have a density with respect to Lebesgue measure, no one point should have a positive probability. We can, of course, artificially manipulate this problem into that form by aliasing values greater than 50 into the interval, say $[50, 51]$. That is, once a particle has penetrated the shield, we artificially place it uniformly on the interval $[50, 51]$ and give it a score of 1.

Thus, we can state this problem in accordance with Section 4.1 by

$$p(x, y) = \begin{cases} 0.6 \cdot \exp\{-(y - x)\} & \text{if} \quad x \leq y < 50 \\ 0.6 \cdot \exp\{-(50 - x)\} & \text{if} \quad x < 50 \leq y \leq 51; \\ 0 & \text{otherwise.} \end{cases}$$

and

$$s(x,y) = \begin{cases} 1 & \text{if } 50 \le y \le 51; \\ 0 & \text{otherwise.} \end{cases}$$

When we are actually running simulations on this problem we would prefer the first formulation as it is more natural. This second formulation is merely to show that the problem falls into the setup of Section 4.1. How the endpoint is treated is irrelevant since there is no uncertainty about that state's importance.

Recall that an analytic solution to this problem was given as

$$\mu(x) = 0.6 \cdot \exp\{-0.4 \cdot (50 - x)\}$$

so that

$$\mu(0) \approx 1.24 \cdot 10^{-9}.$$

If we choose a value d for the number of cells in the partition we wish to impose on $S = [0, 50]$, then we can take

$$C_j = \left[\frac{50 \cdot (j-1)}{d}, \frac{50 \cdot j}{d}\right]$$

as in Figure 3.7.

Note here that we have an analytic form for

$$\int_{C_j} p(\mathbf{x}, \mathbf{z})\, d\mathbf{z}$$

so that the importance scheme (4.19) could be used directly in this problem. However, (4.20) was used to verify that the rate (4.17) is achieved even with this approximation. Since the solution is known is this problem, simulations were run with $\hat{\mu} = \mu$, no "learning" process was attempted here. In the problem given in the next section, the solution is not apparent so that some learning process will be required to obtain $\hat{\mu}$. The number of cells varied from $d = 5000$ to 30,000 in increments of 5000. For each value of d, 10,000 simulations were run using importance sampling with (4.20). Results are given in Figure 4.1. The sample standard deviation is plotted against

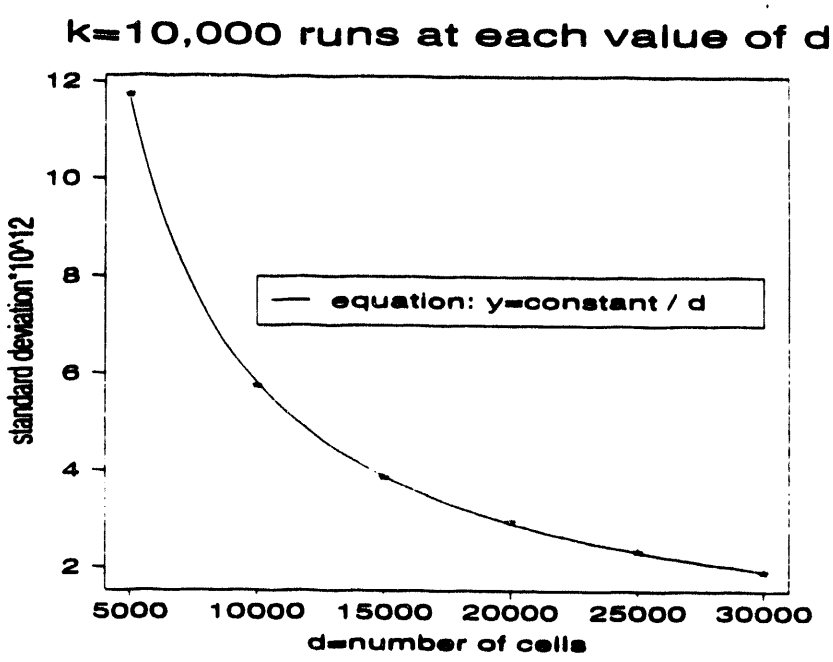


Figure 4.1: Standard deviation is inversely proportional to number of cells in the partition.

number of cells in the partition. A curve inversely proportional to the number of cells is fit to the data suggesting that the rate (4.17) is indeed achieved.

A slight modification to (4.20) was necessary to achieve this rate. When the particle is at state x in the j^{th} interval, C_j we must allow for the possibility that the next state, y, falls in the same interval. Now the particle can not go backwards in this model, so only the portion of C_j to the right of x is eligible for the next transition. If we use (4.20) directly we see that the density $\tilde{q}(x, y, \hat{\mu})$ is proportional to the area to the right of x in C_j. This inflates the variance since when x is close to the right endpoint of C_j, the likelihood ratio

$$\frac{p(x, y)}{\tilde{q}(x, y, \hat{\mu})}$$

is large. The interval C_j containing the current state x is over sampled in (4.20). Its probability should be reduced by the fraction of the interval C_j which is to the right of x. Once this was done the results concurred with (4.17). This, of course would not

be an issue if (4.19) were used directly.

4.5.2 A Two Dimensional Problem

Consider the transport problem depicted in Figure 4.2. A particle starts at the source at the point (25, 0) in two dimensional space. A target occupying the line segment from (20, 50) to (30, 50) is protected by a lead shield. However, an air duct runs though the shield as indicated. Distances travelled between collisions are exponential with cross section (hazard rate) $\lambda = 5$ in the lead, and $\lambda = 1$ in the air duct. A particle is absorbed during a collision with probability 0.9, and scattering angles are uniform $(0, 2\pi)$. We want to estimate the probability that the particle reaches the target before it is absorbed. No attempt to model the particle's energy level was made here.

The arrows in Figure 4.2 denote one path a simulated particle took to reach the target. Although a real particle would show no propensity to stay in the air duct, the simulated particle has "learned" that its best chance for hitting the target is to climb up the air duct getting as close to the target as possible, and then penetrating the remainder of the lead shield.

To perform importance sampling the space was split into $d = 10,000$ cells on an evenly spaced 100×100 grid. Suppose the points $\mathbf{x}$ and $\mathbf{z}$ are both in the air duct. We can transform to polar coordinates with $R = \|\mathbf{z} - \mathbf{x}\|$ and Φ being the angle formed by the line segment $\overline{\mathbf{xz}}$ and a horizontal line. We have

$$Pr\{R \in dr \quad \text{and} \quad \Phi \in d\phi\} = \frac{0.1}{2\pi} \cdot \exp\{-r\}.$$

Translating back to Cartesian space and multplying by the Jacobian, r^{-1}, gives

$$p(\mathbf{x}, \mathbf{z}) = \frac{0.1}{2\pi} \cdot \frac{1}{\|\mathbf{z} - \mathbf{x}\|} \cdot \exp\{-\|\mathbf{z} - \mathbf{x}\|\}. \tag{4.22}$$

Because the transition probabilities are given in terms of inter-collision distances and scattering angles, the problem is more naturally expressed in polar coordinates.

