

VARIANCE REDUCTION WITH
QUASI CONTROL VARIATES

by

Markus Emsermann

B.S., Clemson University, 1991

M.S., Clemson University, 1993

A thesis submitted to the
University of Colorado at Denver
in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Applied Mathematics

2000

This thesis for the Doctor of Philosophy
degree by
Markus Emsermann
has been approved
by

Burton Simon

Stephen Billups

Karen Kafadar

Anatolii Puhalskii

William Wolfe

Date

Emsermann, Markus (Ph.D., Applied Mathematics)

Variance Reduction with Quasi Control Variates

Thesis directed by Professor Burton Simon

ABSTRACT

In a simulation a random variable, Y , can often be identified that is likely to be highly correlated with a random variable of interest, X . If $\mu_Y = E[Y]$ is known then Y can be used as a control variate to estimate $\mu_X = E[X]$ more efficiently than by a direct simulation of X . We propose a method that uses Y to speed up the simulation when μ_Y is unknown. The method is effective when μ_Y can be efficiently estimated in an auxiliary simulation that does not involve X . For a simulation of length $t > 0$ time units, we invest pt units estimating μ_Y with the auxiliary simulation, yielding an estimator \bar{Z}_{pt} . The remaining $q = (1 - p)t$ units are spent on the main simulation yielding estimates $(\tilde{X}_{qt}, \tilde{Y}_{qt})$ for (μ_X, μ_Y) . The two simulations can be interleaved so they are effectively done simultaneously. For each $p \in (0, 1)$ and $\alpha \in \Re$ we have a quasi control variate estimator for μ_X

$$\bar{Q}_t(p, \alpha) = \tilde{X}_{qt} + \alpha(\tilde{Y}_{qt} - \bar{Z}_{pt}), t > 0.$$

We find p and α that minimize the asymptotic variance of $\bar{Q}_t(p, \alpha)$ in terms of statistics that are estimated during the simulations and then describe an easily implemented adaptive procedure that achieves the minimum variance. The adaptive procedure evolves into the optimal quasi control variate scheme if it is more efficient than a direct simulation, $\bar{X}_t \rightarrow \mu_X$; otherwise it develops into the direct simulation. This research is motivated by stochastic linear programs where problem data are random; in this setting, we estimate the expected value of the objective function. We illustrate applications involving petroleum refining and power system reliability evaluation. In the former application the constraint matrix is random and a simple approximation can be constructed to generate an effective control variate. For the power system reliability evaluation illustration, a special “dual” approximation to the primal problem is constructed to form an effective control variate. In both cases, a tremendous improvement in efficiency is realized.

This abstract accurately represents the content of the candidate’s thesis. I recommend its publication.

Signed _____
Burton Simon

DEDICATION

This thesis is dedicated to my father, Hubert Emsermann, who stood by me
for the long haul.

ACKNOWLEDGMENTS

In the words of Paul McCartney, it has been a “long and winding road”. Along the way I have met people and had experiences I will cherish forever and others I hope to forget. Words cannot convey how grateful I am to have had Professor Burt Simon as an advisor and friend. Through a simple email he proposed an interesting research problem that I immediately tackled and developed into my thesis, of which I am quite proud. His tremendous insight, thoughtful direction and mathematical brilliance were instrumental in my success. I now have a better understanding and greater appreciation for what it means to be a true mathematician. I enjoyed his encouragement when I was on the right track and admired his patience when I exhibited characteristics of a blabbering fool. I also extend a special thanks to Professor Tolya Puhalskii for all his helpful suggestions, great ideas and solid support. His input was absolutely invaluable. I greatly appreciate Professor Stephen Billups’ careful proofreading, helpful comments and suggestions regarding my thesis.

I also wish to thank my family, especially my father for his tremendous financial, spiritual and emotional support over the past seven years (not to mention undergraduate and graduate school at Clemson University); without his help, I would have never completed this project. I also wish to thank Caroline for enduring all my moaning and fussing during the painful period when things were just not working out for me. I also thank her for putting her

life on hold and being patient with me while completing this undertaking. Her compassion and encouragement will never be forgotten. What probably kept me sane the most were the great vacations my brother Martin and sister-in-law Marty planned (and funded) every year. All the craziness, laughter and just pure fun we enjoyed reminded me not to take life too seriously and that things will be just fine. How about that dancing to “You Shook Me All night Long” at Charlie Browns?

I also wish to extend a special thanks to Drs. Ruth O’Brien and Maureen McClatchy at the Kempe Prevention Research Center where I was fortunate enough to have been employed for the last two years. I wish them continued success in their child abuse research and nurse-home visitation program.

I also wish to express my appreciation to the family of Lynn Bateman. It was an honor to receive the Lynn Bateman Memorial Teaching Award and be recognized for excellence in teaching, a *weltanschauung* we obviously shared.

And for those that I have not mentioned but had some connection with the production of this thesis: The Lord, Bill Pleau, Norm Lemay, Spinoza, Cogito, Katia, Allen Holder, Joe, John, “The Hook”, Wok to You, Old Chicago, P.K. Sen, my committee, Charles, Alissa Ruelle, Cid Hart, Ralph Kramden, Jean-Luc Picard, Marcia Kelly, Annette Beck and Guillermo Jimenez; I thank you all.

Finally, this acknowledgments page would not be complete if I did not express my appreciation to all those wonderful rentals I resided in while completing my degree. Renting in Denver has provided me much excitement.

The most notable memories range from having the entire Denver Fire Department in my apartment vacuuming a foot of water that had washed in from a broken fire main to having the upstairs neighbor's bath tub almost fall into my kitchen.

Isn't it astonishing that all these secrets have been preserved for so many years just so we could discover them.

-Orville Wright, June 7, 1903

CONTENTS

Figures	xi
Tables	xii
1. Introduction	1
1.1 Background	3
1.2 Literature Review	5
1.3 Multidimensional Integration and Stochastic Programming	6
1.4 The Monte Carlo Method	7
1.4.1 Stochastic Programming Problems	9
1.4.2 Efficiency-Improving Techniques	13
1.4.3 Control Variates	15
2. The Quasi Control Variate Method	19
2.1 Preliminaries	20
2.2 Theoretical Development	26
2.3 Implementation	42
2.4 An Optimal Dynamic QCV procedure	43
2.5 Generic Example	49
3. Stochastic Linear Programs	52
3.1 Petroleum Refinery	52
3.1.1 Refinery	53
3.1.2 Crude Oil Quality	56

3.1.3	Linear Programming Model	57
3.1.4	Numerical Experiments	63
3.2	Stochastic LP with random RHS	66
3.2.1	Construction of the Dual Approximation	70
3.2.2	Deterministic vertex enumeration	71
3.2.3	Random vertex generation	75
3.2.4	Random vs. deterministic generation	75
3.3	Power System Reliability Evaluation	82
3.3.1	Power Systems	83
3.3.2	Fundamentals of Circuit Theory	84
3.3.3	Linear Programming Power Flow Model	88
3.4	Numerical Experiments	92
4.	Conclusion and Future Work	94
4.1	Summary	94
4.2	Future Research	94
	<u>Appendix</u>	
A.	Refining Fractionation Data	95
B.	Power System Bus and Tie-Line data	97
	<u>References</u>	100

FIGURES

3.1	Statistics as a function of Size of Analytical Approximation . . .	67
3.2	AAV as Function of Number of Points in Approximation . . .	68
3.3	Correlations for Equally Probable Case: o-Randomly Generated, x-Enumerated.	77
3.4	Log-linear plot of Time to construct Dual Approximation for Equally Probable Case.	78
3.5	Correlations for Non-Equally Probable Case: o-Randomly Generated, x-Enumerated	80
3.6	Log-linear plot of Time to construct Dual Approximation for Non-Equally Probable Case	81

TABLES

3.1	Topping	59
3.2	Thermic Re-forming	60
3.3	Raw Material Costs	61
3.4	Product Prices	61
3.5	Distilling Capacities	62
3.6	EPNS Speedup	93
3.7	LOLP Speedup	93
A.1	Re-forming	95
A.2	Vacuum	96
A.3	Catalytic Cracking Unit	96
A.4	Catalytic polymerizator	96
B.1	Branch Data	97
B.2	Branch Data (cont.)	98
B.3	Generating Unit Locations	99
B.4	Bus Load Data	99

1. Introduction

We are interested in estimating the expected value of an output random variable, X , involved in a simulation. In addition to obtaining a point estimate of $\mu_X = E[X]$, we would like to make it as reliable as possible. The reliability of a point estimate is generally measured by its variance. In order to increase the reliability of an estimate, we attempt to reduce its variability without increasing the sampling effort, which in simulations is generally measured in computer time. These methods are known as *variance reduction techniques* (VRTs) or *efficiency-improving techniques*.

This thesis generalizes a common VRT, the classical control variate estimation procedure, to the case when the control variate mean is unknown. We give motivation for this generalization, present theoretical results and develop an algorithm for its implementation. In addition, numerical results involving problems in engineering are presented demonstrating the usefulness of our procedure.

Basically, a control variate is a random variable, used to improve efficiency in simulation experiments, whose expectation is known and that is

correlated with a statistical estimator of interest. By way of illustration, one can often identify a random variable Y , also involved in the simulation, that is likely to be highly correlated with X . If so, using Y , an unbiased estimator of X can be formed that is more efficient than a direct simulation of X alone as follows:

$$Q(\alpha) = X + \alpha(Y - \mu_Y),$$

where α is a scalar parameter (whose value is determined later as to not disrupt the flow of the introduction) and μ_Y is *known*.

We investigate the case where a random variable Y might be an effective control variate candidate but where the mean μ_Y of this random variable is unknown. We propose that in such cases it may be beneficial to spend time estimating this unknown control variate mean and then proceed with the classical control variate simulation by using the estimate in place of the true (unknown) mean. We call such an estimator a *quasi control variate* (QCV). In Section 2 we construct our QCV estimator and find its asymptotic variance in terms of optimal parameters that can be estimated during a simulation. We develop an easily implemented adaptive procedure, which achieves the minimum variance, that continually updates estimates of the simulation parameters. The adaptive procedure evolves into the optimal quasi control variate scheme if it is more efficient than a direct simulation, $\bar{X}_t \rightarrow \mu_X$; otherwise it develops into

the direct simulation. In Section 3 we present numerical results involving two “real-world” applications regarding petroleum refining and power system reliability evaluation. In Section 4 we conclude by discussing future work and extensions.

1.1 Background

Scientists are often interested in determining the values of unknown parameters in complex stochastic systems. For instance, in the simulation of a job shop environment, the length of time required to complete the work on a certain task may be a random variable of interest to the experimenter. He or she may wish to estimate the mean of this completion time to analyze the component flow and make decisions to help minimize the Total Flow Cost. In power system reliability evaluation, planners are interested in the system’s ability to supply energy to consumers by calculating a comprehensive set of reliability indices. Probabilistic indices such as loss of load probability (LOLP) and expected power not supplied (EPNS) are common performance measures of electric facilities. These indices are complicated functions of equipment outages and load duration and provide guidance in determining the need for reliability improvements and system reinforcements. Also, financial decision-making problems can often be modeled as stochastic programs. In this case, given

a sequence of investment decisions where the returns are random, a portfolio manager would be interested in the expected utility of terminal wealth. The aforementioned cases illustrate where certain information cannot be directly computed due to the complex and sophisticated nature of the problem.

These complexities are due, in part, to problem dimensionality and complicated function evaluations. The values of these parameters oftentimes are high-dimensional integrals whose integrand properties may be unknown or too complex to apply traditional numerical techniques. Thus, alternative approaches must be used to estimate these parameter values. We are particularly interested in the multidimensional integrals that occur in stochastic programs. Various approaches have been developed to address this problem of numerical integration in stochastic programs. The general problem requires some form of approximation. The most common approximations involve 1) discretizations of the probability distribution to produce bounds on these integrals and 2) Monte Carlo sampling. We are interested in the latter method of approximation since the former is quite restrictive in that it requires a considerable amount of problem structure. Monte Carlo sampling is based on approximating the integral by an average of values of the integrand at randomly selected points. Monte Carlo simulation is quite simple in nature and applies to virtually any problem. Often, the Monte Carlo method is the only useful approach

in estimating integrals, especially when the integrand is not explicitly given. Often experimenters need to estimate a certain quantity within a prescribed level of statistical accuracy. If the cost of achieving this precision is not within the computing time budget constraint, one may need to consider alternate sampling plans that involve some method of variance reduction.

There are several VRTs commonly used in practice; however, the method of control variates, is a popular variance reduction technique due to its simplicity and potential for widespread use. The availability of VRTs distinguishes the Monte Carlo method from the more simplistic sampling experiments that preceded it. One strategy for obtaining a control variate is to use a simpler model's known performance as a control variate to "correct" the principal, complex model's output stream in hopes of reducing variability.

1.2 Literature Review

Little research has been devoted to the concept of QCV estimation, where a portion of a given simulation time is invested in estimating an unknown control variate mean. Quasi control variates (by another name) were considered by Schmeiser, Taaffe and Wang [11] as an alternative to *baised control variates*, Schmeiser, Taaffe and Wang [10]. Their analysis of QCV procedures relies on heuristically determined "cost" measures associated with estimating

the control variate mean and performing the main simulation. Central to their analysis is that the simulation experiments have a finite number of replications; they do not consider asymptotics. We, provide an asymptotic analysis of QCV procedures in terms of asymptotic variance parameters, which can usually be estimated (consistently) in a straightforward manner. Although their QCV procedure is not optimal from our (asymptotic) perspective, they make a strong case for using a biased estimate $\hat{\mu}_Y$ for μ_Y , instead of resorting to an unbiased estimator, in a time constrained simulation experiment where the approximation error, $|\hat{\mu}_Y - \mu_Y|$, is sufficiently small in comparison with the simulation error.

1.3 Multidimensional Integration and Stochastic Programming

The methodology for one-dimensional integration has been extensively developed. Research in numerical analysis has produced various quadrature methods that are extremely effective in approximating definite integrals. Unfortunately, higher dimensional integration does not have such readily available formulae for their evaluation. A few of the major problems are discussed below.

Firstly, the domains of multidimensional integrals can take on an infinite variety of shapes that may not be transformable into simpler regions

which would facilitate integration. Even if these transformations were available, they are usually so clumsy and unmanageable that practitioners do not use them. Secondly, the “curse of dimensionality” is an inevitable problem for which numerical analysis has not offered effective remedies. For instance, if we need M points to achieve a desired level of accuracy in one–dimension using product formulas, the required number of nodes would increase to M^d in d –dimensions. Thus, the amount of work required to evaluate multidimensional integrals grows much faster than the number of dimensions. Also, little or no information may be readily available about the integrand in terms of its value, smoothness and variational characteristics. Lastly, error analysis is much more difficult in higher dimensions than in one–dimension, where strong bounds have been developed. Hence, one possibility to attempt to circumvent this dimensional effect and accuracy question is to use Monte Carlo methods.

1.4 The Monte Carlo Method

Because Monte Carlo simulation appears to offer the best possibilities for higher dimensional integration [20], it seems to be the natural choice for use in stochastic programs. This method of “approximate integration” requires pseudorandom numbers generated from a given distribution; since the integrals involved are in the form of expectations, probability theory gives justification

for this approach. An early application of the Monte Carlo method includes estimating the value of π by throwing a needle at parallel lines at equal distances and then counting the number of times the needle crosses the lines [30]. More descriptively, suppose we want to estimate

$$\mu = E[g(X)] = \int g(x) dP_X(x), \quad (1.1)$$

where g is a function of n variables and $X = (X^1, X^2, \dots, X^n)$ is some random vector having a given distribution function $P_X(\cdot)$. To approximate μ we can generate M independent and identically distributed (i.i.d.) replicates of X , $\{X_1, X_2, \dots, X_M\}$, and then compute

$$\hat{\mu} = \frac{\sum_i^M g(X_i)}{M}. \quad (1.2)$$

By the strong law of large numbers (see Theorem 2.2), we know that

$$\lim_{M \rightarrow \infty} \frac{\sum_i^M g(X_i)}{M} = \mu \quad (1.3)$$

almost surely (*a.s.*). Thus, we can use the average of the generated points $g(X_i), i = 1, \dots, M$ as a *Monte Carlo* estimate of μ . This is the approach we take for estimating high-dimensional integrals. The variance of the estimator (1.2) is

$$\sigma_{g(x)}^2/M, \tag{1.4}$$

which is a measure of the resulting error in a Monte Carlo simulation. This means that the standard error of the estimator is proportional to $1/\sqrt{M}$ which can be regarded as slow convergence. In addition, this error can be large and thus variance reduction techniques are employed in order to obtain more reliable estimates of μ . The implementation of the Monte Carlo method is relatively simple and can apply to virtually any function. Also, an estimate of the deviation (1.4) can be obtained from the Monte Carlo simulation with little additional effort. For further information on Monte Carlo computations, see Hammersley and Handscomb [31]. This estimation procedure is the basis for stochastic programming problems.

Remark 1.1 Note that (1.4) is merely the variance of the “parent” random variable divided by the number of realizations used to calculate the sample mean. This fact will be important to remember in the development of Section 2.2.

1.4.1 Stochastic Programming Problems

The use of quantitative methods proved to be very successful for the analysis of military operations during World War II. In particular, the invention

of the Simplex method by G.B. Dantzig [16] to solve linear programming problems marked the inception of the usage of analytical approaches, which became to be known as operations research, to solving problems in planning, allocation and scheduling. The development of modeling and solution techniques that would allow for stochastic elements within problem data quickly accelerated in the 1950s due, in part, by the rapid growing interest of the great potential mathematical techniques had to “real world” problem solving where uncertainty was prevalent. This uncertainty is usually characterized by a probability distribution on the parameters. In practice it can range from a few possible scenarios to specific and precise joint probability distributions. Early applications include an airline fleet–assignment problem by Ferguson and Dantzig [25], where passenger demand on each route was uncertain, and the classical diet problem first studied by Stigler [55] and latter employed by Dantzig [17] where nutritional variations within food groups were considered. At that time these types of applications were known as *Linear Programming under Uncertainty* and eventually developed into a branch of optimization known as *Stochastic Programming*. This field deals with the theory and methodology of incorporating probabilistic, or stochastic, variations into an optimization model. The necessity of this incorporation is evident in problems such as portfolio management, energy modeling, airline scheduling, production planning and inventory

theory where the assumption of certainty in problem data cannot be justified. Thus there are numerous applications where traditional deterministic mathematical programming models are inadequate and the decision maker is faced with incorporating uncertainty in an appropriate and realistic manner without making solution procedures computationally intractable. The origin of the various stochastic programming models of today stems from the *two-stage linear program* presented independently by Dantzig [15] and Beale [5], the *chance-constrained model* developed by Charnes and Cooper [12] and the *distribution problem* given by Tintner [56] and Mangasarian [42]. In the two-stage programming problem, or what is commonly known as the *recourse problem*, an initial decision is made without the realization of the uncertain parameters; and adjustments are made once a specific realization of the data is observed to minimize total expected cost. An important class of probabilistic programming arises when we assume *chance constraints*, or constraints that place lower limits on the probabilities of satisfying certain inequalities. Often these problems can be converted into deterministic ones. The *distribution* problem of stochastic programming is the calculation of the expected value of the objective function of a mathematical program in which model coefficients are uncertain and the values of the decision variables are chosen after the uncertainty is resolved. Such stochastic programs where decisions can be made after the randomness is

observed are called *wait and see* models. This research deals exclusively with the distribution problem.