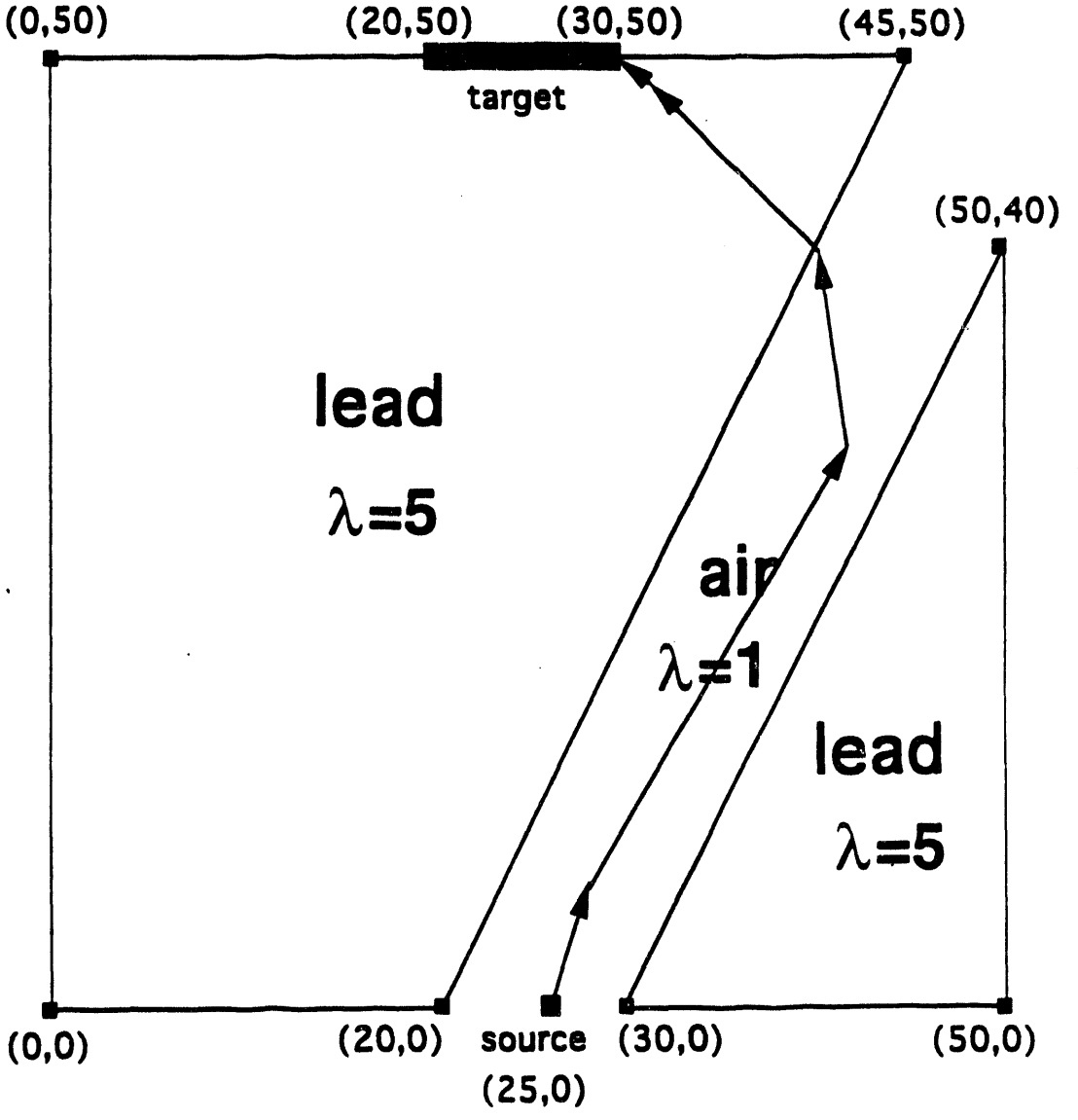


Figure 4.2: A two dimensional transport problem. Cross sections are $\lambda = 5$ in the lead, and $\lambda = 1$ in the air. Absorption probability is 0.9, and the scattering angle is uniform $(0, 2\pi)$.

There is no analytic form for the transition probability to a Cartesian cell, and so we can not use (4.19) directly. Similar results follow when one or both of the points are in the lead shield.

We may try to use (4.20) instead as was done in the previous problem. Unfortunately, this would not yield the same nice results here. If we take $\{C_i\}_{i=1}^d$ to be the Cartesian cells given by our grid, then the importance sampling scheme (4.20) is constant on these cells. Starting a particle at a point $\mathbf{x}$ in C_i, let us condition on the next point, $\mathbf{z}$, being in the same cell. Under (4.20) the conditional distribution would be uniform on this cell. Consider the square of the likelihood ratio

$$\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{x},\mathbf{z})}\right)^2 \propto \frac{1}{\|\mathbf{z}-\mathbf{x}\|^2} \cdot \exp\{-2 \cdot \|\mathbf{z}-\mathbf{x}\|\}.$$

Fixing $\mathbf{x}$ and taking $\mathbf{z}$ uniform on C_i we see that this random variable has infinite expectation. That is, the estimator from this importance sampling scheme has infinite variance no matter how fine the Cartesian partition $\{C_i\}_{i=1}^d$ may be.

As mentioned before, this problem is more naturally expressed in polar coordinates. We avoid the problem of the Jacobian, r^{-1}, blowing up the variance if we take q to be constant on polar cells rather than Cartesian cells. For each cell C_i with center $\mathbf{x}_i$ in our Cartesian partition, let $\{C_j'\}_{j=1}^d$ be a partition of $\mathcal{S}$ by polar coordinates with the reference point, $(r = 0)$, being $\mathbf{x}_i$. That is, C_j' is the set of points in $\mathcal{S}$ whose distance from $\mathbf{x}$ lies in a certain range, and whose angle, ϕ, from $\mathbf{x}$ lies in a certain range. Note each cell in the Cartesian partition $\{C_i\}_{i=1}^d$ has its own partition of $\mathcal{S}$ in polar coordinates.

Suppose for each cell C_i in the Cartesian partition we have an estimate, $\hat{\mu}_i$, of the probability of reaching the target for a particle starting in that cell. Let $\mathbf{x}_j'$ be the center of C_j'. We take

$$q(\mathbf{x},\mathbf{y},\hat{\mu}) = \frac{p(\mathbf{x}_i,\mathbf{x}_j)\cdot\hat{\mu}_{mij}}{p(\mathbf{x}_j,T)+\sum_l p(\mathbf{x}_i,\mathbf{x}_l)\cdot\hat{\mu}_{lij}} \tag{4.23}$$

AIIM

Association for Information and Image Management

1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202

Centimeter

Inches

MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

2 of 2

where i is chosen so that $\mathbf{x} \in C_i$, j so that $\mathbf{y} \in C'_j$, m_{ij} so that $\mathbf{x}'_j \in C_{m_{ij}}$, and l_{ij} so that $\mathbf{x}''_j \in C_{l_{ij}}$. The quantity $p(\mathbf{x}_j, T)$ represents the probability of hitting the target directly from $\mathbf{x}_j$.

Similar to (4.20), the denominator terms may be calculated and stored prior to the simulation runs. This will require d units of storage. Note that the denominator term itself gives a good estimate of μ_i when our previous estimates $\{\hat{\mu}_i\}$ are "close" to the true $\mu(\cdot)$. If we deterministically perform the iteration

$$\hat{\mu}_i^{(n+1)} = p(\mathbf{x}_j, T) + \sum_l p(\mathbf{x}_i, \mathbf{x}_l) \cdot \hat{\mu}_{l_{ij}}^{(n)} \tag{4.24}$$

we obtain estimates $\hat{\mu}^{(n)}$ converging to the solution to a discretized version of this transport problem. We may prefer to use this method as a means of obtaining estimates $\hat{\mu}_i$ for the simulation runs rather than something analogous to Algorithm 3.1. To parallel Algorithm 3.1 simulation runs would have to be made starting from each of the 10,000 cells. This would take a relatively large amount of computing time. The iteration (4.24) would have to be calculated for the denominator terms of $q(\mathbf{x}, \mathbf{y}, \hat{\mu})$ as a first step prior to the simulation runs in each iteration. The deterministic recursion (4.24) alone is much quicker and possibly more stable.

A simple heuristic was used to supply (4.24) with the initial guess, $\hat{\mu}^{(0)}$. Starting from a point $\mathbf{x}$, the particle may head directly towards the target or may climb up the air duct to the point, (40, 40) and then head towards the target. In a one dimensional problem with exponential step sizes, we know that total penetration before absorption is itself an exponential random variable. Its mean is increased by a factor of the reciprocal of the absorption probability (see Section 3.6.2). Thus, we might approximate the probability of a particle starting from $\mathbf{x}$ hitting the target by

$$p(\mathbf{x}, (25, 50))^{0.9}$$

. Similarly, we might approximate the probability of getting to (40, 40) first and then

reaching the target by

$$p(\mathbf{x}, (40, 40))^{0.9} \cdot p((40, 40), (25, 50))^{0.9}.$$

The initial guess was given by

$$\hat{\mu}_j^{(0)} = \max\Big(p(\mathbf{x}_j, (25, 50))^{0.9}, p(\mathbf{x}_j, (40, 40))^{0.9} \cdot p((40, 40), (25, 50))^{0.9}\Big).$$

Twenty iterations of (4.24) were run followed by $k = 100,000$ simulation runs using (4.23). Using the central limit theorem a 95% confidence interval for the probability of reaching the target starting at (25,0) was given by

$$(4.382 \pm 0.475) \cdot 10^{-54}.$$

Appendix A

Differentiating Functionals

Take $\mathcal{S}$ to be a compact subset of a Euclidean space and let $\mathcal{M}$ be an open collection of functions, $\hat{\mu}$, mapping $\mathcal{S} \to \mathcal{R}$. Let $f : \mathcal{M} \to \mathcal{R}$ be a functional mapping functions into real numbers.

Definition A.1 *The functional $f(\cdot)$ is considered to be differentiable at $\hat{\mu}$ if there exists a mapping $\nabla f : \mathcal{M} \times \mathcal{S} \to \mathcal{R}$ such that for a function h*

$$f(\hat{\mu} + h) = f(\hat{\mu}) + \int_{\mathcal{S}} \nabla f(\hat{\mu}, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} + o(\|h\|). \tag{A.1}$$

Here,

$$\|h\| = \text{ess sup}\, h = \inf\Big\{M : L\{\mathbf{y} : h(\mathbf{y}) \geq M\} > 0\Big\}$$

where L denotes Lebesgue measure on $\mathcal{S}$. The functional $o(\cdot)$ is taken so that for any sequence of functions $\{h_n\}$ with $\|h_n\| \to 0$ we have

$$\lim_{n \to \infty} \frac{o(h_n)}{\|h_n\|} = 0. \tag{A.2}$$

The following result is an immediate consequence of Definition A.1.

Theorem A.1 *If f and g are differentiable functionals on $\mathcal{M}$ then $f + g$ and $f \cdot g$ are differentiable and*

$$\nabla[f + g] = \nabla f + \nabla g \quad \textit{and} \quad \nabla[f \cdot g] = f \cdot \nabla g + \nabla f \cdot g.$$

We will need the following analog of the mean value theorem to derive our differentiation rules.

Theorem A.2 Mean Value Theorem.
Suppose f is a differentiable functional on $\mathcal{M}$. If $\hat{\mu} \in \mathcal{M}$ and $\|h\|$ is small enough so that $\hat{\mu} + h \in \mathcal{M}$, then there exists a $\hat{\mu}^* \in \mathcal{M}$ so that

$$f(\hat{\mu} + h) - f(\hat{\mu}) = \int_S \nabla f(\hat{\mu}^*, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y}.$$

Proof: Let b be a function $[0,1] \to \mathcal{R}$ given by

$$b(\alpha) = [f(\hat{\mu} + h) - f(\hat{\mu})] \cdot \alpha - f(\hat{\mu} + \alpha \cdot h).$$

For $\beta \in \mathcal{R}$ sufficiently small,

$$f(\hat{\mu} + (\alpha + \beta) \cdot h) = f([\hat{\mu} + \alpha \cdot h] + \beta \cdot h) =$$

$$f(\hat{\mu} + \alpha \cdot h) + \beta \cdot \int_S \nabla f(\hat{\mu} + \alpha \cdot h, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} + o(\beta \cdot h).$$

Thus, b is differentiable on $(0,1)$ in the regular sense and

$$b'(\alpha) = [f(\hat{\mu} + h) - f(\hat{\mu})] - \int_S \nabla f(\hat{\mu} + \alpha \cdot h) \cdot h(\mathbf{y})\, d\mathbf{y}.$$

Note that

$$b(0) = b(1) = -f(\hat{\mu}).$$

Thus, b must obtain either a local minimum or maximum for some $\alpha^* \in (0,1)$ where we must have $b'(\alpha^*) = 0$. That is

$$f(\hat{\mu} + h) - f(\hat{\mu}) = \int_S \nabla f(\hat{\mu}^*, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y}$$

where $\hat{\mu}^* = \hat{\mu} + \alpha^* \cdot h$. ∎

The next result is used in the proof of Theorem 4.1.

Theorem A.3 *Let θ be a function $\mathcal{S} \times \mathcal{R} \to \mathcal{R}$ which is differentiable in the regular sense with respect to its second element. Let $g(\mathbf{y}, \hat{\mu})$ be a mapping from $\mathcal{S} \times \mathcal{M} \to \mathcal{R}$ such that for each fixed $\mathbf{y} \in \mathcal{S}$, the functional $g(\mathbf{y}, \cdot)$ is differentiable in the sense of (A.1), having derivative $\nabla g(\mathbf{y}, \hat{\mu}, \mathbf{z})$. If there exists a a nonnegative density $\nu(\mathbf{y})$ with*

$$\int_{\mathcal{S}} \nu(\mathbf{y})\, d\mathbf{y} < \infty$$

such that the quantities

$$\nu^{-1}(\mathbf{y}) \cdot \left[|\theta(\mathbf{y}, \hat{\mu}(\mathbf{y}))| + |\theta'(\mathbf{y}, \hat{\mu}(\mathbf{y}))| + |g(\mathbf{y}, \hat{\mu})| + |\nabla g(\mathbf{y}, \hat{\mu}, \mathbf{z})| \right]$$

are uniformly bounded in $(\mathbf{y}, \mathbf{z}) \in \mathcal{S}^2$ and $\hat{\mu} \in \mathcal{M}$, then the functional

$$f(\hat{\mu}) = \int_{\mathcal{S}} \theta(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot g(\mathbf{y}, \hat{\mu})\, d\mathbf{y}$$

is differentiable and

$$\nabla f(\hat{\mu}, \mathbf{y}) = \theta'(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot g(\mathbf{y}, \hat{\mu}) + \int_{\mathcal{S}} \theta(\mathbf{z}, \hat{\mu}(\mathbf{z})) \cdot \nabla g(\mathbf{z}, \hat{\mu}, \mathbf{y})\, d\mathbf{z}.$$

Proof: Let h be a functional with $\|h\|$ small enough so that $\hat{\mu} + h \in \mathcal{M}$. Since θ is differentiable in its second element, we have

$$\theta(\mathbf{y}, \hat{\mu}(\mathbf{y}) + h(y)) = \theta(\mathbf{y}, \hat{\mu}(\mathbf{y})) + \theta'(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot h(\mathbf{y}) + o_\theta(\mathbf{y}, h(\mathbf{y}))$$

where o_θ is a function $\mathcal{S} \times \mathcal{R} \to \mathcal{R}$ such that for each fixed $\mathbf{y}$

$$\lim_{t \to 0} \frac{o_\theta(\mathbf{y}, t)}{t} = 0.$$

Since g is differentiable for each fixed $\mathbf{y}$, we have

$$g(\mathbf{y}, \hat{\mu} + h) = g(\mathbf{y}, \hat{\mu}) + \int_{\mathcal{S}} \nabla g(\mathbf{y}, \hat{\mu}, \mathbf{z}) \cdot h(\mathbf{z})\, d\mathbf{z} + o_g(\mathbf{y}, h)$$

where for fixed $\mathbf{y}$, $o_g(\mathbf{y}, \cdot)$ satisfies (A.2).