The general formulation of a stochastic program is the following:

$$\begin{aligned}
& \min_x && E[f_0(x, \xi)] \\
& \text{s.t.} && E[f_i(x, \xi)] \leq 0, \quad i = 1, \dots, s, \\
& && E[f_i(x, \xi)] = 0, \quad i = s + 1, \dots, m, \\
& && x \in X \subseteq \mathfrak{R}^n.
\end{aligned} \tag{1.5}$$

$f_i, i = 0, \dots, m$ are real valued functions.

The “expectation functionals” are defined as follows:

$$E[f_i(x, \xi)] \stackrel{\text{def}}{=} \int_{\Xi} f_i(x, \xi) dP(\xi) \quad i = 0, \dots, m, \tag{1.6}$$

where ξ is a random vector with support $\Xi \subset \mathfrak{R}^k$ and P is a probability distribution defined on \mathfrak{R}^k .

Remark 1.2 Since we are concerned with only the distribution problem of stochastic programming we will not be directly working with problems in the form of (1.5). This information is provided for clarity and completeness. Although our research directly applies to such problems in estimating the expectation functionals, we are primarily interested in estimating the expected value of stochastic programs with random data.

More specifically, we are interested in estimating the mean $E[X]$ of the objective function value of linear programs in the following form:

$$X = \min_{x \in \mathfrak{R}^n} \{ cx \mid Ax \geq b, x \geq 0 \}, \quad (1.7)$$

where some of the problem data A, b are random.

Only in small specialized cases can the integrals in (1.6) be calculated analytically. Generally, applied problems of interest require the evaluation of high-dimensional integrals, an inherent problem in the field of stochastic programming. Approximation schemes, Monte Carlo estimation and bounding methods are a few approaches used in dealing with their evaluation. We are primarily interested in the Monte Carlo approach.

1.4.2 Efficiency-Improving Techniques

Efficiency-Improving Techniques or variance reduction techniques are experimental design methods used to increase the precision of simulation-sampling based point estimators without increasing the sampling effort. Recent research in stochastic programming involves developing intelligent sampling procedures such as stratification, importance sampling and control variates to obtain at least as “good” estimates with smaller sample sizes. Gaivoronski [Private Communication] uses stratified sampling in stochastic quasi gradient (SQG) methods in the parallel setting. SQG methods are search procedures

that use statistical estimates of gradients and employ analogous deterministic search procedures such as decent methods, except where the actual gradients are replaced by statistical estimates. Infanger [32] used importance sampling as a variance reduction technique in a Monte Carlo sampling–decomposition algorithm for solving large–scale stochastic linear programs. There are many different VRTs; however, we consider only one, the method of *control variates*. Before we motivate the notion of control variates, we give a definition that is important in its development.

Definition 1.3 (Correlation Coefficient) The *correlation coefficient*, denoted by ρ_{XY} , of random variables X and Y is defined to be

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

provided that σ_{XY} , $\sigma_X > 0$ and $\sigma_Y > 0$ exist.

The correlation coefficient is a measure of a linear relationship of X and Y . The correlation coefficient is unitless and satisfies $-1 \leq \rho_{XY} \leq 1$. Values of ρ_{XY} close to 1 indicate a positive linear relationship and values close to -1 indicate a negative linear relationship. Values of ρ_{XY} close to 0 indicate no linear relationship.

1.4.3 Control Variates

The method of control variates is a common and widely used approach to variance reduction. The development of these techniques began during World War II in order to increase the efficiency of the Monte Carlo evaluation of integrals that arose in nuclear particle transport problems. A thorough exposition and detailed description of the method of control variates can be found in [40]. For selected applications where control variates were implemented, see [38, 39, 47, 50, 54]. The idea is to exploit possible correlations between random variables within a Monte Carlo simulation. We consider the one-dimensional case. Let the random variable X be an output random variable in a simulation for which we would like to estimate its mean, μ_X . Suppose there is another random variable Y generated in the same simulation for which we know its mean, μ_Y . We can see that the following random variable is an unbiased estimator of μ_X :

$$Q(\alpha) = X + \alpha(Y - \mu_Y), \tag{1.8}$$

where α is a scalar parameter. The random variable Y is called a *control variate* for X . Also, we find that

$$\sigma_{Q(\alpha)}^2 = \sigma_X^2 + 2\alpha\sigma_{XY} + \alpha^2\sigma_Y^2.$$

Using basic calculus, it can be verified that

$$\alpha^* = -\frac{\sigma_{XY}}{\sigma_Y^2} \quad (1.9)$$

minimizes the variance of $Q(\alpha)$. The amount of variance reduction is quantified in the following expression:

$$\sigma_{Q(\alpha^*)}^2 = (1 - \rho_{XY}^2)\sigma_X^2, \quad (1.10)$$

where ρ_{XY}^2 is the correlation coefficient of X and Y .

Thus, we immediately see that the higher X and Y are correlated in magnitude, the higher the variance reduction of $Q(\alpha^*)$. Unfortunately, α^* is generally not known in advance and must be estimated from the simulation. If n simulation runs are performed resulting in (X_i, Y_i) , $i = 1, \dots, n$, then we can estimate α^* by

$$\hat{\alpha}^* = -\frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}, \quad (1.11)$$

where \bar{X}_n, \bar{Y}_n are the sample means of X and Y , respectively; see [40, page 637].

Of course, before we can employ this method of variance reduction, we must have available a control variate with known mean that is highly correlated with X and relatively easily evaluated. The selection of a control variate is more

of an art than a science. Since Monte Carlo simulations are driven in part by sequences of i.i.d. random variables for which we know the distributions, often simple functions of these input streams serve as good control variables. For example, let X be an output random variable that represents the average waiting time in some queue for the first 100 people; suppose we want to estimate μ_X . A reasonable control variate for X is the average service time, say Y , of the first 99 customers. That is, longer-than-average service times tend to yield longer-than-average waiting times and vice versa. Note that we know μ_Y since we generated the service-times from a known input distribution. Thus, this is a simple example illustrating how functions of the input sequence of random variables can serve as control variates. We refer to such control variates as *internal* since they are “freely” available (within the simulation) and add nothing to the simulation costs.

In some applications, an approximate and analytically solvable model is constructed that reflects the basic characteristics of the larger complex problem. This simplified model has the property that we can precisely determine the value of certain parameters, i.e, the means of output random variables that are not readily available in the complex problem. Both models are employed in the Monte Carlo simulation and driven by the same sequence of input random

variables. Functions of the approximate model can provide good control variates for the output variable of the complex model. Thus, provided that there are strong correlations between random variables of the respective models, the solution of the analytical problem is used to “predict” or “control” the result of the detailed one. These types of control variates are called *external* since they are not costless; that is, they involve another simulation to evaluate the control variate. In our research we employ external control variates.

2. The Quasi Control Variate Method

One of the barriers to widespread application of control variates is the potential difficulty in selecting an appropriate control variate with known expectation. Oftentimes, an experimenter has strong reason to believe that a given random variable Y , whose mean is unfortunately unknown, may be highly correlated with the variable in which he or she is interested. Thus, the scientist must be content with a possibly less effective control variate or must seek alternative variance reduction methods. We propose that in many cases it might be beneficial to spend time estimating this unknown control variate mean and then proceed with the classical control variate approach by using the estimate in place of the true (unknown) mean. The main concern is the allocation of valuable simulation time to a variable whose mean we are not interested in. Why spend effort, which could be used estimating the parameter of interest, on a different variable? One can see that if the proposed control variate is highly correlated with the variable of interest and relatively less expensive to generate realizations, it might be advantageous to apportion some

of the work to estimating the control variate mean. Thus, we propose a generalized control variate technique called the *quasi control variate* (QCV) scheme. In the following sections, we extend the methodology of control variates and present various practical applications.

2.1 Preliminaries

The following are important theorems in probability and can be found in most texts such as [61].

Theorem 2.1 (Convergence in probability) For a sequence X_1, X_2, \dots, X_n of i.i.d. random variables with finite mean μ and variance σ^2 define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, then

$$\lim_{n \rightarrow \infty} Pr(|\bar{X}_n - \mu| > t) = 0 \text{ for any } t > 0. \quad (2.1)$$

Since t can be arbitrarily small, \bar{X}_n becomes arbitrarily close to μ as $n \rightarrow \infty$.

We say that \bar{X}_n converges *in probability* to μ and denote this by

$$\bar{X}_n \xrightarrow{P} \mu \text{ as } n \rightarrow \infty. \quad (2.2)$$

Theorem 2.2 (Almost sure convergence) For a sequence X_1, X_2, \dots, X_n of i.i.d. random variables with finite mean μ ,

$$Pr(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1, \quad (2.3)$$

where \bar{X}_n is given in Theorem 2.1.

We say that \bar{X}_n converges *almost surely* to μ , and denote this by

$$\bar{X}_n \rightarrow \mu \text{ a.s. as } n \rightarrow \infty. \quad (2.4)$$

Almost sure convergence implies convergence in probability.

Definition 2.3 (Convergence in distribution) Suppose that X_n , $n = 1, 2, \dots$, and X are random variables with distribution functions F_n , $n = 1, 2, \dots$, and F respectively. We say that X_n converges in distribution to X , denoted $X_n \Longrightarrow X$, if

$$F_n(x) \rightarrow F(x) \text{ as } n \rightarrow \infty, \quad (2.5)$$

for all x at which F is continuous.

Definition 2.4 (Counting process) A stochastic process $\{N(t), t \geq 0\}$ is said to be a *counting process* if $N(t)$ represents the total number of “events” or “epochs” that have occurred up to time t . Define T_1 to be the time of the first event and for $n > 1$, let T_n denote the elapsed time between the $(n - 1)$ st and n th event.

Now, we give the definition of a counting process called a *delayed renewal process*, which will become important in the development of the quasi control variate approach.

Definition 2.5 If the sequence of nonnegative random variables $\{T_1, T_2, \dots\}$ is independent and $\{T_2, T_3 \dots\}$ is identically distributed, then the counting process $\{N(t), t \geq 0\}$ is said to be a *delayed renewal process*. If T_1 has the same distribution as $\{T_2, T_3 \dots\}$, then the counting process is said to be a *renewal process*.

We shall let \mathbf{F} denote the interarrival distribution, and to avoid trivialities, assume that $Pr\{T_2 = 0\} < 1$, so that T_2 is not identically 0. Furthermore, let $\mu = E[T_2]$.

Note, in the definition above, we allow for T_1 to have a different distribution than the common distribution of $T_2, T_3 \dots$

The following theorems (see [28] and [61] respectively) are important in the development of the variance of our quasi control variate.

Theorem 2.6 (Elementary Renewal Theorem) Let $N(t)$ be defined according to Definition 2.5 and $\mu = E[T_2]$, then

$$\lim_{t \rightarrow \infty} \frac{N(t)}{t} = \frac{1}{\mu} \quad \text{a.s..} \quad (2.6)$$

Theorem 2.7 (Slutsky) If $\{X_n\}, \{Y_n\}$ are sequences of random variables on some probability space with $X_n \implies X$ and $Y_n \xrightarrow{P} c$, where c is a finite constant,

then

$$X_n Y_n \Longrightarrow cX. \quad (2.7)$$

We now state and prove a lemma that will be useful in the development of our quasi control variate.

Lemma 2.8 Let F be the distribution function of a random variable ξ with finite mean μ and variance σ^2 . Also let $\{N(t), t \geq 0\}$ be a delayed renewal process representing the total number of replications of ξ available up to time t , where the expected time to generate an observation of ξ is $0 < \tau < \infty$ and the expected time to initialize the simulation is $0 < \gamma < \infty$. Also, define

$$\bar{\xi}_t = \frac{1}{N(t)} \sum_{i=1}^{N(t)} \xi_i, \quad (2.8)$$

where $\{\xi_1, \xi_2, \dots, \xi_{N(t)}\}$ are replications of ξ .

It follows that

$$\sqrt{t}(\bar{\xi}_t - \mu) \Longrightarrow N(0, \tau\sigma^2). \quad (2.9)$$

We refer to the quantity $\tau\sigma^2$ as the *asymptotic variance parameter* (AVP) for $\bar{\xi}_t$.

Proof: We first prove that $N(t) \rightarrow \infty$ as $t \rightarrow \infty$, *a.s.*. Define

$$W_n = \sum_{i=1}^n D_i \quad n \geq 1, \quad (2.10)$$

where D_1 is the time to set up the Monte Carlo simulation (overhead) *plus* the time to generate ξ_1 ; also, for $i \geq 2$ let D_i be the time required to generate the i^{th} replication of ξ_i . Then, it is clear that

$$N(t) \geq k \iff W_k \leq t. \quad (2.11)$$

The probability distribution of W_n can be calculated by

$$Pr\{W_n \leq x\} = F_n(x), \quad (2.12)$$

where $F_1(x) = F(x)$ is assumed known, and $F_n(x)$ is calculated by the convolution formula:

$$F_n(x) = \int_0^\infty F_{n-1}(x-y)dF(y), \quad n = 2, 3, \dots \quad (2.13)$$

From Statement (2.11) we obtain

$$Pr\{N(t) = k\} = Pr\{N(t) \geq k\} - Pr\{N(t) \geq k+1\} \quad (2.14)$$

$$= Pr\{W_k \leq t\} - Pr\{W_{k+1} \leq t\}. \quad (2.15)$$

Therefore, from (2.13) $Pr\{N(t) = k\} = F_k(t) - F_{k+1}(t)$, $k = 1, \dots$. Now, let k' be given, we find that $Pr\{1 \leq N(t) < k'\} = \sum_{j=1}^{k'-1} (F_j(t) - F_{j+1}(t)) = F_1(t) - F_{k'}(t)$ which converges to 0 as $t \rightarrow \infty$. Since this is true for each k' , $N(t) \rightarrow \infty$ *a.s.* as $t \rightarrow \infty$. Now,

$$\sqrt{t}(\bar{\xi}_t - \mu) = \sqrt{\frac{t}{N(t)}} \sqrt{N(t)}(\bar{\xi}_t - \mu). \quad (2.16)$$

In the previous expression, $\sqrt{\frac{t}{N(t)}}$ converges in probability to $\sqrt{\tau}$ by (2.6) and the remaining portion converges in distribution to $N(0, \sigma^2)$ by the Central Limit Theorem since $N(t) \rightarrow \infty$ as $t \rightarrow \infty$. Thus, it follows from Slutsky's theorem that $\sqrt{t}(\bar{\xi}_t - \mu)$ converges in distribution to $N(0, \tau\sigma^2)$. \square

Remark 2.9 Note that (2.9) is an asymptotically exact result and implies that for large t , $\bar{\xi}_t$ is *approximately* distributed as a normal distribution with mean μ and approximate variance $\frac{\tau\sigma^2}{t-\gamma}$. Recall, γ is the expected time required to initialize the simulation and no realizations of ξ are generated during this time, thus we subtract it from t in the denominator of the approximate variance term. We assume that t is large enough so that $\bar{\xi}_t$ is *approximately* distributed as a normal distribution and that the overhead γ is not negligible. Note that this expression makes intuitive sense. Since the expected time to generate an observation of ξ is τ , the expected number of copies of ξ we should have after t units of time is $\frac{t-\gamma}{\tau}$. Hence, we expect the variance of ξ to be approximately $\frac{\tau\sigma^2}{t-\gamma}$. Since we are interested in how overhead affects the simulation, we will use, for fixed t , $\frac{\tau\sigma^2}{t-\gamma}$ as a measure of simulation efficiency. We will refer to this measure of simulation efficiency and refer to it as the *asymptotic approximation of variance AAV*.

2.2 Theoretical Development

An experimenter is interested in estimating the parameter μ_X . He or she has available the following three simulation programs for the task: simulation A, simulation B and simulation C.

Simulation A is a direct crude Monte Carlo simulation that simply generates $\{X_1, X_2, \dots\}$ i.i.d. replicates of X . This simulation consists of the overhead associated with setting up the experiment and the actual i.i.d. generation of the random variable X . We assume that the time to generate a random observation of X is random. To this end, let $\tau_x^A > 0$ represent the finite mean time it takes to generate a copy of X . Also, the overhead time may also vary and so we let κ_A denote the finite mean of the random overhead time for simulation A. Finally, let $N_A(t)$ be the delayed renewal process associated with the generation of replicates of X . That is, the first event or epoch occurs with the completion of overhead and the generation of X_1 . The next epoch occurs with the generation of X_2 , and so on. Thus, simulation A generates the following statistics:

$$\bar{X}_t = \frac{1}{N_A(t)} \sum_{i=1}^{N_A(t)} X_i,$$

and

$$\sigma_X^2(t) = \frac{t}{N_A(t)} \left(\frac{1}{N_A(t)} \sum_{i=1}^{N_A(t)} X_i^2 - \bar{X}_t^2 \right).$$

Simulation B also estimates μ_X , but it keeps track of additional quantities in the hope that the overhead associated with collecting them pays off in improved simulation efficiency. Simulation B provides the following statistics:

$$(\tilde{X}_t, \tilde{Y}_t, \tilde{\sigma}_X^2(t), \tilde{\sigma}_Y^2(t), \tilde{\sigma}_{XY}(t)),$$

which are defined below. κ_B is the mean overhead time associated with simulation B. Analogously to simulation A, let $t_{xy} > 0$ represent the finite mean time required to generate one replicate of the pair (X, Y) . This simulation also has a delayed renewal process associated with it, namely $N_B(t)$, where the first event represents the time to complete the overhead and generate the first pair (X_1, Y_1) . The remaining events correspond to generating further copies of (X, Y) . Thus, we can write

$$(\tilde{X}_t, \tilde{Y}_t) = \frac{1}{N_B(t)} \sum_{i=1}^{N_B(t)} (X_i, Y_i), \quad (2.17)$$

$$\sigma_{\tilde{X}}^2(t) = \frac{t}{N_B(t)} \left(\frac{1}{N_B(t)} \sum_{i=1}^{N_B(t)} \tilde{X}_i^2 - \tilde{X}_t^2 \right) \quad (2.18)$$

and

$$\sigma_{\tilde{Y}}^2(t) = \frac{t}{N_B(t)} \left(\frac{1}{N_B(t)} \sum_{i=1}^{N_B(t)} \tilde{Y}_i^2 - \tilde{Y}_t^2 \right). \quad (2.19)$$

In the classical method of control variates, simulation B is used to construct

$$\tilde{X}_t^* = \tilde{X}_t - \frac{\tilde{\sigma}_{XY}(t)}{\tilde{\sigma}_Y^2(t)} (\tilde{Y}_t - \mu_Y). \quad (2.20)$$

Note that this method requires that μ_Y is known exactly. There are many examples where μ_Y is unknown, but ρ_{XY}^2 is (potentially) large, suggesting a potential for a significant improvement in efficiency by using simulation B instead of A.