Thus, $f(\hat{\mu}+h) =$

$$f(\hat{\mu}) + \int_{\mathcal{S}} \left[\theta'(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot g(\mathbf{y}, \hat{\mu}) + \int_{\mathcal{S}} \theta(\mathbf{z}, \hat{\mu}(\mathbf{z})) \cdot \nabla g(\mathbf{z}, \hat{\mu}, \mathbf{y})\, d\mathbf{z}\right] \cdot h(y)\, d\mathbf{y} + r(h).$$

It remains to be shown that r satisfies (A.2). Consider the first remainder term

$$r_1(\hat{\mu}, h) = \int_{\mathcal{S}} g(\mathbf{y}, \hat{\mu}) \cdot o_\theta(\mathbf{y}, h(y))\, d\mathbf{y}.$$

Let $\{h_n\}$ be a sequence of functionals with $\|h_n\| \to 0$. Then for almost every fixed $\mathbf{y} \in \mathcal{S}$, $h_n(\mathbf{y}) \to 0$ and hence

$$\frac{o_\theta(\mathbf{y}, h_n(\mathbf{y}))}{h_n(y)} \to 0 \quad \text{as} \quad n \to \infty$$

(take $0/0 = 0$). By the mean value theorem (Theorem 5.10 of Rudin [20])

$$\frac{o_\theta(\mathbf{y}, h_n(\mathbf{y}))}{h_n(\mathbf{y})} = \theta'(\mathbf{y}, \alpha)$$

for some $\alpha \in \mathcal{R}$ and so the quantities

$$\frac{o_\theta(\mathbf{y}, h_n(\mathbf{y}))}{\nu(\mathbf{y}) \cdot h_n(\mathbf{y})}$$

are uniformly bounded except possibly on a set of Lesbesgue measure zero. Now $\|h_n\| \geq h_n(\mathbf{y})$ for almost every $\mathbf{y} \in \mathcal{S}$ so by the dominated convergence theorem

$$\lim_{n\to\infty} \frac{r_1(\hat{\mu}, h_n)}{\|h_n\|} = 0.$$

Consider the remainder term

$$r_2(h) = \int_{\mathcal{S}} \theta(\mathbf{y}) \cdot o_g(\mathbf{y}, h).$$

Let $\{h_n\}$ be a sequence of functions as specified above. By Theorem A.2

$$o_g(\mathbf{y}, h_n) = \int_{\mathcal{S}} \nabla g(\mathbf{y}, \hat{\mu}^*, \mathbf{z}) \cdot h(\mathbf{y})\, d\mathbf{y}$$

for some $\hat{\mu}^* \in \mathcal{M}$. Since $\mathcal{S}$ is compact, it has finite Lebesgue measure and so the quantities

$$\frac{o_g(\mathbf{y}, h_n)}{\nu(\mathbf{y}) \cdot \|h_n\|}$$

are uniformly bounded except possibly on a set of Lebesgue measure zero. By (A.2), for almost every fixed $\mathbf{y} \in \mathcal{S}$,

$$\frac{o_g(\mathbf{y}, h_n)}{\|h_n\|} \to 0$$

and so the dominated convergence theorem implies

$$\lim_{n \to \infty} \frac{r_2(\hat{\mu}, h_n)}{\|h_n\|} = 0.$$

Analogous arguments apply to the other remainder terms. ∎

As an example let us apply this theorem to the function appearing in Section 4.2.

$$f(\mathbf{x}, \hat{\mu}) = \frac{[p(\mathbf{x}, \Delta) \cdot s(\mathbf{x}, \Delta)]^2}{q(\mathbf{x}, \Delta, \hat{\mu})} + \int_{\mathcal{S}} \frac{\left(p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})]\right)^2}{q(\mathbf{x}, \mathbf{y}, \hat{\mu})} d\mathbf{y} - [\mu(\mathbf{x})]^2 =$$

$$c(\mathbf{x}, \hat{\mu}) \cdot \left[p(\mathbf{x}, \Delta) \cdot s(\mathbf{x}, \Delta) + \int_{\mathcal{S}} \frac{p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y})]^2}{s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})} d\mathbf{y} - [\mu(\mathbf{x})]^2 \right]$$

where $c(\mathbf{x}, \mathbf{y})$ is as defined in the proof of Theorem 4.1. Note here, we fix $\mathbf{x} \in \mathcal{S}$, considering $f(\mathbf{x}, \cdot)$ as a functional on $\mathcal{M}$.

Let us first focus on the part

$$f^*(\mathbf{x}, \hat{\mu}) = \int_{\mathcal{S}} \frac{p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y})]^2}{s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})} d\mathbf{y}.$$

Taking

$$g(\mathbf{y}, \hat{\mu}) = 1, \quad \theta(\mathbf{y}, t) = \frac{p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y})]^2}{s(\mathbf{x}, \mathbf{y}) + t} \quad \text{and} \quad \nu(\mathbf{y}) = p(\mathbf{x}, \mathbf{y}) + 1$$

we apply Theorem A.3 to get

$$\nabla f^*(\mathbf{x}, \hat{\mu}, \mathbf{y}) = -\frac{p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y})]^2}{[s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})]^2}.$$

Note $\nabla g = 0$ in this case. Now apply Theorem A.1 and use $\nabla c(\mathbf{x}, \hat{\mu}, \mathbf{y}) = p(\mathbf{x}, \mathbf{y})$ to get

$$\nabla f(\mathbf{y}, \hat{\mu}) = c(\mathbf{x}, \mathbf{y}) \cdot \nabla f^*(\mathbf{x}, \hat{\mu}, \mathbf{y}) + \nabla c(\mathbf{x}, \hat{\mu}, \mathbf{y}) \cdot f(\mathbf{x}, \hat{\mu}) =$$
$$-c(\mathbf{x}, \hat{\mu}) \cdot \frac{p(\mathbf{x}, \mathbf{y}) \cdot [s(\mathbf{x}, \mathbf{y}) + \mu(\mathbf{y})]^2}{[s(\mathbf{x}, \mathbf{y}) + \hat{\mu}(\mathbf{y})]^2} + p(\mathbf{x}, \mathbf{y}) \cdot f(\mathbf{x}, \hat{\mu}).$$

The following result specifies conditions under which we can differentiate infinite sums.

Theorem A.4 *Suppose $\{f_n\}$ is a sequence of differentiable functionals on $\mathcal{M}$ with*

$$f(\hat{\mu}) = \sum_{m=0}^{\infty} f_m(\hat{\mu}).$$

If the sum

$$\sum_{m=0}^{\infty} \int_S \nabla f_m(\hat{\mu}, \mathbf{y})\, d\mathbf{y}$$

converges to a finite limit uniformly on $\mathcal{M}$, then f is differentiable and

$$\nabla f(\hat{\mu}, \mathbf{y}) = \sum_{m=0}^{\infty} \nabla f_m(\hat{\mu}, \mathbf{y}) d\mathbf{y}.$$

Proof: Let $\{h_n\}$ be a sequence of functions such that $\|h_n\| \to 0$. Let $\epsilon > 0$. There exists an integer M such that

$$\sum_{m=M}^{\infty} \left| \int_S \nabla f_m(\hat{\mu}, \mathbf{y})\, d\mathbf{y} \right| < \frac{\epsilon}{3}$$

for all $\hat{\mu} \in \mathcal{M}$. By Theorem A.2, for each pair (m, n) there exists a $\hat{\mu}^*_{m,n}$ such that

$$f_m(\hat{\mu} + h_n) - f_m(\hat{\mu}) = \int_S \nabla f(\hat{\mu}^*_{m,n}, \mathbf{y}) \cdot h_n(\mathbf{y})\, d\mathbf{y}.$$

Thus, for $m \geq M$

$$\sum_{m=M}^{\infty} \frac{|f_m(\hat{\mu} + h_n) - f_m(\hat{\mu})|}{\|h_n\|} < \frac{\epsilon}{3}$$

for all n.

By Definition A.1, we can choose N large enough so that for $n \geq N$

$$\|h_n\|^{-1} \cdot \left| f_m(\hat{\mu} + h_n) - f_m(\hat{\mu}) - \int_S \nabla f_m(\hat{\mu}, \mathbf{y}) \cdot h_n(\mathbf{y}) \right| < \frac{\epsilon}{3^{m+2}} \quad m = 0, \ldots, M.$$

Then for $n \geq N$

$$\|h_n\|^{-1} \cdot \left| f(\hat{\mu} + h_n) - f(\hat{\mu}) - \int_S \left[\sum_{m=0}^{\infty} \nabla f_m(\hat{\mu}, \mathbf{y}) \right] d\mathbf{y} \right| \leq$$

$$\|h_n\|^{-1} \cdot \left[\sum_{m=1}^{M} \left| f_m(\hat{\mu}+h_n) - f_m(\hat{\mu}) - \int_S \nabla f_m(\hat{\mu}, \mathbf{y}) \cdot h_n(\mathbf{y}) \right| + \sum_{m=M}^{\infty} \left| f_m(\hat{\mu}+h_n) - f_m(\hat{\mu}) \right| + \right.$$

$$\left. \sum_{m=M}^{\infty} \left| \int_S \nabla f_m(\hat{\mu}, \mathbf{y}) \cdot h_n(\mathbf{y}) \, d\mathbf{y} \right| \right] \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

That is,

$$\lim_{n \to \infty} \|h_n\|^{-1} \cdot \left| f(\hat{\mu} + h_n) - f(\hat{\mu}) - \int_S \left[\sum_{m=0}^{\infty} \nabla f_m(\hat{\mu}, \mathbf{y}) \right] d\mathbf{y} \right| = 0$$

which establishes the theorem. ∎

Now we need to develop some notion of a second derivative. Unfortunately, we can not simply define it as the derivative of the derivative. This would be too restrictive. For example, consider the functional

$$f(\hat{\mu}) = \int_S \left[\hat{\mu}(\mathbf{y}) \right]^2 d\mathbf{y}.$$

Clearly, this is differentiable with

$$\nabla f(\hat{\mu}, \mathbf{y}) = 2\,\hat{\mu}(\mathbf{y}).$$

Note that for fixed $\mathbf{y} \in S$, $\nabla f(\cdot, \mathbf{y})$ is not differentiable since it depends on the function $\hat{\mu}$ solely through the value $\hat{\mu}(\mathbf{y})$. Here we see that we can not define a quantity $\nabla^2 f(\hat{\mu}, \mathbf{y}, \mathbf{z})$ so that

$$f(\hat{\mu} + h) = f(\hat{\mu}) + \int_S \nabla f(\hat{\mu}, \mathbf{y}) \, d\mathbf{y} + \frac{1}{2} \cdot \iint_{S^2} h(\mathbf{y}) \cdot \nabla^2 f(\hat{\mu}, \mathbf{y}, \mathbf{z}) \cdot h(\mathbf{z}) \, d\mathbf{y} \, d\mathbf{z} + o(\|h\|^2).$$

The quantity $\nabla^2 f$ is supposed to parallel a second derivative matrix in the discrete case. The problem here is that the second derivative matrix for the discrete analog

$$\mathbf{f}(\hat{\mu}) = \hat{\mu}^T \hat{\mu}$$

is diagonal. However, the "diagonal" elememts in our double integral have no contribution since the set $\{(\mathbf{y}, \mathbf{z}) : \mathbf{y} = \mathbf{z}\}$ has Lebesgue measure zero. We must add a second element $\nabla_D^2 f$ to incorporate the "diagonal" contributions of the second derivative.

Definition A.2 *A functional f is twice differentiable if in addition to being differentiable in the sense of (A.1), there exist mappings $\nabla^2 f : \mathcal{M} \times \mathcal{S}^2 \to \mathcal{R}$, and $\nabla_D^2 f : \mathcal{M} \times \mathcal{S} \to \mathcal{R}$ such that*

$$f(\hat{\mu} + h) = f(\hat{\mu}) + \int_{\mathcal{S}} \nabla f(\hat{\mu}, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} + \tag{A.3}$$

$$\frac{1}{2} \cdot \left[\iint_{\mathcal{S}^2} h(\mathbf{y}) \cdot \nabla^2 f(\hat{\mu}, \mathbf{y}, \mathbf{z}) \cdot h(\mathbf{z})\, d\mathbf{y}\, d\mathbf{z} + \int_{\mathcal{S}} \nabla_D^2 f(\hat{\mu}, \mathbf{y}) \cdot [h(\mathbf{y})]^2\, d\mathbf{y} \right] + o(\|h\|^2).$$

We see our example

$$f(\hat{\mu}) = \int_{\mathcal{S}} \left[\hat{\mu}(\mathbf{y})\right]^2 d\mathbf{y}$$

satisfies (A.3) with $\nabla^2 f(\hat{\mu}, \mathbf{y}, \mathbf{z}) = 0$, and $\nabla_D^2 f(\hat{\mu}, \mathbf{y}) = 2$. The following result is immediate.

Theorem A.5 *If f and g are functionals that are twice differentiable in the sense of (A.3), then $f + g$ and $f \cdot g$ are twice differentiable with*

$$\nabla^2 [f + g] = \nabla^2 f + \nabla^2 g$$

$$\nabla_D^2 [f + g] = \nabla_D^2 f + \nabla_D^2 g.$$

$$\nabla^2 [f \cdot g](\hat{\mu}, \mathbf{y}, \mathbf{z}) = f(\hat{\mu}) \cdot \nabla^2 g(\hat{\mu}, \mathbf{y}, \mathbf{z}) + \nabla^2 f(\hat{\mu}, \mathbf{y}, \mathbf{z}) \cdot g(\hat{\mu}) + 2 \nabla f(\hat{\mu}, \mathbf{y}) \cdot \nabla g(\hat{\mu}, \mathbf{z})$$

and

$$\nabla_D^2 [f \cdot g](\hat{\mu}, \mathbf{y}) = f(\hat{\mu}) \cdot \nabla_D^2 g(\hat{\mu}, \mathbf{y}) + \nabla_D^2 f(\hat{\mu}, \mathbf{y}) \cdot g(\hat{\mu}).$$

Here, Theorem A.2 is extended to a second derivative.