We propose to use simulation C in conjunction with simulation B to construct an estimator that is analogous to (2.20) when μ_Y is unknown. Simulation C estimates μ_Y and produces the following statistics:

$$\bar{Z}_t = \frac{1}{N_C(t)} \sum_{i=1}^{N_C(t)} Z_i, \quad (2.21)$$

and

$$\sigma_Z^2(t) = \frac{t}{N_C(t)} \left(\frac{1}{N_C(t)} \sum_{i=1}^{N_C(t)} Z_i^2 - \bar{Z}_t^2 \right). \quad (2.22)$$

where Z is a consistent estimator for μ_Y , $N_C(t)$ is the associated delayed renewal process, κ_C is the associated mean overhead and $\tau_z > 0$ is the finite mean time to generate one copy of Z . Let $\kappa \equiv \kappa_B + \kappa_C$ and $q \equiv 1 - p$. We can divide t units of time between simulations B and C and construct an estimator for μ_X as follows:

$$\bar{Q}_t(p, \alpha) = \tilde{X}_{qt} + \alpha(\tilde{Y}_{qt} - \bar{Z}_{pt}) \text{ for } t > 0. \quad (2.23)$$

Note, we will often eliminate the subscripts of \tilde{X}_{qt} , \tilde{Y}_{qt} and \bar{Z}_{pt} to facilitate the presentation.

The implication of the Remark 2.9 is that now we have derived the AAV of $\bar{\xi}_t$. Thus, fortified with this knowledge we apply it to calculating the AAV of $\bar{Q}_t(p, \alpha)$ given as follows:

$$v^2(p, \alpha) = \frac{\tau_{xy}\sigma_X^2}{(t - \kappa)(1 - p)} + \alpha^2 \left[\frac{\tau_{xy}\sigma_Y^2}{(t - \kappa)(1 - p)} + \frac{\tau_z\sigma_Z^2}{(t - \kappa)p} \right] + \frac{2\alpha\tau_{xy}\sigma_{XY}}{(t - \kappa)(1 - p)}. \quad (2.24)$$

where p is the fraction of time executing simulation C, $q \equiv 1 - p$ is the fraction of time spent executing simulation B and $\alpha \in \mathfrak{R}$.

Thus, we have found $v^2(p, \alpha)$ as a function of p and α . Note, that this expression is the AAV of $\bar{Q}_t(p, \alpha)$ as long as $0 < p < 1$. If $p = 0$ then we must have $\alpha = 0$ so that $\bar{Q}_t(p, \alpha)$ reduces to \bar{X}_t . If $p = 1$ then $\bar{Q}_t(p, \alpha)$ does not estimate μ_X , so this case is not relevant.

Remark 2.10 In order to avoid trivial or pathological situations, we assume that σ_X^2, σ_Y^2 and σ_Z^2 are positive and finite; otherwise, the quasi control variate procedure would not be applicable or appropriate. In addition, we assume that $0 < \rho_{XY}^2 < 1$.

One natural question arises. How long should we estimate the control variate mean with (2.21) (simulation C) and continue with our estimation of μ_X with the main simulation (2.17) (simulation B) to arrive at our final estimate \bar{Q}_t ? We sacrifice the time that could be spent estimating μ_X , by allocating too

much effort to the estimation of μ_Y by \bar{Z}_{pt} . Likewise, if too little time is spent estimating μ_Y , then \bar{X}_t (simulation A) may be a better estimate of μ_X than \bar{Q}_t . Hence, we expect there exists an optimal proportion of time that should be devoted to estimating each mean. We formulate this problem by obtaining the optimal proportion p^* and coefficient α^* (the optimal QCV parameters) required to minimize the AAV of the estimator \bar{Q}_t .

We first offer a definition and a common theorem [41] that relate to optimization.

Definition 2.11 Given a convex set Ω , a function f on the set Ω is strictly convex if $x^1, x^2 \in \Omega$, $x^1 \neq x^2$ implies

$$f(\lambda x^1 + (1 - \lambda)x^2) < \lambda f(x^1) + (1 - \lambda)f(x^2) \text{ where } 0 < \lambda < 1.$$

Theorem 2.12 Let $f \in C^2$ be a strictly convex function defined on a region Ω in which the point x^* is an interior point. Suppose in addition that $\nabla f(x^*) = 0$. Then x^* is the unique global minimum of f over Ω .

The following theorem establishes an important property of v^2 and will be useful in finding the the optimal pair (p^*, α^*) :

Theorem 2.13 $v^2(p, \alpha)$ defined by (2.24) is a strictly convex function over the convex set $S = \{(p, \alpha) | -\infty < \alpha < \infty, 0 < p < 1\}$.

Proof: It is sufficient to show that the Hessian of v^2 , $\nabla v^2(p, \alpha)$, is positive definite for all points in S. Recall from linear algebra that the Hessian of $v^2(p, \alpha)$,

$$\nabla^2 v^2(p, \alpha) = \begin{bmatrix} \frac{\partial^2 v^2}{\partial \alpha^2} & \frac{\partial^2 v^2}{\partial \alpha \partial p} \\ \frac{\partial^2 v^2}{\partial p \partial \alpha} & \frac{\partial^2 v^2}{\partial p^2} \end{bmatrix},$$

is positive definite on S if and only if

$$\begin{aligned} \text{Det}(\nabla^2 v^2(p, \alpha)) &= \frac{\partial^2 v^2(p, \alpha)}{\partial \alpha^2} \frac{\partial^2 v^2(p, \alpha)}{\partial p^2} - \left(\frac{\partial^2 v^2(p, \alpha)}{\partial p \partial \alpha} \right)^2 > 0 \\ &\quad \frac{\partial^2 v^2(p, \alpha)}{\partial \alpha^2} > 0 \text{ and} \\ &\quad \forall (\alpha, p) \in S. \end{aligned}$$

We find that

$$\begin{aligned} \frac{\partial^2 v^2(p, \alpha)}{\partial \alpha^2} &= 2 \left[\frac{\tau_{xy} \sigma_Y^2}{(t - \kappa)(1 - p)} + \frac{\tau_z \sigma_Z^2}{(t - \kappa)p} \right] > 0, \\ \frac{\partial^2 v^2(p, \alpha)}{\partial p^2} &= \frac{2(\alpha p^2 + \alpha(\alpha(\tau_{xy} \sigma_X^2 p^3 - \tau_z \sigma_Z^2 (p - 1)^3) + 2\tau_{xy} \sigma_{XY} p^3))}{(t - \kappa)p^3(p - 1)^3} \end{aligned}$$

and

$$\frac{\partial^2 v^2(p, \alpha)}{\partial \alpha \partial p} = \frac{2(\alpha(\tau_{xy} \sigma_Y^2 p^2 - \tau_z \sigma_Z^2 (p^2 - 2p + 1)) + \tau_{xy} \sigma_{XY} p^2)}{(t - \kappa)p^2(p - 1)^2} \quad \forall p \in (0, 1).$$

We find

$$\begin{aligned}
& \text{Det}(\nabla^2 v^2(p, \alpha)) \\
&= 4 \left[\tau_{xy} \sigma_X^2 p^2 (\tau_{xy} \sigma_Y^2 p + \tau_z \sigma_Z^2 (1-p)) + \alpha^2 \tau_{xy} \sigma_Y^2 \tau_z \sigma_Z^2 (1-p) \right. \\
&+ \left. \tau_{xy} \sigma_{XY} p (2\alpha \tau_z \sigma_Z^2 (1-p) - \tau_{xy} \sigma_{XY} p^2) \right] / ((t - \kappa)^2 p^3 (1-p)^4).
\end{aligned}$$

Note that for $p \in (0, 1)$ this expression is positive if and only if the numerator is positive which is given as follows:

$$\begin{aligned}
& 4[\tau_{xy} \sigma_X^2 p^2 (\tau_{xy} \sigma_Y^2 p + \tau_z \sigma_Z^2 (1-p)) + \alpha^2 \tau_{xy} \sigma_Y^2 \tau_z \sigma_Z^2 (1-p) \\
&+ \tau_{xy} \sigma_{XY} p (2\alpha \tau_z \sigma_Z^2 (1-p) - \tau_{xy} \sigma_{XY} p^2)] \\
&= 4[\tau_{xy}^2 (\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2) p^3 + \tau_{xy} \sigma_X^2 \tau_z \sigma_Z^2 p^2 (1-p) + \alpha^2 \tau_{xy} \sigma_Y^2 \tau_z \sigma_Z^2 (1-p) \\
&+ 2\alpha \tau_{xy} \sigma_{XY} \tau_z \sigma_Z^2 p (1-p)] \\
&= 4^2 \tau_{xy} (\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2) p^3 + 4(1-p) \tau_{xy} \tau_z \sigma_Z^2 [(\sigma_X t)^2 + (\sigma_Y \alpha)^2 + 2\sigma_{XY} p \alpha] \\
&= 4\tau_{xy}^2 (\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2) p^3 + 4(1-p) \tau_{xy} \tau_z \sigma_Z^2 \sigma_X \sigma_Y \left(\frac{\sigma_X}{\sigma_Y} p^2 + \frac{\sigma_Y}{\sigma_X} \alpha^2 + 2 \frac{\sigma_{XY}}{\sigma_X \sigma_Y} p \alpha \right) \\
&= 4\tau_{xy}^2 \sigma_X^2 \sigma_Y^2 (1 - \rho^2) p^3 + 4(1-p) \tau_{xy} \tau_z \sigma_Z^2 \sigma_X \sigma_Y \left(\frac{\sigma_X}{\sigma_Y} p^2 + \frac{\sigma_Y}{\sigma_X} \alpha^2 + 2\rho p \alpha \right) \\
&\left\{ \begin{aligned} &= 4(1-p) \tau_{xy} \tau_z \sigma_Z^2 \sigma_X \sigma_Y (p \sqrt{\frac{\sigma_X}{\sigma_Y}} + \alpha \sqrt{\frac{\sigma_Y}{\sigma_X}})^2 & \text{if } \rho = 1, \\ &= 4(1-p) \tau_{xy} \tau_z \sigma_Z^2 \sigma_X \sigma_Y (p \sqrt{\frac{\sigma_X}{\sigma_Y}} - \alpha \sqrt{\frac{\sigma_Y}{\sigma_X}})^2 & \text{if } \rho = -1, \\ &\geq (2\tau_{xy} \sigma_X \sigma_Y)^2 (1 - \rho^2) p^3 > 0 & \text{otherwise.} \end{aligned} \right. \quad (2.25)
\end{aligned}$$

Thus, from the above expressions we can conclude that $\nabla^2 v^2(p, \alpha)$ is positive definite on S if $|\rho| < 1$. \square

We now give the biggest result from this section, namely the Theorem that establishes what the optimal QCV parameters, (p^*, α^*) are.

Theorem 2.14 Let $r = \frac{\sqrt{\tau_z}\sigma_Z}{\sqrt{\tau_{xy}}\sigma_Y}$. The function $v^2(p, \alpha)$ has a unique global minimum at (p^*, α^*) subject to the condition $p \in [0, 1)$ where

$$p^* = \begin{cases} \frac{\sqrt{r^2\rho_{XY}^2(1-r^2)(1-\rho_{XY}^2)}-r^2(1-r^2)}{(1-r^2)(1-r^2-\rho_{XY}^2)} & \text{if } r^2 + \rho_{XY}^2 \neq 1 \text{ and } \rho_{XY}^2 > r^2 \\ 1 - \frac{1}{2\rho_{XY}^2} & \text{if } r^2 + \rho_{XY}^2 = 1 \text{ and } \rho_{XY}^2 > r^2 \\ 0 & \text{if } \rho_{XY}^2 \leq r^2, \end{cases} \quad (2.26)$$

and $\alpha^* = -\frac{\sigma_{XY}}{\sigma_Y^2}\tilde{\alpha}^*$, where

$$\tilde{\alpha}^* = \begin{cases} \frac{\sqrt{\rho_{XY}^2(1-r^2)(1-\rho_{XY}^2)}+r(r^2-1)}{(1-r^2)[\sqrt{\rho_{XY}^2(1-r^2)(1-\rho_{XY}^2)}-r\rho_{XY}^2]} & \text{if } r^2 + \rho_{XY}^2 \neq 1 \text{ and } \rho_{XY}^2 > r^2 \\ \frac{2\rho_{XY}^2-1}{\rho_{XY}^2} & \text{if } r^2 + \rho_{XY}^2 = 1 \text{ and } \rho_{XY}^2 > r^2 \\ 0 & \text{if } \rho_{XY}^2 \leq r^2. \end{cases} \quad (2.27)$$

Proof: We first show that if $\rho_{XY}^2 > r^2$ then p^* is an interior point of $[0, 1)$.

We consider the following three cases: (i) $\rho_{XY}^2 > \max(r^2, 1 - r^2)$, (ii) $r^2 < \min(\rho_{XY}^2, 1 - \rho_{XY}^2)$ and (iii) $r^2 + \rho_{XY}^2 = 1, \rho_{XY}^2 > r^2$. In case (i) we have that $1 - r^2 < \rho_{XY}^2$ which implies that $\sqrt{1 - r^2} < |\rho_{XY}|$ and $\sqrt{1 - \rho_{XY}^2} < r$. Thus,

we have $\sqrt{(1 - r^2)(1 - \rho_{XY}^2)} < r|\rho_{XY}| \Rightarrow$

$$\begin{aligned} & 1 - \rho_{XY}^2 < r|\rho_{XY}|\sqrt{\frac{1 - \rho_{XY}^2}{1 - r^2}} \\ \Rightarrow & 1 - \rho_{XY}^2 - r^2 < r|\rho_{XY}|\sqrt{\frac{1 - \rho_{XY}^2}{1 - r^2}} - r^2 \\ \Rightarrow & (1 - r^2)(1 - \rho_{XY}^2 - r^2) < \sqrt{r^2\rho_{XY}^2(1 - r^2)(1 - \rho_{XY}^2)} - r^2(1 - r^2) \\ \Rightarrow & 1 > p^*, \end{aligned}$$

where the last implication follows by dividing both sides of the inequality by the negative quantity $(1-r^2)(1-\rho_{XY}^2-r^2)$. Case (ii) is similar to case (i) except that the sense of the inequalities in the series of implications above is reversed with the exception of the last statement. Case (iii) is trival. Thus, we have proven that $p^* < 1$. We now show $p^* > 0$. Suppose we are in cases (i) or (ii). Note that the numerator of p^* is $(\geq) 0 \iff \sqrt{r^2\rho_{XY}^2(1-r^2)(1-\rho_{XY}^2)}$
 $(\geq) r^2(1-r^2) \iff \sqrt{\rho_{XY}^2(1-\rho_{XY}^2)} (\geq) \sqrt{r^2(1-r^2)} \iff \rho_{XY}^2(1-\rho_{XY}^2)$
 $(\geq) r^2(1-r^2) \iff \rho_{XY}^2(1-\rho_{XY}^2) - r^2(1-r^2) (\geq) 0$. Thus, $p^* > 0$ if and only if the product of the previous expression and the denominator of p^* is positive. We find $(\rho_{XY}^2(1-\rho_{XY}^2) - r^2(1-r^2))(1-r^2)(1-\rho_{XY}^2-r^2) = (1-r^2)(\rho_{XY}^2-r^2)(r^2+\rho_{XY}^2-1)^2$ which is positive if and only if $\rho_{XY}^2 > r^2$. Case (iii) implies that $\rho_{XY}^2 > r^2 = 1-\rho_{XY}^2 \Rightarrow \rho_{XY}^2 > 1/2$ thus, (2.26b) gives $p^* > 0$.

We now show that $\nabla v^2(p^*, \alpha^*) = 0$. We now find and set $\nabla v^2(p, \alpha)$ equal to the zero vector.

$$\begin{aligned} \frac{\partial v^2(p, \alpha)}{\partial \alpha} &= \frac{1}{(t-\kappa)} \left(2\alpha \left[\frac{\tau_{xy}\sigma_Y^2}{1-p} + \frac{\tau_z\sigma_Z^2}{p} \right] \right. \\ &\quad \left. + \frac{2\tau_{xy}\sigma_{XY}}{(1-p)} \right) = 0 \text{ and (2.28)} \\ \frac{\partial v^2(p, \alpha)}{\partial p} &= \frac{1}{(t-\kappa)} \left(\frac{\tau_{xy}\sigma_X^2}{(1-p)^2} + \alpha^2 \left[\frac{\tau_{xy}\sigma_Y^2}{(1-p)^2} - \frac{\tau_z\sigma_Z^2}{p^2} \right] \right) \end{aligned}$$

$$+ \frac{2\alpha\tau_{xy}\sigma_{XY}}{(1-p)^2} = 0. \quad (2.29)$$

Solving for α in (2.28) above yields the following:

$$\alpha = \frac{-\tau_{xy}\sigma_{XY}p}{\tau_{xy}\sigma_Y^2p + \tau_z\sigma_Z^2(1-p)} \quad (2.30)$$

$$= \frac{-\sigma_{XY}}{\sigma_Y^2} \left(\frac{p}{p + (1-p)r^2} \right). \quad (2.31)$$

For notational convenience let

$$\beta(p) \equiv \frac{p}{p + (1-p)r^2}, \quad (2.32)$$

so that

$$\alpha = \frac{-\sigma_{XY}}{\sigma_Y^2} \beta(p). \quad (2.33)$$

Now, substituting this into equation (2.29) above yields the following:

$$\begin{aligned} & \tau_{xy}(\sigma_X^2(\tau_{xy}\sigma_Y^2p + (1-p)\tau_z\sigma_Z^2)^2 - \tau_{xy}\sigma_{XY}^2(\tau_{xy}\sigma_Y^2p^2 \\ & \qquad \qquad \qquad + (1-p)(\tau_z\sigma_Z^2(p+1)))) = 0 \end{aligned} \quad (2.34)$$

$$\begin{aligned} \iff & \sigma_X^2\sigma_Y^2 \left[\left(\frac{\tau_{xy}\sigma_Y^2p + (1-p)\tau_z\sigma_Z^2}{\sigma_Y} \right)^2 - \tau_{xy}\rho_{XY}^2(\sigma_Y^2\tau_{xy}p^2 \right. \\ & \qquad \qquad \qquad \left. + (1-p)(\tau_z\sigma_Z^2(p+1))) \right] = 0 \end{aligned} \quad (2.35)$$

$$\Leftrightarrow \left[\left(\frac{\tau_{xy}\sigma_Y^2 p + (1-p)(\tau_z\sigma_Z^2)}{\sigma_Y} \right)^2 - \tau_{xy}^2 \rho_{XY}^2 \sigma_Y^2 (p^2 + (1-p)(r^2(p+1))) \right] = 0 \quad (2.36)$$

$$\Leftrightarrow \left[\left(\frac{\tau_{xy}\sigma_Y^2 p + (1-p)(\tau_z\sigma_Z^2)}{\sigma_Y^2} \right)^2 - \tau_{xy}^2 \rho_{XY}^2 (p^2 + (1-p)(r^2(p+1))) \right] = 0 \quad (2.37)$$

$$\Leftrightarrow \left[\left(\tau_{xy}p + (1-p)\tau_{xy}(r^2) \right)^2 - \tau_{xy}^2 \rho_{XY}^2 (p^2 + (1-p)(r^2(p+1))) \right] = 0 \quad (2.38)$$

$$\Leftrightarrow \tau_{xy}^2 \left[\left(p + (1-p)(r^2) \right)^2 - \rho_{XY}^2 (p^2 + (1-p)r^2) \right] = 0 \quad (2.39)$$

$$\Leftrightarrow \rho_{XY}^2 = \frac{\beta(p^2)}{\beta(p)^2}. \quad (2.40)$$

Using the quadratic formula we can solve equation (2.40) for p and obtain expressions (2.26a,b). Now, using equation (2.31) we find α^* to arrive at expressions (2.27a,b). Thus, it follows from Theorems 2.12 and 2.13 that (p^*, α^*) is the unique global minimum of v^2 over S when $\rho_{XY}^2 > r^2$. If $\rho_{XY}^2 \leq r^2$ then we note from expression (2.31) that for fixed $p \in (0, 1)$ the optimal QCV coefficient is the following:

$$\alpha_p^* = \frac{-\sigma_{XY}}{\sigma_Y^2} \left(\frac{p}{p + (1-p)r^2} \right). \quad (2.41)$$

Substituting this expression into (2.24) yields

$$v^2(p, \alpha_p^*) = \frac{\tau_{xy}\sigma_X^2}{(t-k)(1-p)} \left(1 - \frac{\rho}{p + (1-p)r^2}\right). \quad (2.42)$$

It follows that

$$v^2(p, \alpha_p^*) < \sigma_X^2 \iff 0 < p < \frac{\rho_{XY}^2 - r^2}{1 - r^2}. \quad (2.43)$$

Thus, if $r^2 \geq \rho_{XY}^2$, then $v^2(p, \alpha_p^*) \geq \sigma_X^2$ for every $p \in (0, 1)$, which implies that $p^* = 0$ and therefore $\alpha^* = 0$ as well. \square

The significance of the previous theorem is that we now have in closed form the values of (p^*, α^*) in terms of a common statistical parameter, ρ_{XY}^2 and a parameter r^2 whose interpretation is given next. Also, it is important to note that the clever algebra and factoring in (2.34) through (2.40) renders this possible; otherwise, the formulas for (p^*, α^*) would have been confusing, disordered functions of $\tau_{xy}, \tau_z, \sigma_X^2, \sigma_Y^2, \sigma_Z^2$ and σ_{XY} .

Remark 2.15 The definition of r has an important interpretation. One should think of r as the ratio of the “effort” required to generate a realization of the control variate Z versus a copy of (X, Y) in the main simulation. As will be seen in the following sections, for the quasi control variate approach to be beneficial, r needs to be small.

The following proposition offers an important interpretation of $\rho_{XY}^2\beta(p^*)$. We can interpret $\rho_{XY}^2\beta(p^*)$ as the asymptotic squared correlation

between \tilde{X}_{q^*t} and $\tilde{Y}_{q^*t} - \bar{Z}_{p^*t}$.

Proposition 2.16 Let $\bar{W}_t = \tilde{Y}_{q^*t} - \bar{Z}_{p^*t}$, then $\rho_{\tilde{X}_{q^*t} \bar{W}_t}^2 \rightarrow \rho_{\tilde{X} \tilde{Y}}^2 \beta(p^*)$ a.s.

as $t \rightarrow \infty$.

Proof: (Note: to facilitate notation, we will often omit subscripts when the context is clear.) We first show that $\sigma_{\tilde{X}, \tilde{Y} - \bar{Z}}^2 = \sigma_{\tilde{X} \tilde{Y}}^2$ to justify the step from (2.44) to (2.45). Independent of t we have

$$\begin{aligned}
\sigma_{\tilde{X}, \tilde{Y} - \bar{Z}}^2 &= E[\tilde{X}(\tilde{Y} - \bar{Z})] - E[\tilde{X}]E[\tilde{Y} - \bar{Z}] \\
&= E[\tilde{X} \tilde{Y}] - E[\tilde{X} \bar{Z}] - E[\tilde{X}]E[\tilde{Y} - \bar{Z}] \\
&= E[\tilde{X} \tilde{Y}] - E[\tilde{X}]E[\bar{Z}] - E[\tilde{X}](E[\tilde{Y}] - [\bar{Z}]) \\
&= E[\tilde{X} \tilde{Y}] - E[\tilde{X}]E[\bar{Z}] \\
&= E[\tilde{X} \tilde{Y}] - E[\tilde{X}]E[\tilde{Y}] \\
&= \sigma_{\tilde{X} \tilde{Y}}^2.
\end{aligned}$$

Thus,

$$\lim_{t \rightarrow \infty} \rho_{\tilde{X} \bar{W}_t}^2 = \lim_{t \rightarrow \infty} \rho_{\tilde{X}, \tilde{Y} - \bar{Z}}^2(t) \quad (2.44)$$

$$= \lim_{t \rightarrow \infty} \frac{(\sigma_{\tilde{X} \tilde{Y}})^2}{\sigma_{\tilde{X}}^2 (\sigma_{\tilde{Y}}^2 + \sigma_{\bar{Z}}^2)} \quad (2.45)$$

$$= \lim_{t \rightarrow \infty} \frac{\left(\frac{\sigma_{\tilde{X} \tilde{Y}}}{N_B(tq^*)}\right)^2}{\frac{\sigma_{\tilde{X}}^2}{N_B(tq^*)} \left(\frac{\sigma_{\tilde{Y}}^2}{N_B(tq^*)} + \frac{\sigma_{\bar{Z}}^2}{N_C(tp^*)}\right)} \quad (2.46)$$

$$= \lim_{t \rightarrow \infty} \frac{\frac{\sigma_{\tilde{X} \tilde{Y}}^2}{N_B(tq^*)}}{\sigma_{\tilde{X}}^2 \sigma_{\tilde{Y}}^2 \left[\frac{N_C(tp^*) + N_B(tq^*) r^2 \tau_{xy}}{N_B(tq^*) N_C(tp^*) \tau_z} \right]} \quad (2.47)$$

$$= \lim_{t \rightarrow \infty} \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2 \left[1 + \frac{N_B(tq^*) \tau_{xy} r^2}{N_C(tp^*) \tau_z} \right]} \quad (2.48)$$

We now show that the expression in brackets in equation (2.48) converges a.s.

to $1/\beta(p^*)$. Note that

$$\begin{aligned} & 1 + \frac{N_B(tq^*) \tau_{xy} r^2}{N_C(tp^*) \tau_z} \\ &= 1 + \frac{tp^*}{N_C(tp^*)} \frac{N_B(tq^*) \tau_{xy}}{tq^*} \frac{tq^*}{\tau_z} \frac{r^2}{tp^*} \end{aligned}$$

and that it follows from the Elementary Renewal Theorem (2.6) that the first

two terms in this last expression converge to $\frac{\tau_z}{\tau_{xy}}$ a.s. as $t \rightarrow \infty$. Hence, $\rho_{\bar{X}}^2 \bar{W}(t)$

converges to

$$\frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2 \left[1 + \frac{q^* r^2}{p^*} \right]},$$

which is equal to $\rho_{XY}^2 \beta(p^*)$. \square

The following Lemma is extremely important in that it establishes precisely what the AAV of $Q_t(p^*, \alpha^*)$ is.

Lemma 2.17 The optimal AAV for $Q_t(p^*, \alpha^*)$ is the following:

$$\frac{\tau_{xy} \sigma_X^2}{(t - \kappa)(1 - p^*)} (1 - \rho_{XY}^2 \beta(p^*)) \quad (2.49)$$

Proof:

$$\begin{aligned} & v^2(p^*, \alpha^*) \\ &= \frac{\tau_{xy} \sigma_X^2}{(t - \kappa)(1 - p^*)} + \left(\frac{-\sigma_{XY}}{\sigma_Y^2} \beta(p^*) \right)^2 \frac{1}{t - \kappa} \left[\frac{\tau_{xy} \sigma_Y^2}{(1 - p^*)} + \frac{\tau_z \sigma_Z^2}{p^*} \right] \end{aligned}$$

$$+ \frac{2\left(\frac{-\sigma_{XY}}{\sigma_Y^2}\beta(p^*)\right)\tau_{xy}\sigma_{XY}}{(t-\kappa)(1-p^*)} \quad (2.50)$$

$$= \frac{1}{t-\kappa} \left[\frac{\tau_{xy}\sigma_X^2}{1-p^*} + \frac{\tau_{xy}\sigma_{XY}^2(\beta^2(p^*) - 2\beta(p^*))}{\sigma_Y^2(1-p^*)} + \frac{\tau_z\sigma_Z^2\sigma_{XY}^2\beta^2(p^*)}{\sigma_Y^4 p^*} \right] \quad (2.51)$$

$$= \frac{1}{t-\kappa} \left[\frac{\tau_{xy}\sigma_X^2}{1-p^*} + \tau_{xy}\sigma_X^2 \left(\frac{\rho_{XY}^2(\beta^2(p^*) - 2\beta(p^*))}{(1-p^*)} + \frac{\rho_{XY}^2\beta^2(p^*)r^2}{p^*} \right) \right] \quad (2.52)$$

$$= \frac{1}{t-\kappa} \left[\frac{\tau_{xy}\sigma_X^2}{1-p^*} + \tau_{xy}\sigma_X^2 \left(\frac{(\beta(p^*) - 2)}{1-p^*} + \frac{\beta(p^*)r^2}{p^*} \right) \rho_{XY}^2\beta(p^*) \right] \quad (2.53)$$

$$= \frac{1}{t-\kappa} \left[\frac{\tau_{xy}\sigma_X^2}{1-p^*} - \frac{\tau_{xy}\sigma_X^2\rho_{XY}^2\beta(p^*)}{1-p^*} \right] \quad (2.54)$$

$$= \frac{\tau_{xy}\sigma_X^2}{(t-\kappa)(1-p^*)} (1 - \rho_{XY}^2\beta(p^*)). \quad \square \quad (2.55)$$

We now determine for which values of ρ and r that $p^* > 0$.

Theorem 2.18 The following statements are equivalent:

(a) $p^* > 0$ (It is optimal to devote some time to simulation C.),

(b) $v^2(p^*, \alpha^*) < \frac{\tau_{xy}\sigma_X^2}{t-k}$ (The optimal QCV procedure has a lower AAV than simulation B alone.),

(c) $r^2 < \rho_{XY}^2$.

Proof: The proof of Theorem (2.14) shows that (a) and (c) are equivalent.

Since p^* and α^* are the unique optimal parameters, we have $v^2(p^*, \alpha^*) <$

$v^2(0, 0) = \frac{\tau_{xy}\sigma_X^2}{t-k}$ when $p^* > 0$. If $p^* = 0$ then (2.43) implies that $\alpha^* = 0$,

so (a) and (b) are equivalent. \square

We now state a Central Limit Theorem for $\bar{Q}_t(p^*, \alpha^*)$. This theorem is important in that we establish the approximate distribution of \bar{Q}_t for large

t . This is important in constructing confidence intervals for \bar{Q}_t .

Theorem 2.19 As $t \rightarrow \infty$, $\sqrt{t}(\bar{Q}_t - \mu_X) \Rightarrow N\left(0, \frac{\tau_{xy}\sigma_X^2}{(1-p^*)}(1 - \rho_{XY}^2\beta(p^*))\right)$.

Proof:

$$\begin{aligned}
& \sqrt{t}(\bar{Q}_t(p^*, \alpha^*) - \mu_X) \\
&= \sqrt{t}(\tilde{X}_{tq^*} + \alpha^*(\tilde{Y}_{tq^*} - \bar{Z}_{tp^*}) - \mu_X) \\
&= \sqrt{t}(\tilde{X}_{tq^*} - \mu_X) + \alpha^*\sqrt{t}(\tilde{Y}_{tq^*} - \mu_Y) - \alpha^*\sqrt{t}(\bar{Z}_{tp^*} - \mu_Y) \\
&= \frac{\sqrt{t}\sqrt{N_C(tp^*)}}{\sqrt{N_C(tp^*)}} \left[(\tilde{X}_{tq^*} - \mu_X) + \alpha^*(\tilde{Y}_{tq^*} - \mu_Y) \right] \\
&\quad - \alpha^* \frac{\sqrt{t}\sqrt{N_B(tq^*)}}{\sqrt{N_B(tq^*)}} (\bar{Z}_{tp^*} - \mu_Y) \\
&= \frac{\sqrt{t}}{\sqrt{N_C(tp^*)}} \left[\sqrt{N_C(tp^*)} (\tilde{X}_{tq^*} + \alpha^*\tilde{Y}_{tq^*} - (\mu_X + \alpha^*\mu_Y)) \right] \\
&\quad - \alpha^* \frac{\sqrt{t}}{\sqrt{N_B(tq^*)}} \left[\sqrt{N_B(tq^*)} (\bar{Z}_{tp^*} - \mu_Y) \right] \\
&= \frac{\sqrt{tq^*}}{\sqrt{N_C(tp^*)}} \frac{\sqrt{t}}{\sqrt{tq^*}} \left[\sqrt{N_C(tp^*)} (\tilde{X}_{tq^*} + \alpha^*\tilde{Y}_{tq^*} - (\mu_X + \alpha^*\mu_Y)) \right] \\
&\quad - \alpha^* \frac{\sqrt{tp^*}}{\sqrt{N_B(tq^*)}} \frac{\sqrt{t}}{\sqrt{tp^*}} \left[\sqrt{N_B(tq^*)} (\bar{Z}_{tp^*} - \mu_Y) \right]
\end{aligned}$$

Note, the coefficients of the bracketed expressions in the last equality:

$$\frac{\sqrt{tq^*}}{\sqrt{N_C(tp^*)}} \frac{\sqrt{t}}{\sqrt{tq^*}}$$

and

$$\frac{\sqrt{tp^*}}{\sqrt{N_B(tq^*)}} \frac{\sqrt{t}}{\sqrt{tp^*}}$$

converge a.s. to $\sqrt{\frac{\tau_z}{p^*}}$ and $\sqrt{\frac{\tau_{xy}}{q^*}}$, respectively as $t \rightarrow \infty$. Also, the associated bracketed expressions converge in distribution to a $N(0, \sigma_X^2 + \sigma_Y^2 + 2\alpha^* \sigma_{XY})$ and $N(0, \sigma_Z^2)$, respectively. Thus, it follows from Slutsky's Theorem, and the independence of (X, Y) and Z that as $t \rightarrow \infty$,

$$\sqrt{t}(\bar{Q}_t - \mu_X) \implies N\left(0, \frac{\tau_{xy}}{1-p^*}(\sigma_X^2 + (\alpha^* \sigma_Y)^2 + 2\alpha^* \sigma_{XY}) + \frac{\tau_z}{p}(\alpha^* \sigma_Z)^2\right).$$

However, the variance given in the distribution above can be simplified to $\frac{\tau_{xy} \sigma_X^2}{(1-p^*)}(1 - \rho_{XY}^2 \beta(p^*))$. Note, the algebra for this simplification is given in (2.50)–(2.55) and we do not reproduce it here. \square

2.3 Implementation

Recall that the classical control variate estimator for μ_X was given as follows:

$$Q(\alpha^*) = X + \alpha^*(Y - \mu_Y). \quad (2.56)$$

where α^* is an unknown scalar parameter that must be estimated from the simulation. That is, if n replications of (X, Y) are generated resulting in (X_i, Y_i) , $i = 1, \dots, n$, then we can estimate α^* by

$$\hat{\alpha}^* = -\frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}, \quad (2.57)$$

where \bar{X}_n, \bar{Y}_n are the sample means of X and Y , respectively. In QCV analysis

one must estimate two unknown parameters, α^* and p^* . Here, an implementation of the QCV estimation procedure is more complex. The simulation must “simultaneously” estimate α^* and determine how long to estimate the control variate mean, by estimating p^* . [21] develops a dynamic QCV procedure that continually updates estimates of r and ρ_{XY}^2 and adaptively changes the amount of CPU time that is devoted to estimating the QCV mean and performing the main simulation. We now describe this procedure in detail.

2.4 An Optimal Dynamic QCV procedure

In order to estimate μ_X an experimenter has available the three simulations A, B and C as described in Section 2.2. An optimal experimental design is developed using these simulations where the criterion for optimality is to minimize the AAV. The optimal QCV procedure basically runs all three simulations continually collecting and updating statistics in order to direct the overall simulation by dynamically changing the amount of time spent on each of the simulations. We divide the simulation into n time segments and allocate different portions to simulations A, B and C. Without loss of generality we assume that the time segment is unit length and the n th time step occurs at $t = n$, $n = 0, 1, 2, \dots$ (We assume that one unit of time is enough to allow for any overhead associated with our procedure.) After the $(n - 1)$ st time

segment we update estimators of r and ρ_{XY} based on all simulation data up to that point in order to revise our scheme for the n th segment. The strategy for the n th segment is represented by the fractions of time spent on simulations A, B and C denoted by $(\tau_n^A, \tau_n^B, \tau_n^C)$. Define a_t, b_t and c_t to be the CPU times devoted to simulations A, B and C up to time t and $\hat{\tau}_x^A \hat{\sigma}_X(a_t)$, $\hat{\tau}_{xy} \hat{\sigma}_X(b_t)$, $\hat{\tau}_{xy} \hat{\sigma}_Y(b_t)$, $\hat{\tau}_{xy} \hat{\sigma}_{XY}(b_t)$, $\hat{\tau}_z \hat{\sigma}_Z(c_t)$, to be estimates of the unknown $\tau_x^A \sigma_X$, $\tau_{xy} \sigma_X$, $\tau_{xy} \sigma_Y$, $\tau_{xy} \sigma_{XY}$, $\tau_z \sigma_Z$, so that we can define

$$\hat{r}_t = \frac{\sqrt{\hat{\tau}_z \hat{\sigma}_Z(c_t)}}{\sqrt{\hat{\tau}_{xy} \hat{\sigma}_Y(b_t)}} \quad \text{and} \quad \hat{\rho}_{XY}(t) = \frac{\hat{\sigma}_{XY}(b_t)}{\hat{\sigma}_X(b_t) \hat{\sigma}_Y(b_t)}, \quad (2.58)$$

as the estimates of r and ρ_{XY} , respectively, based on all simulation data up to time t . Now, based on Theorem 2.14, define:

$$p_t = \begin{cases} \frac{\sqrt{\hat{r}_t^2 \hat{\rho}_{XY}^2(t)(1-\hat{r}_t^2)(1-\hat{\rho}_{XY}^2(t)) - \hat{r}_t^2(1-\hat{r}_t^2)}}{(1-\hat{r}_t^2)(1-\hat{r}_t^2 - \hat{\rho}_{XY}^2(t))}, & \text{if } \hat{r}_t^2 + \hat{\rho}_{XY}^2(t) \neq 1 \text{ and } \hat{\rho}_{XY}^2(t) > \hat{r}_t^2 \\ 1 - \frac{1}{2\hat{\rho}_{XY}^2(t)}, & \text{if } \hat{r}_t^2 + \hat{\rho}_{XY}^2(t) = 1 \text{ and } \hat{\rho}_{XY}^2(t) > \hat{r}_t^2 \\ 0 & \text{if } \hat{\rho}_{XY}^2(t) \leq \hat{r}_t^2, \end{cases} \quad (2.59)$$

and $\alpha_t = -\frac{\hat{\sigma}_{XY}(b_t)}{\hat{\sigma}_Y^2(b_t)} \tilde{\alpha}$, where

$$\tilde{\alpha} = \begin{cases} \frac{\sqrt{\hat{\rho}_{XY}^2(t)(1-\hat{r}_t^2)(1-\hat{\rho}_{XY}^2(t)) + \hat{r}_t(\hat{r}_t^2 - 1)}}{(1-\hat{r}_t^2)[\sqrt{\hat{\rho}_{XY}^2(t)(1-\hat{r}_t^2)(1-\hat{\rho}_{XY}^2(t)) - \hat{r}_t \hat{\rho}_{XY}^2(t)}]}, & \text{if } \hat{r}_t^2 + \hat{\rho}_{XY}^2(t) \neq 1 \text{ and } \hat{\rho}_{XY}^2(t) > \hat{r}_t^2 \\ \frac{2\hat{\rho}_{XY}^2(t) - 1}{\hat{\rho}_{XY}^2(t)}, & \text{if } \hat{r}_t^2 + \hat{\rho}_{XY}^2(t) = 1 \text{ and } \hat{\rho}_{XY}^2(t) > \hat{r}_t^2 \\ 0 & \text{if } \hat{\rho}_{XY}^2(t) \leq \hat{r}_t^2, \end{cases} \quad (2.60)$$

to be the estimates of p^* and α^* up to time t ; let $q_t \equiv 1 - p_t$ and construct the estimate of $v^2(p^*, \alpha^*)$ to be

$$v_t^2 = \frac{\hat{\tau}_{xy} \hat{\sigma}_X^2}{(b_t + c_t - \kappa_B - \kappa_C)(1 - p_t)} \left(1 - \frac{p_t}{p_t + q_t \hat{r}_t^2} \hat{\rho}_{XY}^2(t) \right). \quad (2.61)$$

Also, at time t we can estimate the AAV of \bar{X}_t from simulation A as

$$u_t^2 = \frac{\hat{\tau}_X^A \hat{\sigma}_X^2(a_t)}{a_t - \kappa_A}. \quad (2.62)$$

Let

$$\Delta_t = u_t^2 - v_t^2.$$

If a_t, b_t and c_t grow without bound, estimates of all the parameters will converge almost surely to their exact values, which are needed to decide whether \bar{X}_t or $\bar{Q}_t(p^*, \alpha^*)$ is more efficient. On the other hand, we want to spend an asymptotically negligible fraction of time running a suboptimal experimental design.