Theorem A.6 *Suppose f is a twice differentiable functional in the sense of (A.3). Let $\hat{\mu} \in \mathcal{M}$ and $\|h\|$ be small enough so that $\hat{\mu} + h \in \mathcal{M}$. Then there exists a $\hat{\mu}^* \in \mathcal{M}$ such that*

$$f(\hat{\mu} + h) - f(\hat{\mu}) = \int_S \nabla f(\hat{\mu}, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} +$$

$$\frac{1}{2} \cdot \left[\iint_{S^2} h(\mathbf{y}) \cdot \nabla^2 f(\hat{\mu}^*, \mathbf{y}, \mathbf{z}) \cdot h(\mathbf{z})\, d\mathbf{y}\, d\mathbf{z} + \int_S \nabla_D^2 f(\hat{\mu}^*, \mathbf{y}) \cdot [h(\mathbf{y})]^2\, d\mathbf{y} \right].$$

Proof: Let $b : [0, 1] \to \mathcal{R}$ be defined by

$$b(\alpha) = f(\hat{\mu} + \alpha \cdot h).$$

Then

$$b(\alpha + \beta) = f([\hat{\mu} + \alpha \cdot h] + \beta \cdot h) = f(\hat{\mu} + \alpha \cdot h) + \beta \cdot \int_S \nabla f(\hat{\mu} + \alpha \cdot h, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} +$$

$$\frac{\beta^2}{2} \cdot \left[\iint_{S^2} h(\mathbf{y}) \cdot \nabla^2 f(\hat{\mu} + \alpha \cdot h, \mathbf{y}, \mathbf{z}) \cdot h(\mathbf{z})\, d\mathbf{y}\, d\mathbf{z} + \int_S \nabla_D^2 f(\hat{\mu} + \alpha \cdot h, \mathbf{y}) \cdot [h(\mathbf{y})]^2\, d\mathbf{y} \right] +$$

$$o(\beta^2 \cdot \|h\|^2).$$

Therefore, the function b is twice differentiable with

and $$b'(\alpha) = \int_S \nabla f(\hat{\mu} + \alpha \cdot h, \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y}$$

$$b''(\alpha) = \iint_{S^2} h(\mathbf{y}) \cdot \nabla^2 f(\hat{\mu} + \alpha \cdot h, \mathbf{y}, \mathbf{z}) \cdot h(\mathbf{z})\, d\mathbf{y}\, d\mathbf{z} + \int_S \nabla_D^2 f(\hat{\mu} + \alpha \cdot h, \mathbf{y}) \cdot [h(\mathbf{y})]^2\, d\mathbf{y}.$$

Now apply Taylor's theorem to b. ∎

The following analog of Theorem A.3 allows us to compute second derivatives in the proof of Theorem 4.2.

Theorem A.7 *Let $\theta : \mathcal{S} \times \mathcal{R} \to \mathcal{R}$ be a mapping which is twice differentiable in the regular sense with respect to its second element. Let $g : \mathcal{S} \times \mathcal{M} \to \mathcal{R}$ be a mapping such that for each fixed $\mathbf{y} \in \mathcal{S}$, $g(\mathbf{y}, \cdot)$ is twice differentiable in the sense of (A.3). If there exists exists a a nonnegative density $\nu(\mathbf{y})$ with*

$$\int_{\mathcal{S}} \nu(\mathbf{y})\, d\mathbf{y} < \infty$$

such that the quantities

$$\nu^{-1}(\mathbf{y}) \cdot \Big[|\theta(\mathbf{y}, \hat{\mu}(\mathbf{y}))| + |\theta'(\mathbf{y}, \hat{\mu}(\mathbf{y}))| + |\theta''(\mathbf{y}, \hat{\mu}(\mathbf{y}))| +$$

$$|g(\mathbf{y}, \hat{\mu})| + |\nabla g(\mathbf{y}, \hat{\mu}, \mathbf{z})| + |\nabla^2 g(\mathbf{y}, \hat{\mu}, \mathbf{z}, \mathbf{w})| + |\nabla_D^2 g(\mathbf{y}, \hat{\mu}, \mathbf{z})| \Big]$$

are uniformly bounded in $(\mathbf{y}, \mathbf{z}, \mathbf{w}) \in \mathcal{S}^3$ and $\hat{\mu} \in \mathcal{M}$, then the functional

$$f(\hat{\mu}) = \int_{\mathcal{S}} \theta(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot g(\mathbf{y}, \hat{\mu})\, d\mathbf{y}$$

is twice differentiable with

$$\nabla^2 f(\hat{\mu}, \mathbf{y}, \mathbf{z}) = 2\theta'(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot \nabla g(\mathbf{y}, \hat{\mu}, \mathbf{z}) + \int_{\mathcal{S}} \theta(\mathbf{w}, \hat{\mu}(\mathbf{w})) \cdot \nabla^2 g(\mathbf{w}, \hat{\mu}, \mathbf{z}, \mathbf{y})\, d\mathbf{w}$$

and

$$\nabla_D^2 f(\hat{\mu}, \mathbf{y}) = \theta''(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot g(\mathbf{y}, \hat{\mu}) + \int_{\mathcal{S}} \theta(\mathbf{z}, \hat{\mu}(\mathbf{z})) \cdot \nabla_D^2 g(\mathbf{z}, \hat{\mu}, \mathbf{y})\, d\mathbf{z}.$$

Proof: Since θ is twice differentiable we have

$$\theta(\mathbf{y}, \hat{\mu}(\mathbf{y}) + h(\mathbf{y})) = \theta(\mathbf{y}, \hat{\mu}(\mathbf{y})) + \theta'(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot h(\mathbf{y}) + \theta''(\mathbf{y}, \hat{\mu}(\mathbf{y})) \cdot \frac{[h(\mathbf{y})]^2}{2} + o_\theta(\mathbf{y}, h(\mathbf{y}))$$

where o_θ satisfies

$$\lim_{t \to \infty} \frac{o_\theta(\mathbf{y}, t)}{t^2} = 0$$

for each fixed $\mathbf{y} \in \mathcal{S}$.

Similarly

$$g(\mathbf{y}, \hat{\mu} + h) = g(\mathbf{y}, \hat{\mu}) + \int_{\mathcal{S}} \nabla g(\mathbf{y}, \hat{\mu}, \mathbf{z}) \cdot h(\mathbf{z})\, d\mathbf{z} +$$

$$\frac{1}{2}\cdot\left[\iint_{S^2} h(\mathbf{z})\cdot\nabla^2 g(\mathbf{y},\hat{\mu},\mathbf{z},\mathbf{w})\cdot h(\mathbf{w})\,d\mathbf{w}\,d\mathbf{z} + \int_S \nabla_D^2 g(\mathbf{y},\hat{\mu},\mathbf{z})\cdot[h(\mathbf{z})]^2\,d\mathbf{z}\right] + o_g(\mathbf{y},h)$$

where for fixed $\mathbf{y}\in\mathcal{S}$

$$\lim_{\|h\|\to 0}\frac{o_g(\mathbf{y},h)}{\|h\|^2} = 0.$$

Now proceed as in the proof of Theorem A.3. ∎

We also need the following analog of Theorem A.4 in the proof of Theorem 4.2.

Theorem A.8 *Suppose $\{f_n\}$ is a sequence of twice differentiable functionals on $\mathcal{M}$ with*

$$f(\hat{\mu}) = \sum_{m=0}^{\infty} f_m(\hat{\mu}).$$

If the sums

$$\sum_{m=0}^{\infty}\int_S \nabla f_m(\hat{\mu},\mathbf{y})\,d\mathbf{y},\quad \sum_{m=0}^{\infty}\iint_{S^2}\nabla^2 f_m(\hat{\mu},\mathbf{y},\mathbf{z})\,d\mathbf{y}\,d\mathbf{z}\quad \textit{and}\quad \sum_{m=0}^{\infty}\int_S \nabla_D^2 f_m(\hat{\mu},\mathbf{y})\,d\mathbf{y}$$

all converge to a finite limit uniformly on $\mathcal{M}$, then f is twice differentiable with

$$\nabla^2 f(\hat{\mu},\mathbf{y},\mathbf{z}) = \sum_{m=0}^{\infty}\nabla^2 f_m(\hat{\mu},\mathbf{y},\mathbf{z})$$

and

$$\nabla_D^2 f(\hat{\mu},\mathbf{y}) = \sum_{m=0}^{\infty}\nabla_D^2 f_m(\hat{\mu},\mathbf{y}).$$

Proof: The argument is analogous to the proof of Theorem A.4, using Theorem A.6 in place of Theorem A.2. ∎

Appendix B

Fortran Code for Algorithm 3.1

```
c
c   This program runs uses an adaptive importance
c   sampling technique to simulate the expected
c   cumulative score of a Markov Chain subject
c   to absorption.  Each time the chain moves
c   from state i to j, the amount s(i,j) is added
c   to the total score.
c
c   To use this program create a file called fort.1
c   The first line should have three numbers:
c      # of states, # of simulations per iteration,
c      and # of iterations.
c    This should be followed by a n by n matrix of
c    transition probabilities with each row being
c    one line in the file.
c    This should be followed by an n by (n+1) matrix
c    of scores.(s(i,j) being the amount scored
     jumping from i to j, the last column
c    corresponding to absorption scores).
c    The last line of the file should contain initial
c    guesses for the expected scores. (n numbers).
c    Output is put in a file called fort.2
c
c    This program computes the exact solution.  At
c    each iteration the  actual error of the estimate
c    and the estimated standard deviation are
c    written to a file called fort.3
```

```
      implicit double precision(a-h,o-z)
      common/ibuff/nhit
      double precision p(100,101),s(100,101),score(100)
      double precision sd,s2(100),b(100,100),e(100)
      double precision oldscore(100)
      integer n,k,it,nhit(100)
      alpha=-1.d0
      do while((alpha.lt.0.d0).or.(alpha.gt.1.d0))
      write(*,*) 'enter value of alpha= wt given'
     &,'to newest runs'
      read(*,*) alpha
      if((alpha.lt.0.d0).or.(alpha.gt.1.d0))
     &   write(*,*) 'need 0<alpha<=1'
      end do
      beta=1.d0-alpha
      alpha2=alpha*alpha
      beta2=beta*beta
      write(2,5)
5     format('THIS IS ALGO3.1'/'INFORMATION IS USED',
     &' TO ESTIMATE EXP. SCORE FOR EVERY STATE THAT',
     &' GETS HIT IN THE RUN.'/)
      read(1,*) n,k,it
      do 10 i=1,n
         read(1,*)(p(i,j),j=1,n)
         p(i,n+1)=1.
         do 15 j=1,n
            p(i,n+1)=p(i,n+1)-p(i,j)
15       continue
         if(p(i,n+1).lt.1e-10) then
            if(p(i,n+1).gt. -1e-10) then
                 p(i,n+1)=0.
            else
                 write(*,16) i
                 write(2,16) i
16     format('ERROR- transistion probabilities'
     & ,' for state',i3,'have sum >1')
             endif
         endif
10    continue
      do 20 i=1,n
         read(1,*)(s(i,j),j=1,n+1)
20    continue
      read(1,*)(score(i),i=1,n)
      score(n+1)=0.
      write(2,25) n
25    format('number of states :',i4/)
      write(2,30)
30    format('Transision matrix P:'/)
```

```
      do 40 i=1,n
         write(2,50) (p(i,j),j=1,n+1)
49       format(100f15.6)
50       format(100(1pe10.2))
40    continue
      write(2,53)
53    format(/'Score matrix S:'/)
      do 55 i=1,n
         write(2,50) (s(i,j),j=1,n+1)
55    continue
      do 57 i=1,n
         s2(i)=p(i,n+1)*s(i,n+1)
      do 57 j=1,n
         s2(i)=s2(i)+p(i,j)*s(i,j)
57    continue
      write(2,60) k
60    format(/'# simulations per iteration per state:',i4)
c
c       calculate exact solution
c
      do 62 i=1,n
      do 63 j=1,n
         b(i,j)=-p(i,j)
63    continue
         b(i,i)=b(i,i)+1.
         e(i)=s2(i)
62    continue
      call rsolve(n,b,e,sing)
      write(2,64)
64    format(/'exact solution:')
      write(2,49)(e(i),i=1,n)
      write(2,70)
70    format(/'initial guess at exp scores:')
      write(2,49) (score(i),i=1,n)
      write(2,80)
80    format(/'iteration',5x,'expected score',10x,'sd')
c
c     run the algorithm
c
      do 90 m=1,it
          call iteration(n,k,p,s,score,var)
          if(m.ne.1) then
             sd=beta2*oldvar+alpha2*var
             oldvar=sd
             sd=dsqrt(sd)
             do 95 l=1,n
                score(l)=alpha*score(l)+beta*oldscore(l)
                oldscore(l)=score(l)
```

```
95          continue
         else
            oldvar=var
            sd=dsqrt(var)
            do 97 l=1,n
               oldscore(l)=score(l)
97          continue
            end if
         write(2,140) m,score(1),sd
         do 150 l=2,n
            write(2,145) score(l)
150      continue
         write(3,200) score(1)-e(1),sd,nhit(1)
200      format(2(1pe10.2),i4)
145      format(13x,f15.6)
140      format(/i6,7x,f15.6,5x,1pe11.3)
90    continue
      end
c
c*****************************************************
      subroutine iteration(n,k,p,s,score,var)
c
c   performs k simulations of MC with trans. matrix p
c   n is # of states, s is the payoff
c   vector. estimates of expected score are given in score.
c   var is sample variance of estimator for score(1).
c
c   k runs are run starting from each state for a total of
c   nk runs.  Information is used to estimate expected score
c   for every state that gets hit in the run.
c   SMP says that's legitimate
c*********************************************************
c
      implicit double precision(a-h,o-z)
      common/ibuff/nhit
      double precision p(100,101),q(100,101),r(100),s(100,101)
      double precision score(100),var,score2(10000),dk
      real x
      integer n,k,i,ix,nhit(100),nh
      logical hit(100)
      data ix/27928/
      dk=dble(k)
c
c      nhit(.) keeps track of how many runs in this iteration have
c      hit the particular state
c
      do 2 i=1,10000
         score2(i)=0.d0
```

```
2     continue
      call import(n,p,q,s,score)
      do 5 l=1,n
         score(l)=0.
         nhit(l)=0
5     continue
      do 10 istart=1,n
c
c     run k simulations starting in state istart
c
      do 10 ii=1,k
         nh=nhit(1)
c
c      r(l) represents the R-N deriv. between p and q
c      starting from when chain 1st hit state l .
c
         do 15 l=1,n
            r(l)=1.
15       continue
c
c    hit(l) keeps track of whether this particular
c    run has ever hit state l
c
         nhit(istart)=nhit(istart)+1
         do 17 l=1,n
            hit(l)=.false.
17       continue
         hit(istart)=.true.
         i=istart
         do while (i.le.n)
c
c        state n+1 corresponds to absorption
c        now in state i. choose next state j according to q
c
             j=1
             x=rangen(ix)
             do while ((x.gt.q(i,j)).and.(j.le.n))
                x=x-q(i,j)
                j=j+1
             end do
c
c    update scores for all states that have been hit
c
             do 30 l=1,n
                if(hit(l)) then
                   r(l)=r(l)*p(i,j)/q(i,j)
                   score(l)=score(l)+s(i,j)*r(l)
                end if
```

```
30          continue
c
c       score2(.) keeps track of all scores from state
c       #1 so that the sd may be calculated later from
c       a 2-pass algorithm
c
c
            nh=nhit(1)
            if((hit(1)).and.(nh.le.10000))
     &      score2(nh)=score2(nh)+s(i,j)*r(1)
            i=j
c
c         state j has now been hit
c
            if((.not.(hit(j))).and.(j.le.n)) then
                nhit(j)=nhit(j)+1
                hit(j)=.true.
            endif
c
c          if j.ne.n+1 repeat!
c
          end do
c
c     ii-th simulation is finished (particle has been absorbed)
c
10     continue
c
c     simulations for this iteration are over. normalize score
c     estimates and calculate sd
c
      do 50 l=1,n
         score(l)=score(l)/dfloat(nhit(l))
50    continue
      var=0.d0
      do 60 ii=1,nh
         var=var+(score2(ii)-score(1))**2
60    continue
      if (nh.gt.1) var=var/dfloat(nh-1)
      return
      end
```

```
c**************************************
      subroutine import(n,p,q,s,score)
c
c     computes importance probabilities q
c**************************************
c
      implicit double precision(a-h,o-z)
      double precision q(100,101),p(100,101),total
      double precision score(100),s(100,101)
      integer n,i,j
      do 5 i=1,n
         total=0.d0
      do 10 j=1,n+1
         q(i,j)=p(i,j)*(s(i,j)+score(j))
         total=total+q(i,j)
10    continue
      do 20 j=1,n+1
         q(i,j)=q(i,j)/total
20    continue
5     continue
      return
      end
c
c*****************************
    function rangen(ix)
c
c   uniform(0,1) random number
c*****************************
c
    integer a,p,ix,b15,b16,xhi,xalo,leftlo,fhi,k
    data a/16807/, b15/32768/, b16/65536/, p/2147483647/
    xhi=ix/b16
    xalo=(ix-xhi*b16)*a
    leftlo=xalo/b16
    fhi=xhi*a+leftlo
    k=fhi/b15
    ix=(((xalo-leftlo*b16)-p)+(fhi-k*b15)*b16)+k
    if (ix .lt. 0) ix=ix+p
    rangen=float(ix)*4.656612875e-10
    return
    end
```

```
c*********************************************
      subroutine lu(m,b,p,sing)
c
c  does lu facorization of matrix b
c  stores u in upper half, l in lower half
c  p is the permutation factors
c  m is dim of b
c  sing true if matrix is singular
c*********************************************
c
      implicit double precision(a-h,o-z)
      double precision bnorm,sum,temp,b(100,100)
      double precision pres,xmax
      integer ipivot,p(100),m
      logical sing
      sing=.false.
      pres=1.e-10
c
c     calculate norm of b
c
      bnorm=0.
      do 10 j=1,m
         sum=0.
         do 20 i=1,m
            sum=sum+abs(b(i,j))
20       continue
         if(sum.gt.xnorm) xnorm=sum
10    continue
c
c   do pivoting
c
      do 30 j=1,m
c
c        find row with largest abs val in jth col
c
         xmax=0.
         do 35 k=j,m
            if(abs(b(k,j)).gt.xmax) then
               xmax=abs(b(k,j))
               ipivot=k
            end if
35        continue
          if(abs(b(ipivot,j)).le.pres) then
            sing=.true.
             write(2,70)
             goto 80
          end if
          p(j)=ipivot
```

```
          if(ipivot.eq.j) goto 50
c
c         switch rows j and ipivot
c
          do 40 k=j,m
             temp=b(j,k)
             b(j,k)=b(ipivot,k)
             b(ipivot,k)=temp
40        continue
c
c        clear jth col below (j,j)th elem.
c        store L factors below diagonal
c
50         do 60 i=j+1,m
             b(i,j)=b(i,j)/b(j,j)
           do 60 k=j+1,m
             b(i,k)=b(i,k)-b(i,j)*b(j,k)
60         continue
           if(abs(b(m,m)).le.pres) then
             sing=.true.
             write(2,70)
           end if
30      continue
70         format('MATRIX IS SINGULAR')
80         return
           end
c
c****************************************
      subroutine rsolve(m,a,b,sing)
c
c      solves ax=b puts answer in b
c      sing is true if a is singular
c****************************************
c
      implicit double precision(a-h,o-z)
      double precision a(100,100),b(100),temp
      integer m,p(100)
      logical sing
      call lu(m,a,p,sing)
c
c      forward elimination
c
       if (.not.(sing)) then
       do 10 j=1,m-1
             temp=b(j)
             b(j)=b(p(j))
             b(p(j))=temp
          do 20 k=j+1,m
```

```
            b(k)=b(k)-a(k,j)*b(j)
20        continue
10      continue
c
c       backward substution
c
        do 30 j=m,1,-1
           b(j)=b(j)/a(j,j)
        do 30 k=1,j-1
           b(k)=b(k)-a(k,j)*b(j)
30      continue
        end if
        return
        end
```