We now describe the recommended QCV procedure. Let $\delta_i \in (0, \frac{1}{3})$ $i = 1, 2, \dots$ be a sequence satisfying the following conditions:

$$\delta_i \rightarrow 0,$$

$$\sum_{i=1}^{\infty} \delta_i = \infty$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \rightarrow 0.$$

Now, assign $\tau_0^A > 0$, $\tau_0^B > 0$ and $\tau_0^C > 0$ time units to simulations A, B and C in the zeroth segment where $\tau_0^A + \tau_0^B + \tau_0^C = 1$. For $n \geq 1$, the simulations are assigned

$$\tau_n^B = \begin{cases} (1 - 2\delta_n)\max(q_n, \delta_n), & \text{if } \Delta_n > 0 \\ \delta_n & \text{if } \Delta_n < 0, \end{cases} \quad (2.63)$$

$$\tau_n^C = \begin{cases} (1 - 2\delta_n)\max(p_n, \delta_n), & \text{if } \Delta_n > 0 \\ \delta_n & \text{if } \Delta_n < 0. \end{cases} \quad (2.64)$$

and

$$\tau_n^A = 1 - \tau_n^B - \tau_n^C \quad (2.65)$$

time units in the n th segment. At time t we can estimate μ_X by the following weighted average of \bar{X}_{a_t} (from simulation A) and $\tilde{X}_{b_t} + \alpha_t(\tilde{Y}_{b_t} - \bar{Z}_{c_t})$ (from simulations B and C). For $t > 0$ let

$$Q'_t = \frac{a_t}{t} \bar{X}_{a_t} + \frac{b_t + c_t}{t} \left(\tilde{X}_{b_t} + \alpha_t(\tilde{Y}_{b_t} - \bar{Z}_{c_t}) \right), \quad (2.66)$$

and let $v_{Q'_t}^2$ be the AAV of Q'_t .

In summary, pseudo code for the QCV procedure for $t > 0$ units of simulation time is given as follows:

Initialize $\tau_0^A > 0$, $\tau_0^B > 0$, $\tau_0^C > 0$ such that $\tau_0^A + \tau_0^B + \tau_0^C = 1$; let $n = 0$.

begin

while ($n < t$) **do**

[1] Run simulations A, B and C for τ_n^A , τ_n^B and τ_n^C units of time resp.

[2] Obtain estimates of p_n and α_n using (2.58), (2.59) and (2.60).

[3] Find $\Delta_n = u_n^2 - v_n^2$ using (2.61) and (2.62).

[4] Let $n \leftarrow n + 1$.

[5] Using (2.63), (2.64) and (2.63), assign values for τ_n^A , τ_n^B and τ_n^C .

end while

output:

$$Q'_t = \frac{a_t}{t} \bar{X}_{a_t} + \frac{b_t + c_t}{t} \left(\tilde{X}_{b_t} + \alpha_t (\tilde{Y}_{b_t} - \bar{Z}_{c_t}) \right)$$

and

$$v_{Q'_t}^2 = \frac{\hat{\tau}_{xy} \hat{\sigma}_X^2}{(b_t + c_t - \kappa_B - \kappa_C)(1 - p_t)} \left(1 - \frac{p_t}{p_t + q_t \hat{r}_t^2} \hat{\rho}_{XY}^2(t) \right).$$

end

We now offer justification why the recommended QCV procedure is a reasonable algorithm. Note, for the following results, we reference [21]. So far, all of our analyses regarding simulation efficiency allowed for overhead. To warrant the previous algorithm, we need to understand how the algorithm behaves as $t \rightarrow \infty$. Thus, asymptotically, overhead is negligible and we consider the asymptotic variance parameters (see Remark 2.9) of each of the estimators

\bar{X}_t and \bar{Q}_t , namely

$$u^2 = \tau_X^A \sigma_X^2$$

and

$$v^2 = \frac{\tau_{xy} \sigma_X^2}{(1 - p^*)} \left(1 - \frac{p^*}{p^* + q^* r^2} \rho_{XY}^2 \right),$$

which are estimated by

$$u_t^2 = \hat{\tau}_X^A \hat{\sigma}_X^2(a_t)$$

and

$$v_t^2 = \frac{\hat{\tau}_{xy} \hat{\sigma}_X^2}{(1 - p_t)} \left(1 - \frac{p_t}{p_t + q_t \hat{r}_t^2} \hat{\rho}_{XY}^2(t) \right),$$

respectively.

Simply put, in the QCV algorithm, replace the AAV of each estimator with its asymptotic variance parameter (AVP). (Notice that the asymptotic variance parameters are independent of t .) Having established that, we can now justify our algorithm. First, a_t, b_t and c_t each grow without bound as $t \rightarrow \infty$. Thus, $\hat{r}_t \rightarrow r$, $\hat{\rho}_{XY}(t) \rightarrow \rho_{XY}$, $\hat{\tau}_X^A \hat{\sigma}_X^2(a_t) \rightarrow \tau_X^A \sigma_X^2$, $\bar{X}_t \rightarrow \mu_X$, $(\tilde{X}_t, \tilde{Y}_t) \rightarrow (\mu_X, \mu_Y)$, $\bar{Z}_t \rightarrow \mu_Y$, $p_t \rightarrow p^*$ and $\alpha_t \rightarrow \alpha$ and $v_t^2 \rightarrow v^2$. Thus, all estimators converge to their exact values. Now, define

$$\Delta = u^2 - v^2$$

to be the true difference between the corresponding asymptotic variance parameters. An optimal estimator has AVP

$$v_*^2 = \min(u^2, v^2),$$

i.e., use the QCV procedure if it is superior to simulation A; otherwise use simulation A.

If $\Delta > 0$, then the AVP of our QCV estimator is less than that of \bar{X}_t from simulation A. In this case it can be shown that $\frac{c_t}{t} \rightarrow p^*$, $\frac{b_t}{t} \rightarrow (1-p^*)$ and $\frac{a_t}{t} \rightarrow 0$ *a.s.* as $t \rightarrow \infty$. Thus, the QCV procedure spends asymptotically the “right” amount of time doing the correct simulations and an asymptotically negligible amount of time running a suboptimal one. Likewise, if $\Delta < 0$, then the AVP of our QCV estimator is more than that of \bar{X}_t from simulation A. In this case it can be shown that $\frac{b_t+c_t}{t} \rightarrow 0$, and $\frac{a_t}{t} \rightarrow 1$ *a.s.* as $t \rightarrow \infty$. If $\Delta = 0$, it is unclear how the QCV procedure will behave and the method is not guaranteed to work. However, when $\Delta \neq 0$ it can be shown that $v_{Q'_t} = v_*^2$.

2.5 Generic Example

We begin this section by describing a “generic” implementation of the QCV procedure. Suppose one is interested in estimating $\mu = E[f(\phi)]$ where $\phi = (\phi^1, \phi^2, \dots, \phi^m)$ is a random vector and $f : \mathfrak{R}^m \rightarrow \mathfrak{R}$ is a well behaved function, e.g., uniformly continuous. The “generic” approach begins by choosing

$\psi_j \in \mathfrak{R}^m$, $j = 1, \dots, M$ and computing and storing $f(\psi_j)$ $j = 1, \dots, M$. Now, let ϕ_1, ϕ_2, \dots be i.i.d. replicates of ϕ and define $\pi_i = \operatorname{argmin}_{j \in \{1, 2, \dots, M\}} \|\phi_i - \psi_j\|$ to be the index of the element of $\Psi = \{\psi_1, \psi_2, \dots, \psi_M\}$ closest to ϕ_i . (In case of a tie select the last index found.)

Simulation A estimates μ by Monte Carlo simulation. Let N_t be the number of replicates of $f(\phi_i)$ the simulation generates up to time t . Then simulation A provides

$$\bar{X}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} f(\phi_i)$$

and

$$\sigma_X^2(t) = \frac{t}{N_t} \left(\frac{1}{N_t} \sum_{i=1}^{N_t} f(\phi_i)^2 - \bar{X}_t^2 \right).$$

Simulation B estimates μ and $\mu_Y = E[f(\psi_{\pi_i})]$ along with the asymptotic covariance matrix by

$$(\tilde{X}_t \tilde{Y}_t) = \frac{1}{M_t} \sum_{i=1}^{M_t} (f(\phi_i), f(\psi_{\pi_i})), \quad (2.67)$$

and

$$\begin{pmatrix} \tilde{\sigma}_X^2(t) & \tilde{\sigma}_{XY}(t) \\ \tilde{\sigma}_{XY}(t) & \tilde{\sigma}_Y^2(t) \end{pmatrix} = \frac{t}{M_t} \left(\frac{1}{M_t} \sum_{i=1}^{M_t} \begin{bmatrix} f(\phi_i) \\ f(\psi_{\pi_i}) \end{bmatrix} [f(\phi_i) f(\psi_{\pi_i})] - \begin{pmatrix} \tilde{X}_t^2 & \tilde{X}_t \tilde{Y}_t \\ \tilde{X}_t \tilde{Y}_t & \tilde{Y}_t^2 \end{pmatrix} \right),$$

where M_t is the number of pairs $(f(\phi_i), f(\psi_{\pi_i}))$ simulation B generates up to time t .

Simulation C estimates μ_Y and its asymptotic variance parameter by

$$\bar{Y}_t = \frac{1}{K_t} \sum_{i=1}^{K_t} f(\psi_{\pi_i})$$

and

$$\sigma_Y^2(t) = \frac{t}{K_t} \left(\frac{1}{K_t} \sum_{i=1}^{K_t} f(\psi_{\pi_i})^2 - \bar{Y}_t^2 \right),$$

where K_t is the number of replicates of $f(\psi_{\pi_i})$ that are generated up to time t . Since $\{f(\psi_j)\}$, $j = 1, 2, \dots, M$ is evaluated before the main simulation starts, there is no need to evaluate $f(\cdot)$ in simulation C. Evaluating π_i involves finding the closest element of Ψ to ϕ_i , which has complexity $O(M)$, and is therefore fast (unless M is too big). Thus, r is small; see Remark 2.15. Also, if f is fairly smooth and Ψ is “representative” of ϕ then $f(\phi_i)$ and $f(\psi_{\pi_i})$ will be highly correlated. (A reasonable way to choose Ψ is to generate M independent replicates of ϕ). This approach is particularly effective if f is expensive or too time consuming to evaluate.

3. Stochastic Linear Programs

We present two “real-world” applications. The first is an illustration of the generic QCV scheme previously described and involves estimating expected profits in oil refining, where the operations, i.e., blending and distilling can be modeled using a linear program. Here, the technology matrix (“A-matrix”) is random. The second application involves power system reliability evaluation and employs a linear program with a random right-hand side. In this situation a special “dual approximation” procedure suggested by [44] can be implemented which be discussed in Section 3.2.

3.1 Petroleum Refinery

Oil companies have been using linear programming to plan refinery operations for more than 30 years. The activity to be planned is the optimal blend and processing management of varying crude oils. Crude oil or just crude, as petroleum directly out of the ground is called, is a remarkably varied substance, both in its use and composition. It can be a straw-colored liquid or tar-black solid. Red, green and brown hues are not uncommon. Crude oils also vary in their physical characteristics, e.g., density, sulphur content, waxiness,

etc., and price. To produce finished products, the crude oil goes through a number of processes such as fractional distillation, which separates out the light components from the thicker heavier oils; cracking, which breaks down the heavy constituents of the crude oil into lighter components; and reforming, which changes the chemical structure of other components in order to meet product specifications. This is just a brief description of the operations of a refinery, and the actual processes are remarkably complex and will be discussed in the next section. The refinery can be operated with varying blends of crude oils, which generally results in sub-optimal profits. However, the aim of the refinery economist is to determine the optimal blend of the various crude oils to process in the distiller, the amount of the resulting heavier fractions to feed to the cracker, etc., in order to find the most profitable way of meeting the expected market for gasoline, diesel, kerosene, propane and other products.

3.1.1 Refinery

Essentially, a refinery is a factory that takes crude oil and transforms it into gasoline, kerosene-type jet fuel and many other useful products. Refining breaks crude oil down into various components, which then are selectively reconfigured into new products. Modern separation involves piping the crude oil through hot furnaces. The resulting liquids and vapors are discharged into

distillation towers. Inside the towers, the liquids and vapors separate into components or fractions according to weight and boiling point. The lightest fractions, including gasoline and liquid petroleum gas (LPG), vaporize and rise to the top of the tower, where they condense back to liquids. Medium weight liquids, including kerosene and diesel oil distillates, stay in the middle. Heavier liquids, called gas oils, separate lower down, while the heaviest fractions with the highest boiling points settle at the bottom.

The refining process consists of the following six distilling towers:

- (1) the topping,
- (2) the re-forming,
- (3) the thermic re-forming,
- (4) the vacuum,
- (5) the catalytic cracking unit and
- (6) the catalytic polymerizator.

Topping The first step in refining crude oil is called fractionation or topping where the crude oil blend is heated to certain temperatures and distilled into fractions of different boiling ranges. The main fractions of the topping are gasoline, propane, butane, Petroleum Fuel Distillates (PFD), benzolene, naphthalene, kerosene, high boiling residue and some non-vaporized fraction which is removed from the bottom of the fractionation tower.

Re-forming Some fractions of the topping, especially benzolene, naphthalene and kerosene can undergo a second distillation in re-forming. Re-forming is a process that transforms heavy benzene fractions with high boiling point and low octane number into lighter benzene fractions with higher octane number.

Thermic Re-forming The thermic re-forming process is similar to re-forming except that temperatures are higher; this results in higher gasoline production than in reforming. The input to the thermic re-forming comes from the PFD, benzolene and naphthalene fractions of the topping.

Vacuum Some of the heated residue from the topping is flushed into the vacuum distillation column and undergoes a distilling with reduced pressures. This residue is further fractionated to produce a light fuel oil fraction and a very heavy black fuel oil fraction.

Catalytic cracking unit A liquid fraction of the vacuum called WD is “cracked” in the catalytic cracking unit. Large chains of molecules of high-boiling hydrocarbons are broken up and changed into smaller gasoline molecules. This process changes the higher-boiling fractions of the petroleum distillation into lower-boiling gasoline.

Catalytic polymerizator The very light fractions of the catalytic cracking unit are treated in the catalytic polymerizator. In this unit one obtains

gasoline of very high quality.

3.1.2 Crude Oil Quality

Variations in the quality of crude oils used by a refinery results in varying production yields, which in turn affect profits. Several parameters influence the degree of crude oil variability. In addition to crude oil location, aging production reservoirs, changes in relative field production rates, gathering system mixing of crude, pipeline degradation and injection of significantly different quality streams into common specification crude streams contribute to crude oil quality variation. “Value” to a refinery is based on the expected output minus the operating costs to be incurred to achieve the desired yield [59]. Ensuring that the quality of crude oil received is equivalent to the purchased quality is one of the greatest challenges facing the industry today. Analytical testing of a received batch of crude oil can be performed whereby a complete physical distillation is done on a sample in the laboratory. The results of this analysis will determine how the crude is represented in the refinery’s linear program model. For instance, suppose a refinery receives several batches of crude oil from various reserves. The refiners do not know how these new crudes will fractionate. That is, they do not have beforehand knowledge of the quantities of the various products that will be produced. For example, the amount of

gasoline that can be produced can vary from 13 to 45 percent of the input and the amount of butane that can be distilled can vary from 0 to 8 percent of the input. Similar variations exist for the other fractions. However, as was mentioned, laboratory tests can be conducted to determine how various crudes will fractionate in the refining process. Once these experiments are completed, the refiners know what data to input into their LP model to determine optimal blend recipes for refining. Hence, we can view this situation as a *wait-and-see* solution to a stochastic program where the randomness occurs in the technology matrix of the LP. That is, optimal decisions can be made after the randomness is resolved.

3.1.3 Linear Programming Model

We now illustrate the use of linear programming in a petroleum refinery. The data is based on a Belgian refinery given in [63]. The structure of the linear programming model is simple. The objective function is profit maximization; it will reflect the crude oil cost, additional manufacturing costs and the market value for finished products. The constraints of the LP model the production flow for the refinery, describing how the inputs can be used for a variety destinations. Three types of crude oil, (x_1, x_2 and x_3 in tons), are used as input into the topping. The following eight fractions are produced in

the topping: gas, propane, butane, Petroleum Fuel Distillates (PFD), benzene and naphthalene, kerosene, residue and loss. These fractions may have as many as 26 destinations having decision variables (x_4 through x_{29}). Table 3.1 gives the percentages of each crude oil that will be transformed into the various fractions in topping; for instance, 6.0 percent of crude oil 1 (x_1), 8.0 percent of crude oil 2 (x_2) and 6.8 percent of crude oil 3 (x_3) are transformed into the PFD fraction. (Later we will treat these values as random variables reflecting the varying crude oil qualities.)

This table also indicates that this PFD fraction can be further processed by any combination of the following:

- (1) (x_{11}) tons transformed into finished product Combustible PFD,
- (2) (x_{12}) tons sent to Thermic re-forming for further processing,
- (3) (x_{13}) tons converted to finished product Gasoline I,
- (4) (x_{14}) tons converted to finished product Gasoline II or
- (5) (x_{15}) tons left as finished product PFD.