Bibliography

[1] Billingsley, P. (1986). *Probability and Measure* 2d ed. New York, John Wiley & Sons.

[2] Booth, T.E. (1985). "Exponential convergence for Monte Carlo particle transport?". *Transactions of the American Nuclear Society* 50: 267-268.

[3] Booth, T.E. (1985) "A sample problem for variance reduction in MCNP". LA-10363-MS, Los Alamos National Laboratory.

[4] Booth, T.E. (1986). "A Monte Carlo learning/biasing experiment with intelligent random numbers". *Nuclear Science and Engineering* 92: 465-481.

[5] Booth, T.E. (1988). "The intelligent random number technique in MCNP". *Nuclear Science and Engineering* 100: 248-254.

[6] Booth, T.E. (1989). "Zero-variance solutions for linear Monte Carlo". *Nuclear Science and Engineering* 102: 332-340.

[7] Bucklew, J.A. (1990). *Large Deviation Techniques in Decision, Simulation, and Estimation.* New York, Wiley Interscience.

[8] Cramer, S.N., Gonnord, J., & Hendricks, J.S. (1986). "Monte Carlo techniques for analyzing deep-penetration problems". *Nuclear Science and Engineering* 92: 280-288.

[9] Glynn, P.W. and Iglehart, D.L. (1989) "Importance sampling for stochastic simulations". *Management Science* 35: 1367-1392 .

[10] Halton, J.H. (1962). "Sequential Monte Carlo". *Procedings of the Cambridge Philosophical Society* 58: 57-78.

[11] Lehtonen, T. and Nyrhinen, H. (1992). "Simulating level-crossing probabilities by importance sampling". *Advances in Applied Probability* 24: 858-874.

[12] Lewis, E.E. and Miller, W.F. (1984). *Computational Methods of Neutron Transport.* New York, John Wiley & Sons.

[13] Lux, I. and Koblinger, L. (1991). *Monte Carlo Particle Transport Methods: Neutron and Photon Calculations.* Boca Raton, CRC Press, Inc.

[14] "MCNP- A general Monte Carlo code for neutron and photon transport". LA-7396-M, Los Alamos National Laboratory (April 1981).

[15] Murthy, K.P.N. & Indira, R. (1986). "Analytical results of variance reduction characteristics of biased Monte Carlo for deep-penetration problems". *Nuclear Science and Engineering* 92: 482-487.

[16] Osborn, R.K. and Yip, S. (1966). *The Foundations of Linear Transport Theory.* New York, Gordon & Breach.

[17] Parzen, E. (1954). "On uniform convergence of families of sequences of random variables". *University of California Publications in Statistics* Volume 2, Number 2, 23-54.

[18] Ripley, B.D. (1987). *Stochastic Simulation.* New York, John Wiley & Sons.

[19] Rockafellar, R.T. (1970). *Convex Analysis.* New Jersey, Princeton University Press.

[20] Rudin, W. (1976). *Principles of Mathematical Analysis* 3d ed., McGraw-Hill Book Company, New York.

[21] Siegmund, D. (1976). "Importance sampling in the Monte Carlo study of sequential tests". *Annals of Statistics* 4: 673-684.

[22] Siegmund, D. (1985). *Sequential Analysis.* New York, Springer-Verlag.

[23] Troubetzkoy, E.S. (1991). "Optimization of linear Monte Carlo calculations". *Nuclear Science and Engineering* 107: 359-364.

[24] Ullo, J.J. (1986). "Use of multidimensional transport methodology on nuclear logging problems". *Nuclear Science and Engineering* 92: 228-239.

[25] Williams, D. (1991). *Probability with Martingales.* New York, Cambridge University Press.

DATE
FILMED
9/12/94

END