The aforementioned is modeled in the LP as the following constraint:

$$0.06x_1 + 0.08x_2 + 0.068x_3 = x_{11} + x_{12} + x_{13} + x_{14} + x_{15} = 0$$

or equivalently

$$0.06x_1 + 0.08x_2 + 0.068x_3 - x_{11} - x_{12} - x_{13} - x_{14} - x_{15} = 0. \quad (3.1)$$

Fraction	Yield of input, %			Destination	Model Variable
	x_1	x_2	x_3		
Gas	0.45	0.27	0.30	Combustible	x_4
Propane	0.25	0.03	0.30	Propane	x_5
				Combustible	x_6
Butane	0.90	0.80	0.90	Butane	x_7
				Gasoline I	x_8
				Gasoline II	x_9
				Combustible	x_{10}
PFD	6.00	8.00	6.80	Combustible PFD	x_{11}
				Thermic re-forming	x_{12}
				Gasoline I	x_{13}
				Gasoline II	x_{14}
				PFD	x_{15}
Bensolene & naphthalene	9.00	11.50	10.20	Thermic re-forming	x_{16}
				JP4	x_{17}
				Re-forming	x_{18}
Kerosene	25.60	32.40	31.50	Re-forming	x_{19}
				JP4	x_{20}
				Kerosene	x_{21}
				Gas oil	x_{22}
Residue	57.30			Vacuum	x_{23}
				Fuel	x_{24}
		46.50		Vacuum	x_{25}
				Fuel	x_{26}
		49.50		Vacuum	x_{27}
				Fuel	x_{28}
Loss	0.50	0.50	0.50	Loss	x_{29}

Table 3.1. Topping

For the quantity (x_{12}) of PFD that is potentially piped to the Thermic re-forming unit as input (see item (2) in the list above), Table 3.2 shows it will be further transformed into gas, gasoline, residue and loss.

Fraction	Yield of input, % (x_{12}, x_{16})	Destination	Model Variable
Gas	29.00	Combustible	x_{38}
Gasoline	65.00	Gasoline I	x_{39}
		Gasoline II	x_{40}
Residue	5.00	Fuel	x_{41}
Loss	1.00	Loss	x_{42}

Table 3.2. Thermic Re-forming

Thus, we see that the initial fractions have innumerable destinations and Tables A.1– A.4, given in the appendix, present analogous data for the remaining distillation units. Since varying production plans result in variable product yields, the refinery wishes to find the one that will be most beneficial. The refinery will, of course, choose the one that maximizes profit. Here, profit is the amount of revenue that can be generated through the market value of the products less the raw material and production costs. So, the objective function will reflect the crude oil cost and the revenue obtained from selling the following finished products: propane, butane, gas oil, gasoline, JP4 (jet fuel), pitch and kerosene. Table 3.3 gives the cost per ton for each of the various crudes. The quality of the crude is reflected in its price. Also, Table 3.4 lists the market price for each of the products.

Crude Oil	Cost/ton	Variable
1	\$118	x_1
2	\$150	x_2
3	\$170	x_3

Table 3.3. Raw Material Costs

Product	Price/ton	Variables
Propane	\$275	x_5, x_{31}, x_{58}
Butane	\$285	x_7, x_{33}, x_{60}
Gas Oil	\$420	$x_{22}, x_{44}, x_{53}, x_{55}$
Gasoline I	\$580	$x_8, x_{13}, x_{35}, x_{39}, x_{51}, x_{62}$
Gasoline II	\$538	$x_9, x_{14}, x_{36}, x_{40}, x_{52}, x_{63}$
JP4	\$260	x_{17}, x_{20}
Pitch	\$207	x_{45}
Kerosene	\$277	x_{21}

Table 3.4. Product Prices

Thus the objective is to maximize the following function:

$$\begin{aligned}
 f = & -118x_1 - 150x_2 - 170x_3 + 275(x_5 + x_{31} + x_{58}) \\
 & + 285(x_7 + x_{33} + x_{60}) + 420(x_{22} + x_{44} + x_{53} + x_{55}) \\
 & + 580(x_8 + x_{13} + x_{35} + x_{39} + x_{51} + x_{62}) \\
 & + 538(x_9 + x_{14} + x_{36} + x_{40} + x_{52} + x_{63}) + 260(x_{17} + x_{20}) \\
 & + 207x_{45} + 277x_{21}.
 \end{aligned}$$

Other constraints that are included in the model involve distilling capacities. The topping unit can hold at most 5000 tons of crude and thus we include the following constraint in our model:

$$x_1 + x_2 + x_3 \leq 5000. \tag{3.2}$$

The other capacity constraints included in the model are described in Table 3.5

Column	Input variables	Capacity (tons)
Topping	x_1, x_2, x_3	5000
Re-forming	x_{18}, x_{19}	500
Thermic Re-forming	x_{12}, x_{16}	250
Vacuum	x_{23}, x_{25}, x_{27}	None
Catalytic cracking	x_{43}	400
Catalytic polymerizator	x_{50}	None

Table 3.5. Distilling Capacities

This concludes the description of the linear program model for the

petroleum refinery problem. Note that the fractionation Tables 3.1– 3.2 and Tables A.1– A.4 present the fractionation percentages as known constants. As was stated before, we wish to incorporate uncertainty in the LP model. Thus, we will consider the data presented in the tables as the mean fractionation percentages and allow for a 20% uniform deviation. This uncertainty distribution was our choice and although this deviation may not be realistic, it suited our purposes to demonstrate the QCV procedure. For instance, in Table 3.1 note that 25.6% of crude oil 1 will fractionate into kerosene. In the LP model this will be represented as a uniform random variable that ranges from 20.48 to 30.72. All other data in the table will be treated similarly.

3.1.4 Numerical Experiments

We now demonstrate the “generic” version of the QCV procedure described in Section 2.5. Our goal is to estimate the expected value of the objective function of an LP whose technology or constraint matrix is random. The linear programming and simulation code was written in C using the compiler Microsoft Visual C++ 6.0. The first operation of the estimation procedure is to construct and store the representative set $\Psi \equiv \{\psi_1, \psi_2, \dots, \psi_M\}$ (see section 2.5) which in our setting is merely M independent replicates of the random technology matrix. For each of these M matrices we compute and store the

corresponding LP objective function value. The time spent on this construction is the overhead and increases as a function of the number of points. The function f is the objective function of the LP.

The experiment consisted of allotting $t = 1$ minute time unit for the entire simulation including the overhead and the QCV procedure. Each time segment was five seconds long (see Section 2.4). We are interested in determining how various parameters of the simulation behave as a function of the number of points in Ψ . For instance, one would expect that the correlation ρ would increase as a function of the number, M , of points in Ψ . Also, it would seem reasonable that the ratio r should also increase. Thus, given a finite simulation time of one minute, an item of interest is how many points to include in Ψ . If too few points are used, one would not expect a large correlation and the QCV will not be effective. If many points are involved, too much time will be spent in the overhead and very little time will be spent performing the main simulation. Hence, one would expect an “optimal” choice for the size of Ψ .

Figure 3.1 illustrates how the various parameters behave as a function of M . Note that since the placement of the points in Ψ is random, the resulting parameters, i.e., r , ρ and the AAV will also be random. Thus, the points on the graphs are averages taken over 20 replications of the experiment. As expected,

we see that the time it takes to build the approximation is a linear function of M . The correlation function is the most encouraging. Notice that it rises quickly in a concave downward fashion reaching 0.90 with 100 points in the approximation. The graph then tends to flatten out. Thus, we get more “bang for the buck” early in the construction achieving a high correlation with only a few points in the approximation. We also have displayed the ratio r and the optimal proportion p^* as a function of M . We note that the more points we use in our approximation, the more time is spent on simulation C in estimating the control variate mean. The question then becomes the following: “How many points should we include in the approximation?”. Figure 3.2 plots the AAV variance as a function of M . We see that if we have too few points, e.g., less than 50, our variance will be large. The graph reaches a minimum at about 100 points and then begins to increase with the addition of more points. This is because the marginal gain in higher correlation is not worth the marginal effort of adding more points; also, with the addition of more points the time spent on constructing the approximation becomes a significant part of the allotted simulation time and little is left for the main experiment.

3.2 Stochastic LP with random RHS

As was illustrated in the previous section, the estimation of the mean of the objective function of an LP whose constraint matrix is random is a great example of the “generic” application of the QCV procedure. We now illustrate the QCV scheme in a situation with a stochastic linear program with random right hand sides. In this case there is a special utilization of the QCV proposed by Oliveira, Pereira, Pinto and Cunha [44].

Let A be an $m \times n$ matrix and let $c \in \Re^n$. For $b \in \Re^m$, let

$$P(b) = \min_{x \in \Re^n} \{ cx \mid Ax \geq b, x \geq 0 \}. \quad (3.3)$$

We will refer to this problem as the primal problem. One might be interested in estimating $\mu = E[P(\phi)]$, where $\phi = (\phi^1, \phi^2, \dots, \phi^m)'$ is a random (right hand side) vector. Of course one approach would be to generate N i.i.d. replications $\{\phi_1, \phi_2, \dots, \phi_N\}$ of ϕ and solve N linear programming problems to obtain $P(\phi_i)$, $i = 1, \dots, N$. Our estimate of $E[P(\phi)]$ would be

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N P(\phi_i), \quad (3.4)$$

which is obtained in simulation A.

If the linear program is large, one may spend much time obtaining each solution. Consider the dual problem to (3.3)

$$D(b) = \max_{\pi \in \Re^m} \{ b\pi \mid \pi A \leq c, \pi \geq 0 \}, \quad (3.5)$$

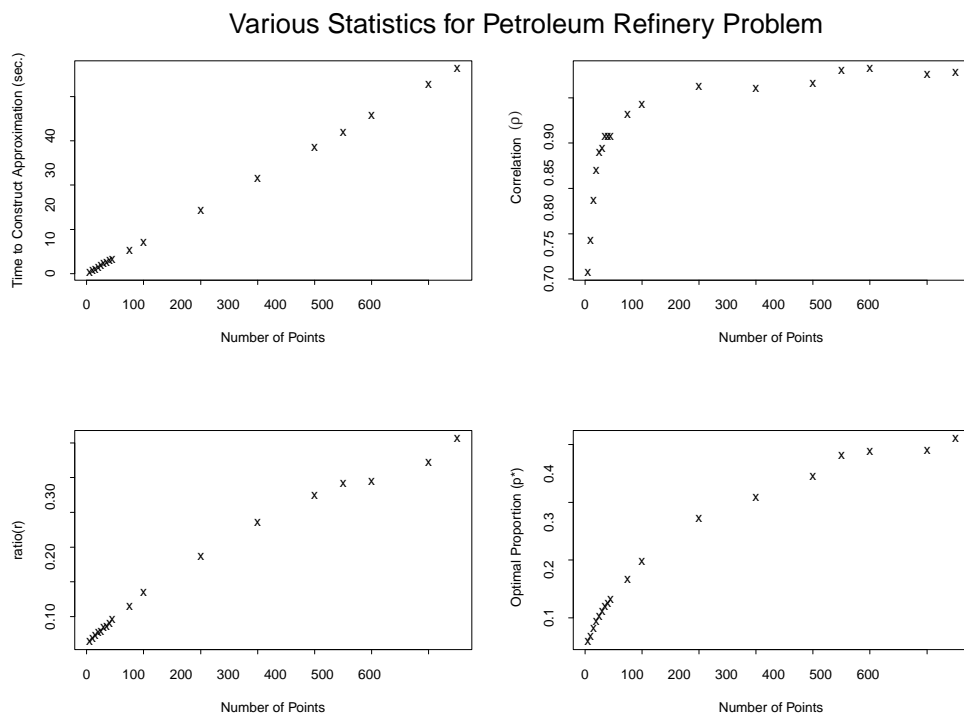


Figure 3.1. Statistics as a function of Size of Analytical Approximation

The AAV as a function of Number of Points

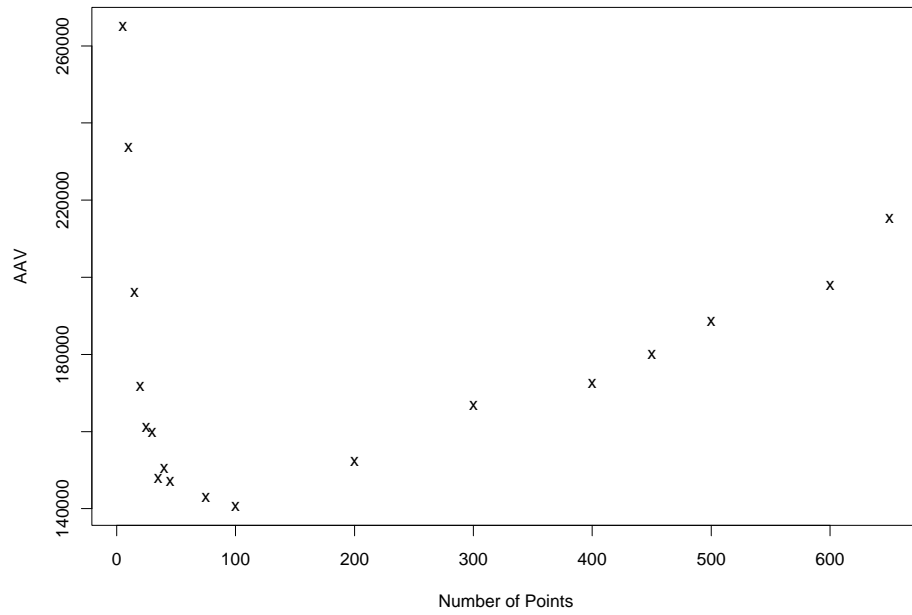


Figure 3.2. AAV as Function of Number of Points in Approximation

and note that the feasible region is independent of b .

Our goal is to construct a control variate that is highly correlated with $P(\phi)$, but can be obtained much quicker than solving the primal problem. Thus, we need to find a problem that retains the primal problem's characteristics yet is much simpler to solve. Our simpler model will be a dual approximation to the primal problem. Consider the polyhedron $H = \{\pi | \pi A \leq c, \pi \geq 0\}$, which we consider to be bounded and nonempty to facilitate the exposition, and its extreme points $Ext(H) = \{\pi_1, \pi_2, \dots, \pi_n\}$. It follows from the Fundamental Theorem of Linear Programming and the equivalence of extreme points and basic feasible solutions [41] that

$$P(b) = \max_i \{b\pi_i | \pi_i \in Ext(H)\}. \quad (3.6)$$

This suggests how we should construct our simpler model. Instead of enumerating all the extreme points of H , construct a potentially small subset $H' \subset Ext(H)$ and let

$$D^*(b) = \max_i \{b\pi_i | \pi_i \in H'\}, \quad (3.7)$$

be the dual approximation or control variate for the primal problem. Thus, provided that $|H'|$ is not prohibitively large, obtaining a solution to (3.7) is much faster than solving the original primal problem for a given replication of

ϕ . Thus, our QCV equation for this application becomes

$$P^*(\phi) = P(\phi) + \alpha^*(D^*(\phi) - \bar{D}^*). \quad (3.8)$$

where $(P(\phi), D^*(\phi))$ are the random pair generated in simulation B and \bar{D}^* is the estimate of $E[D^*(\phi)]$ computed in simulation C.

3.2.1 Construction of the Dual Approximation

Two methods are considered for constructing a dual approximation to problem (3.6), that is, to generating the set H' . One approach is to enumerate the desired number of vertices of H using a deterministic algorithm and the other is to randomly generate them from the density of ϕ . It is not readily clear which approach would be better suited for our needs. Deterministic vertex enumeration would guarantee obtaining a new vertex at each iteration whereas a random selection approach might spend valuable simulation time producing the same vertex. However, since we will be generally dealing with very large problems, the chance of generating a previously obtained vertex would be very small. Also, generating the vertices randomly may “probabilistically” produce a better dual approximation since it would generate those that would most likely appear in the actual Monte Carlo simulation. In contrast, the deterministic algorithms simply pivot from one vertex to another and may be able to find an abundant number of vertices quickly. In random generation,

a linear program would have to be solved for each vertex. Finally, the purpose of deterministic algorithms is to find *all* the vertices and extreme rays. We merely want to generate a subset of them. As will be seen in Section 3.2.4, a surprisingly small subset of dual vectors is required to induce a high correlation between the primal and dual approximation. Numerical experiments clearly indicated that the random generation of the vertices is superior.

3.2.2 Deterministic vertex enumeration

Much literature and research are devoted to polyhedral vertex enumeration; it has applications in robotics, quantum chemistry and multicommodity flows. Fukuda and Avis [4] develop a clever pivoting algorithm for vertex enumeration, which is based on “inverting” finite pivoting algorithms for linear programming. We coded and employed an adaptation of their pivoting algorithm for our purpose.

We first give motivation for and then a verbal description of the vertex enumeration algorithm. Suppose we wish to enumerate all vertices of the following non-empty, bounded, non-degenerate, n dimensional polyhedron

$$H = \{x | Ax = b \ x \geq 0\}, \tag{3.9}$$

where A is a full rank, $m \times n$ matrix and $b \in \Re^m$.

We first give a definition.

Definition 3.1 Given the set (3.9), let B be a nonsingular $m \times m$ submatrix made up of columns of A , such that $B^{-1}b \geq 0$. Then, if all $n - m$ components of x not associated with columns of B are set equal to zero, the solution to the resulting set of equations is said to be a *basic feasible solution* to (3.9) with respect to the basis B . The components of x associated with columns of B are called *basic variables*; the remaining components are called *non-basic variables* whose associated columns are represented in the submatrix N of A .

Assume that the following linear program has a unique optimal solution x^* :

$$\min_{x \in \mathbb{R}^n} \{ cx \mid x \in H \}. \quad (3.10)$$

The motivation and philosophy of their algorithm is based on the simplex method of linear programming. Starting from some initial vertex, the simplex method essentially traverses a sequence of *adjacent* vertices of H until it reaches x^* . We arrive at each adjacent vertex by performing a *pivot* whereby a non-basic variable is introduced into the current basis and a basic variable leaves. The path chosen from the initial vertex depends on the pivot rule which must be chosen to avoid cycling. Anti-cycling pivot rules guarantee that we will reach the optimum in a finite number of steps. We use a particularly simple rule, known as Bland's rule [9], which guarantees a unique path from

any starting vertex to the optimum vertex. Conceptually, we could implement this simplex procedure from *every* vertex of H to arrive at our optimum x^* . Now, if we look at the set of all paths from all the vertices of H , we obtain a spanning forest of the graph of adjacent vertices of H . The root of each subtree of the forest is the optimum vertex. However, if the polyhedron is non-degenerate (as we assume), each vertex lies on exactly n hyperplanes. In this case, the spanning forest has one component, which is a spanning tree of the “skeleton” of the polyhedron, and each vertex is produced once. Before we describe how the algorithm works, we give another definition.

Definition 3.2 Let B be a basis for (3.10). Given a column v in B and a non-basic column u in N , a *valid reverse pivot* is one that if we were to pivot u into the basis to obtain the new feasible basis $B - v + u$ and then apply Bland’s rule to the updated tableau we would pivot back to the original basis B .

The vertex enumeration algorithm proceeds by first finding the optimum vertex x^* . We now have a basic feasible solution and basis B for this vertex. The vertex enumeration algorithm now proceeds by fixing the first column of B and attempts to perform a valid reverse pivot for *each* nonbasic column with this fixed column of B . If it finds no valid reverse pivot, the

algorithm then holds the second column of B fixed and again attempts to find a valid reverse pivot with each of the non–basic columns, and so forth. If in the course of the aforementioned iteration it finds a valid reverse pivot, the algorithm pivots to that vertex, saves it, updates the basis to B' and starts the entire procedure over by holding the first column of this basis B' fixed and attempting to find a valid reverse pivot with the new set of non–basic columns. Now, if the algorithm does not find a valid reverse pivot for some given basis, it applies Bland’s rule and performs a simplex pivot moving back up the tree basically “revisiting” a basis. Here, the algorithm continues where it left off (when it last encountered this vertex before descending into the tree) by scanning the remaining non–basic columns to find valid reverse pivots. So, the algorithm proceeds to trace out the entire tree in depth first order by “reversing” Bland’s rule.

Avis and Fukuda [4] analyze the time-complexity of the previously described algorithm. We reproduce the arguments here. Note, (3.9) can have at most $\binom{n}{m}$ basic feasible solutions. For each basis, we may evaluate $m(n - m)$ candidates for valid reverse pivots, each candidate requiring $O(m + n)$ time. The pivot requires $O(m(n - m))$ time per execution. Therefore, the overall time–complexity of the enumeration procedure is

$$O\left((m+n)mn\binom{n}{m}\right).$$

3.2.3 Random vertex generation

The random generation of the dual vertices of H is quite simple. One merely generates independent replicates $\{\phi_1, \phi_2, \dots\}$ of ϕ and solves $D(\phi_i)$ to obtain a vertex $y_i, i = 1, \dots$ and if this vertex is not already on the current list it is added.

3.2.4 Random vs. deterministic generation

Numerical experiments clearly suggest (at least for the test problems and enumeration algorithms we used) that for our purposes a random generation approach is superior to a deterministic approach. We performed various numerical experiments to determine the number of dual vertices to be included in an approximation.

We considered two generic polytopes, one that would have “equally” probable vertices and one with non-equal probable vertices. We chose these artificially constructed problems for a couple of reasons. Firstly, several real world problems that were initially tested were demonstrated to have high correlation with only a few dual vertices and did not “challenge” the QCV scheme.

Secondly, we wanted control and full knowledge of the polytopes so that accurate statements could be made about the numerical results. For example, in both of the test problems below, we know that each has 2^n vertices in the dual problem.

Test Problem 1. (“Equally” Probable Vertices):

$$\min_{x \in \mathfrak{R}^n} \{ ex \mid x \geq \xi, x \geq 0 \}. \quad (3.11)$$

where $e = [1, \dots, 1]$ is the unit vector of length n and $\xi_i \sim U[-99, 101]$, $i = 1, \dots, n$.

For the “equally” probable case we would expect for the correlations that resulted from enumerating the vertices or generating them randomly to be comparable. Figure (3.3) does suggest this since the corresponding correlations are similar. However, we do notice a consistently larger correlation for those that were randomly generated. What is important however is how *long* it takes to build the dual approximations. Notice that Figure 3.4 gives the time required to build a dual approximation where the x -axis represents the number of dual vertices in the approximation. The graph which is based on a log-linear scale clearly indicates that in order to achieve comparable correlation that the random enumeration procedure yields, one would have to spend a substantially longer amount of time with the vertex enumeration procedure.

Correlation vs. Size of Dual Approximation for Equally Probable Case

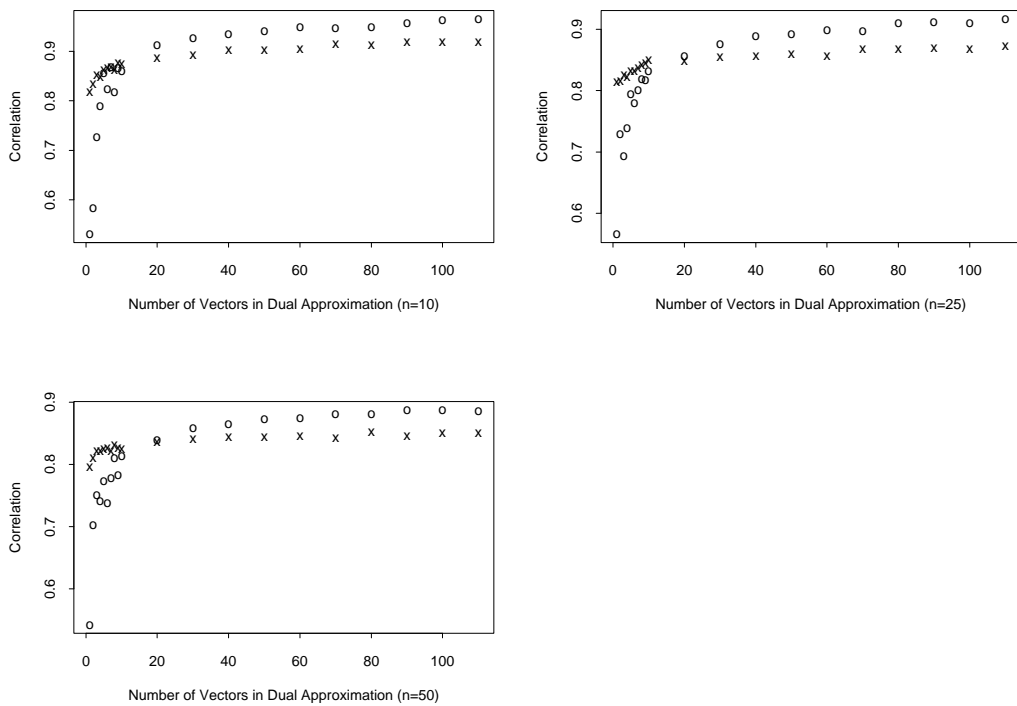


Figure 3.3. Correlations for Equally Probable Case: o-Randomly Generated, x-Enumerated.

Construction Time vs. Size on a Log-Linear Scale for Equal Probable Case

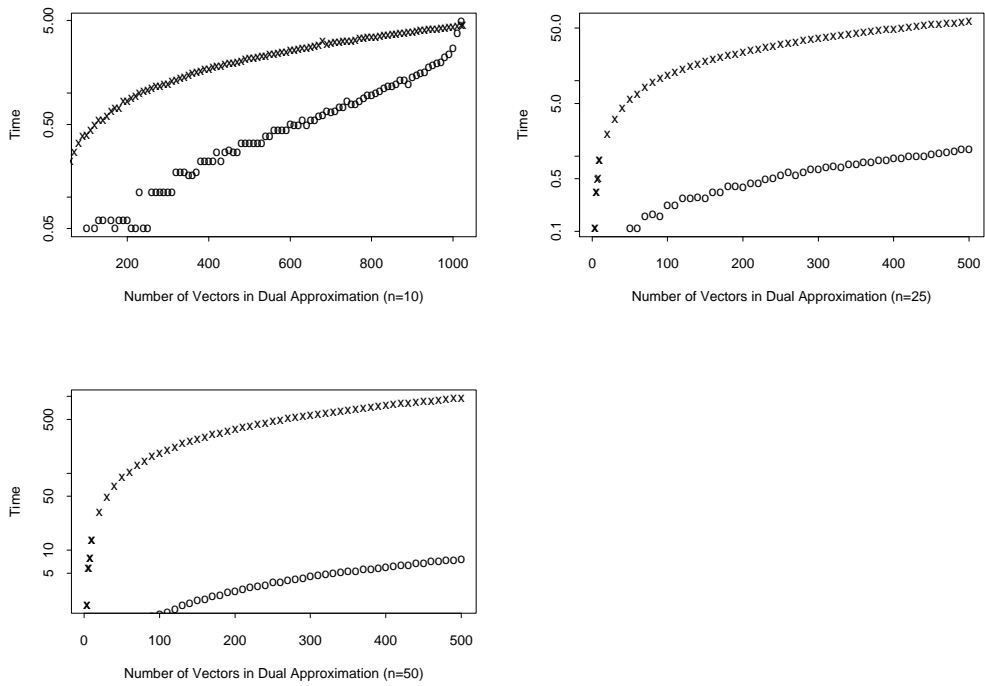


Figure 3.4. Log-linear plot of Time to construct Dual Approximation for Equally Probable Case.

For clarity, we wish to emphasize that $n = 10, 25$ or 50 corresponds to the linear program having 2^n extreme points, basic feasible solutions or vertices in the dual problem.

Test Problem 2. (Non-Equally Probable Vertices):

$$\begin{aligned}
 \min \quad & nx_1 + (n-1)x_2 + \dots + x_n \\
 \text{st} \quad & x_1 \geq n\xi_1 + n \\
 & x_1 + x_2 \geq (n-1)\xi_2 + (n-1) \\
 & \vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \\
 & x_1 + x_2 + \dots + x_n \geq \xi_n + 1 \\
 & x_i \geq 0, \quad i = 1, \dots, n
 \end{aligned}$$

where $\xi_i \sim U[-1/2, 1/2]$.

The non-equally probable test problem is more interesting. In this linear program, only a small subset of the dual vertices will actually be generated in the Monte Carlo simulation. Thus, one would expect that vertex enumeration would be inferior to the Monte Carlo approach. That is, enumeration might produce unwanted dual vertices that the Monte Carlo simulation would never generate, thus not aiding in the effort of increasing correlation. Notice that Figure 3.5 suggests this since the correlations that were generated by enumeration were below 0.3 as they did not appear on the graph for $n = 25$ and 50 . Again, Figure (3.6), as in the “equally” probable case, indicates that the amount of time required to build the approximation

Correlation vs. Size of Dual Approximation for Non-Equal Probable Case

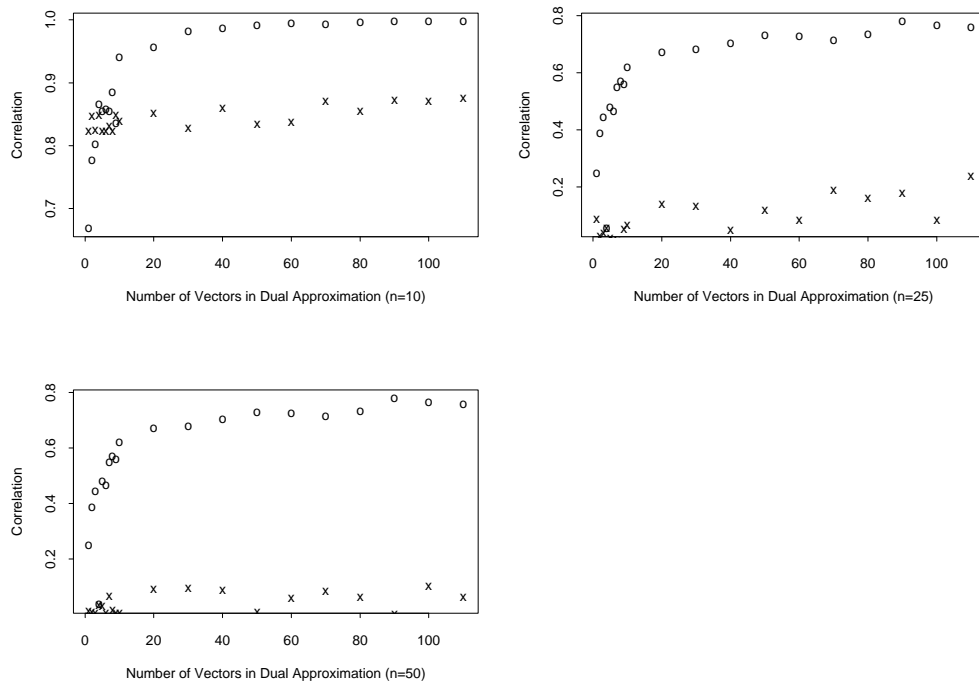


Figure 3.5. Correlations for Non-Equaly Probable Case: o-Randomly Gen-
erated, x-Enumerated

Construction Time vs. Size on a Log-Linear Scale for Non-Equal Probable Case

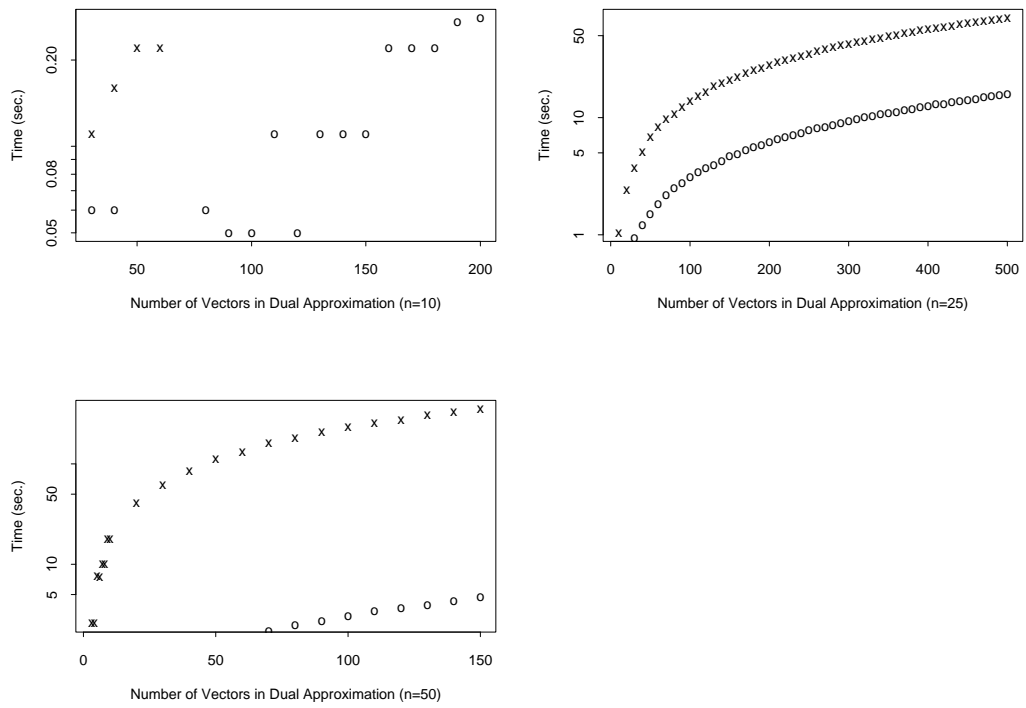


Figure 3.6. Log-linear plot of Time to construct Dual Approximation for Non-Equally Probable Case

is much larger than for random generation. These experiments indicate that random generation is superior, at least for the cases investigated, to enumeration for the construction of the dual approximation.

3.3 Power System Reliability Evaluation

The primary function of an electric power system is to provide electricity to its customers as economically as possible with minimal disruption. Due to random component failures of the system that may be outside the control of the power system personnel, the continuous supply of electrical energy may not be available on demand. The supply of electricity generally involves a complex and highly integrated system; failures in any part can cause interruptions which range from inconveniencing a few local residents to widespread outages. The need for probabilistic evaluation of power system behavior has been recognized at least in the last forty years and has now evolved to the point at which most utilities use these techniques in one or more areas of their planning, design and operation. Many of the methods used are based on analytical models and evaluation procedures [6]. However, the continual advancement of high-speed computing power has created the opportunity to analyze more complex and intricate models using stochastic simulation methods, and in the last decade there has been increased interest and use of Monte Carlo simulation in quantitative power system reliability applications.

Power system reliability evaluation has been extensively developed using

probabilistic methods where a wide range of appropriate indices have been determined. There are many possible indices that can be used to measure adequacy or reliability of a power system [7, section 2.6]. The most popular reliability indices are the following: *loss of load probability* (LOLP) and *expected power not supplied* (EPNS). Many power system planning applications can be modeled using linear programming problems with stochastic right-hand sides [7, 44, 52]. Here the linear program represents the power network equations and constraints. The objective function is to minimize the interruption of power supply to customers where the right-hand side contains the capacities of the system generators, which are subject to random failures. We describe in detail this model in section 3.3.3.

3.3.1 Power Systems

The major parts of an electric power system are the generation, transmission and distribution systems. One may think of a power system as an extremely complicated electrical network. Power systems are basically a network of buses or nodes (buses and nodes will be used interchangeably) interconnected by transmission lines that carry the power flow. A bus is essentially an assembly of conductors for collecting electric currents and distributing them throughout the network. Also, each bus has an associated load that corresponds to a regional customer power demand that it serves; furthermore, some buses have generators that inject power into the network. Each bus has the following four associated parameters:

- (1) Voltage magnitude (V),

- (2) Voltage Angle (δ),
- (3) Real Power Injection (P) and
- (4) Reactive Power Injection (Q).

These parameters along with a basic coverage of circuit theory and the mathematical representation of load flow between the buses in the network will be discussed in further detail in the next section.

3.3.2 Fundamentals of Circuit Theory

For the following discussion, we advise consulting an engineering text on circuit analysis such as [34] for the technical description of the electrical terms and concepts; see also appendix C in [7]. One may think of power as the rate of change of energy with respect to time. The unit of electrical power is the watt (W), which is a joule per second. A *load* is an electrical device connected to a power source. The electrical power consumed by a load depends on two things: how much voltage is applied to the device, and how much current (flow of electricity) flows through the device. For example, a 100 watt light bulb consumes 100 joules of electric energy each second.

The mathematical representation of power flow through a network is a system of nonlinear equations. These equations are based on Kirchhoff's current and voltage laws. The current law basically states that the current entering and leaving a bus or node must be equal and the voltage law states that the total voltage around a closed loop must be zero.

Power flow in a network has two main components: the “real” power P in watts (W) and the “reactive” power Q in voltamperes–reactive (var). In general, electrical engineers refer to P as the resistance and Q as the reactance. Let Ω_i be the set of all nodes connected to bus i , $i = 1, \dots, n$. The real (3.12) and reactive (3.13) power injections at bus i are given by the following *alternating current load flow equations*:

$$P_i = \sum_{j \in \Omega_i} P_{ij} = V_i \sum_{j \in \Omega_i} V_j (G_{ij} \cos \delta_{ij} + B_{ij} \sin \delta_{ij}) \quad (i = 1, \dots, n) \quad (3.12)$$

$$Q_i = \sum_{j \in \Omega_i} Q_{ij} = V_i \sum_{j \in \Omega_i} V_j (G_{ij} \sin \delta_{ij} - B_{ij} \cos \delta_{ij}) \quad (i = 1, \dots, n), \quad (3.13)$$

where V_i and δ_i are the magnitude and angle of the voltage at bus i ; $\delta_{ij} = \delta_i - \delta_j$ is known as the power factor angle. G_{ij} and B_{ij} are called *conductance* and *susceptance* of circuit ij , respectively, and n is the number of system buses. It can be shown that

$$G_{ij} = -\frac{R_{ij}}{R_{ij}^2 + X_{ij}^2} \quad \text{and} \quad B_{ij} = \frac{X_{ij}}{R_{ij}^2 + X_{ij}^2}. \quad (3.14)$$

where R_{ij} and X_{ij} are the resistance and reactance of circuit ij .

Note that each bus has four variables, V_i , δ_i , P_i and Q_i and that the nonlinear system has $2n$ equations with $4n$ variables. Thus two of the aforementioned variables for each bus must be prespecified; generally, P_i and Q_i are specified. These loadflow equations can be solved numerically for V_i and δ_i using an iterative method such as the Gauss–Seidel or Newton-Raphson method. From the solution, the total real and reactive power flows in line ij become

$$P_{ij} = V_i V_j (G_{ij} \cos \delta_{ij} + B_{ij} \sin \delta_{ij}) \quad (3.15)$$

$$Q_{ij} = V_i V_j (G_{ij} \sin \delta_{ij} - B_{ij} \cos \delta_{ij}). \quad (3.16)$$

However, since reliability evaluation studies require that load flows must be repeated for each state or scenario of the system, e.g., tie-line failures or generator outages, assumptions are made in power networks that greatly simplify the nonlinear complexities of (3.12)-(3.13). Often, reactive power is of no concern and thus the system (3.13) can be dismissed. That is, many of the important reliability indices are associated with real power load curtailments and calculating these only requires real power related information. Furthermore, in power systems, power factor angles δ_{ij} are generally very small and thus the following approximations can be made: $\sin \delta_{ij} \approx \delta_i - \delta_j$ and $\cos \delta_{ij} \approx 1$. Also, circuit reactances are normally much larger than circuit resistances ($X_{ij} \gg R_{ij}$) thus, reliability engineers generally use the following approximation from (3.14) for circuit conductance and susceptance:

$$G_{ij} = -\frac{R_{ij}}{R_{ij}^2 + X_{ij}^2} \approx 0 \quad \text{and} \quad B_{ij} = \frac{X_{ij}}{R_{ij}^2 + X_{ij}^2} \approx \frac{1}{X_{ij}}. \quad (3.17)$$

Network reactances, resistances, real power, reactive power and voltage are usually expressed in a “per unit” (p.u.) system wherein a reference voltage and usually (mega watts) MW level are specified and all variables are normalized with respect to these references. In a per unit system, the voltage magnitudes are usually

unity. Here, all bus voltage magnitudes are assumed to be 1.0 p.u. Thus, based on the above assumptions and using (3.15) the real line flow in a branch can be calculated by

$$P_{ij} = \frac{\delta_i - \delta_j}{X_{ij}}, \quad (3.18)$$

therefore bus real power injections are

$$P_i = \sum_{j \in \Omega_i} P_{ij} = B_{ii}\delta_i - \sum_{j \in \Omega_i} B_{ij}\delta_j \quad (i = 1, \dots, n), \quad (3.19)$$

where

$$B_{ij} = \frac{1}{X_{ij}}, \quad B_{ii} = \sum_{j \in \Omega_i} B_{ij} \quad \text{and} \quad (3.20)$$

Ω_i is the set of all branches connected to bus i .

Now, equations (3.19) can be represented in matrix form:

$$P = B\delta, \quad (3.21)$$

where $B = [B_{ij}]$ is the *susceptance* matrix and $\delta = (\delta_1, \delta_2, \dots, \delta_n)'$, where $'$ means transpose. In addition, we assume that lines are lossless, that is, $P_1 + P_2 + \dots + P_n = 0$. This ensures that the load demanded is greater than or equal to the load generated. Note that (3.21) is a linear system and the real power injections depend only on the bus voltage angles; however, B as defined in (3.20) is singular. Since the equation for node n is the negative of the sum of the equations for nodes $1, \dots, n-1$ the linear

system involves only $n - 1$ independent equations. Essentially, the last equation is redundant. Thus, engineers typically will assign a zero node voltage angle to node n and then solve the following system:

$$P_0 = B_0 \delta_0, \tag{3.22}$$

where B_0 is the $(n - 1) \times (n - 1)$ submatrix of B obtained by deleting the last row and column of B , $\delta_0 = (\delta_1, \delta_2, \dots, \delta_{n-1})$ and $P_0 = (P_1, P_2, \dots, P_{n-1})$. Here $P_0 = P_1 + P_2 + \dots + P_{n-1}$ to ensure the equality of supply and demand and hence node n is often referred to as the *slack* bus.

The solution of (3.22) involves matrix inversion which can be accomplished directly and is, therefore, much faster than the iterative methods needed for solving the original nonlinear equations. This method of determining the real flows through solving first for the bus angles is often referred to as the linearized DC method of load flows, in contrast with the exact, non-linear solution, which is termed the AC solution. The term “DC load flow” arose because the linear relationship between P and δ is analogous to the relationship between current and voltage in a direct current network, which contains only resistors.

3.3.3 Linear Programming Power Flow Model

Power system reliability evaluation can be modeled as a linear program with stochastic right hand sides. There are two main classes of constraints: (1)

power flow equations and (2) operating constraints. Using the model we developed in the previous section, the power flow equations are modeled by the following set of linear equations:

$$B_0 \delta_0 + g_0 = d_0, \quad (3.23)$$

where B_0 is the *susceptance* matrix, δ_0 is the vector of node voltage angles, $g_0 = (g_1, g_2, \dots, g_{n-1})$ is the active power generation vector and $d_0 = (d_1, d_2, \dots, d_{n-1})$ is the load vector. That is, g_{0_i} is the generating capacity in MW of the i^{th} bus. If a given bus does not have a generator, then the corresponding generating capacity is 0. Likewise, d_{0_i} is the load at bus i . By rewriting (3.23) as $B_0 \delta_0 = d_0 - g_0$ we can compare it to (3.22) so that $d_0 - g_0$ corresponds to P_0 . A positive element of $d_0 - g_0$ represents a bus load and a negative value represents available capacity. Thus, given g_0 and d_0 , the solution vector δ_0 to (3.23) is found and active power flow in circuit ij is given as $P_{ij} = (\delta_i - \delta_j)/X_{ij}$; see (3.18). Consequently, these equations model the distribution of electrical power in the network circuits. Load uncertainty always exists in actual power systems and it has long been recognized that they can have a great impact in power system reliability evaluation. An accepted approach to simulating load uncertainty is to model the load as normal random variables. The mean values are estimated load means based on historical data. The standard deviation is assigned according to the perceived load forecast uncertainty, such as 5%.

The operating constraints place limits on power generation $g_{0_i} \leq \bar{g}_{0_i}$ and

power flow $P_{ij} \leq \bar{P}_{ij}$. The generating units in the capacity vector are subject to random failures and are thus modeled as random variables. For example, suppose a generator on a bus, say i , is comprised of three units each having a generating capacity of 100 MW. Each of these units is prone to failure and so we can model the total capacity available at generator i as the following random variable:

$$\bar{g}_i = \sum_{k=1}^3 100\chi_k, \quad (3.24)$$

where $\chi_k, k = 1, \dots, 3$ are independent Bernoulli random variables. If $\chi_k = 0$ for some k , then the corresponding generating unit is down, otherwise it is operating. If $\chi_k = 0$ for all k , then no power can be supplied. This modeling technique is similarly used for all buses with generators. Likewise, weather related outages of transmission lines can cause entire power interruption between system buses. To model power line outages between circuit ij , we use the following random variable.

$$\bar{P}_{ij} = \sum_{k=1}^3 CAP_{ij}\chi_{ij}, \quad (3.25)$$

where CAP_{ij} is the flow capacity of circuit ij and χ_{ij} is a Bernoulli random variable.

Overloads caused by generator or circuit outages can often be eliminated by rescheduling the system generators. In some severe situations, it may be necessary to curtail load in the system; that is, to potentially redirect energy through the network to minimize load curtailment. Load curtailment is represented by “fictitious” generators placed at each bus. That is, when the required load at a given

bus cannot be completely satisfied due to the insufficient total availability generating capacity and/or limits of tie line capacities, the “fictitious generator variables” can provide the required load such that power balance at each bus is always guaranteed. Thus, the fictitious generator variables are load curtailment variables for each associated bus and therefore the upper limit is assigned as the associated bus load. The generator variables, the fictitious generator variables and tie lines constitute a generator–transmission system. The objective is to minimize total load curtailment while satisfying the power flow equations and operating constraints. The following linear program [1, 43, 44, 52] can be used for this purpose:

$$L(\bar{d}, \bar{g}, \bar{f}) = \min \quad \sum_{i=1}^n r_i$$

$$\text{s.t.} \quad B\delta + g + r = \bar{d} \quad (3.26)$$

$$g \leq \bar{g} \quad (3.27)$$

$$r \leq \bar{d} \quad (3.28)$$

$$|P_{ij}| \leq \bar{P}_{ij} \text{ for all circuits } ij \quad (3.29)$$

$$g, r \geq 0, \delta \text{ free.} \quad (3.30)$$

where r_i is the “fictitious generator variable” at bus i . Hence EPNS = $E[L(\bar{d}, \bar{g}, \bar{f})]$ and LOLP = $E[1_{L(\bar{d}, \bar{g}, \bar{f}) > 0}]$.

We employ the load curtailment linear programming model presented in the previous section to demonstrate the use of quasi control variates. The IEEE

Subcommittee on the Application of Probability Methods has developed a Reliability Test System [48] (RTS) which includes both generation and major transmission facilities. The main objective was to provide a standard basic model which could be used to test or compare methods for reliability analysis of power systems. Using a modified version of the RTS we estimate the EPNS and LOLP. Tie line and bus generating capacity data are given in Tables B.1 and B.2. Notice that this table also provides the probability failure of a given tie line and its capacities in MW. Generating unit capacities are given in Table B.3 and bus load data are provided in Table B.4.

3.4 Numerical Experiments

Using the algorithm described in Section 2.4 we allot 60 seconds for the n th segment and allow for a total simulation time of $t = 30$ minutes. We consider various dual approximations where ν , given in the Table below, represents the number of dual vertices. Also, ρ is the correlation between the objective functions of the analytical problem and the dual approximation; r is the relative effort involved in obtaining the objective function value of the approximate problem to that of solving the linear program in the main simulation. Results are given in Tables 3.6 and 3.7 where the Speedup is defined as the estimated variance of the “crude” Monte Carlo estimate divided by the variance of the QCV estimator. Also, due to the length of the simulation time, overhead costs are insignificant.

ν	EPNS Speedup	ρ	r
2	2.31	0.829764	0.018577
5	2.70	0.853090	0.020492
10	3.66	0.871843	0.025262
15	3.35	0.876689	0.025749
20	5.31	0.921301	0.027271
25	6.40	0.941632	0.028887
30	6.38	0.935184	0.031484
40	6.12	0.942480	0.034804
50	6.87	0.950231	0.035146
80	8.72	0.963415	0.041050

Table 3.6. EPNS Speedup

ν	LOLP Speedup	ρ	r
2	4.73	0.900753	0.020569
5	4.86	0.907416	0.021833
10	6.29	0.931023	0.023687
15	5.31	0.918041	0.025076
20	5.78	0.931357	0.026676
25	6.03	0.931847	0.028752
30	8.10	0.950192	0.031000
40	10.8	0.967068	0.032710
50	8.51	0.954699	0.036998

Table 3.7. LOLP Speedup

4. Conclusion and Future Work

4.1 Summary

We have generalized the notion of classical control variate methodology to the case where the control variate mean is unknown and have developed a procedure for its implementation in applications. We have demonstrated through a couple of “real-word” illustrations that the QCV scheme potentially provides tremendous speedup.

4.2 Future Research

In our research, we considered having only one QCV estimator. A possible extension to the QCV scheme is the potential of multiple estimators. Also, we used the QCV scheme to estimate the mean of the objective functions of stochastic linear programs. It would be interesting to incorporate the QCV method into algorithms for actually solving stochastic programs. One such algorithm may involve stochastic quasi gradient methods whereby we attempt to improve efficiency in estimating the quasi gradients. We demonstrated examples that involve exclusively linear programming models. We would like to see new applications for the QCV method, i.e., stochastic integer problems, stochastic networks or stochastic vehicle routing problems.

A. APPENDIX Refining Fractionation Data

Fraction	Yield of input, %	Destination	Model
	(x_{18}, x_{19})		Variable
Gas	8.00	Combustible	x_{30}
Propane	4.00	Propane	x_{31}
		Combustible	x_{32}
Butane	4.50	Butane	x_{33}
		Combustible	x_{34}
Gasoline	81.50	Gosoline I	x_{35}
		Gasoline II	x_{36}
Loss	2.00	Re-forming	x_{37}

Table A.1. Re-forming

Fraction	Yield of input, %	Destination	Model
	(x_{23}, x_{25}, x_{27})		Variable
WD	47.00	Catalytic cracking	x_{43}
Gas Oil	5.00	Gas oil	x_{44}
Residue	43.00	Pitch	x_{45}
		Fuel	x_{46}
		Combustible	x_{47}
Loss	5.00	Loss	x_{48}

Table A.2. Vacuum

Fraction	Yield of input, %	Destination	Model
	(x_{43})		Variable
Gas	3.50	Combustible	x_{49}
Catalytic polymerizator	7.50	Catalytic polymerizator	x_{50}
Gasoline	22.00	Gasoline I	x_{51}
		Gasoline II	x_{52}
LCO	32.00	Gas Oil	x_{53}
		Fuel	x_{54}
HYCO	28.50	Gas Oil	x_{55}
		Fuel	x_{56}
Loss	6.50	Loss	x_{57}

Table A.3. Catalytic Cracking Unit

Fraction	Yield of input, %	Destination	Model
	(x_{43})		Variable
Propane	10.00	Propane	x_{58}
		Combustible	x_{59}
Butane	30.00	Butane	x_{60}
		Combustible	x_{61}
Gasoline	60.00	Gasoline I	x_{62}
		Gasoline II	x_{63}

Table A.4. Catalytic polymerizator

B. APPENDIX Power System Bus and Tie-Line data

From bus	To bus	Susceptance (p.u. 100 MW base)	Capacity (MW)	Failure Probability
1	2	71.9	175	0.00044
1	3	4.7	175	0.00058
1	5	11.8	175	0.00038
2	4	7.9	175	0.00045
2	6	5.2	175	0.00055
3	9	8.4	175	0.00043
3	24	11.9	400	0.00175
3	24	11.9	400	0.00175
4	9	9.6	175	0.00041
5	10	11.3	175	0.00039
6	10	16.5	175	0.00132
7	8	16.3	175	0.00034
8	9	6.1	175	0.00050
8	10	6.1	175	0.00050

Table B.1. Branch Data

From bus	To bus	Susceptance (p.u. 100 MW base)	Capacity (MW)	Failure Probability
9	11	11.9	400	0.00175
9	12	11.9	400	0.00175
10	11	11.9	400	0.00175
10	12	11.9	400	0.00175
11	13	21.0	500	0.00050
11	14	23.9	500	0.00050
12	13	21.0	500	0.00050
12	23	10.4	500	0.00065
13	23	11.6	500	0.00062
14	26	25.7	500	0.00048
15	16	57.8	500	0.00041
15	21	20.4	500	0.00051
15	24	19.3	500	0.00051
16	17	38.6	500	0.00044
16	19	43.3	500	0.00044
17	18	69.4	500	0.00040
17	22	9.5	500	0.00068
18	21	36.8	500	0.00044
19	20	25.3	500	0.00048
20	23	46.3	500	0.00043
21	22	14.7	500	0.00057

Table B.2. Branch Data (cont.)

Bus	Unit 1 p.u.	Unit 2 p.u.	Unit 3 p.u.	Unit 4 p.u.	Unit 5 p.u.	Unit 6 p.u.
1	0.20	0.20	0.76	0.76		
2	0.20	0.20	0.76	0.76		
7	0.10	0.10	0.10	0.10		
13	1.97	1.97	1.97			
15	0.12	0.12	0.12	0.12	0.12	1.55
16	1.55					
18	4.00					
21	4.00					
22	0.50	0.50	0.50	0.50	0.50	0.50
23	1.55	1.55	1.55			

Table B.3. Generating Unit Locations

Bus	Load (p.u.)	Bus	Load (p.u.)
1	1.08	10	1.95
2	0.97	13	2.65
3	1.80	14	1.94
4	0.74	15	3.17
5	0.71	16	1.00
6	1.36	18	3.33
7	1.25	19	1.81
8	1.71	20	1.28
9	1.75		

Table B.4. Bus Load Data

References

- [1] M.V.F. Pereira A.C.G. Melo and A.M. Leite da Silva. Frequency and duration calculations in composite generation and transmission reliability evaluation. *IEEE Transactions on Power Systems*, 7(10):469–475, May 1992.
- [2] S. Asmussen and R.Y. Rubenstein. Complexity properties of steady-state rare events simulation in queueing models. In J. Dshalalow, editor, *Advances in Queueing*, pages 429–462. CRC Press, 1995.
- [3] D. Avis. A c implementation of the reverse search vertex enumeration algorithm. Technical report, Kyoto University, 1994.
- [4] D. Avis and K. Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete and Computational Geometry*, 8:295–313, 1992.
- [5] E.M.L. Beale. On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society, Series B*, 17:173–184, 1955.
- [6] R. Billinton and R.N. Allan. *Reliability Evaluation of Power Systems*. Plenum Press, New York, New York, 1984.
- [7] R. Billinton and W. Li. *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods*. Plenum Press, NY, 1994.
- [8] J.R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer-Verlag New York, Inc, 1997.
- [9] R.G. Bland. New finite pivoting rules for the simplex method. *Mathematics of Operations Research*, 2(2):103–107, May 1977.
- [10] M.R. Taaffe B.W. Schmeiser and J. Wang. Biased control-variate estimation. *To appear in IIE Transactions*, 2000.

- [11] M.R. Taaffe B.W. Schmeiser and J. Wang. Control–variate estimation using estimated control means. *To appear in IIE Transactions*, 2000.
- [12] A. Charnes and W.W. Cooper. Chance-constrained programming. *Management Science*, 6:73–79, 1959.
- [13] A. Charnes and W.W. Cooper. Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations Research*, 11:18–39, 1963.
- [14] R.C.H. Cheng and G.M. Feast. Control variables with known mean and variance. *Journal of the Operational Research Society*, 31:51–56, 1980.
- [15] G.B. Dantzig. Linear programming under uncertainty. *Management Science*, 1:197–206, 1955.
- [16] G.B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ, 1963.
- [17] G.B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ, 1963.
- [18] G.B. Dantzig and P.W. Glynn. Parallel processors for planning under uncertainty. *Annals of Operations Research*, 22:1–21, 1990.
- [19] G.B. Dantzig and P. Wolfe. On the solution of two–stage linear programs under uncertainty. In J. Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 165–167, Berkeley, CA, 1961. University of California Press.
- [20] I. Deák. Multidimensional integration and stochastic programming. In Y. Ermoliev and R.J-B Wets, editors, *Numerical Techniques for Stochastic Optimization*, chapter 7, pages 187–200. Springer Verlag, Berlin, 1988.
- [21] M. Emsermann and B. Simon. Improving simulation efficiency with quasi control variates. *Submitted to Stochastic Models*, 2000.
- [22] J. Endrenyi. *Reliability Modeling in Electric Power Systems*. John Wiley & Sons, New York, New York, 1978.
- [23] Y. Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics*, 9:1–36, 1983.

- [24] Y. Ermoliev and A. Gaivoronsk. Stochastic quasigradient methods for optimization of discrete systems. *Annals of Operations Research*, 39:1–39, 1992.
- [25] A. Ferguson and G.B. Dantzig. The allocation of aircraft to routes: An example of linear programming under uncertain demands. *Management Science*, 3:45–73, 1956.
- [26] A. Gaivoronski. Implementation of stochastic quasigradient methods. In Y. Ermoliev and R.J-B Wets, editors, *Numerical Techniques for Stochastic Optimization*, chapter 16, pages 313–351. Springer Verlag, Berlin, 1988.
- [27] J.D. Glover and M. Sarma. *Power System Analysis and Design*. PWS Publishing Company, Boston, MA, 1994.
- [28] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, 1992.
- [29] C.A. Gross. *Power System Analysis*. John Wiley & Sons, New York, New York, 1979.
- [30] A. Hall. On an experiment determination of π . *Messeng. Math.*, 2:113–114, 1873.
- [31] J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods*. Methuen, London, 1964.
- [32] G. Infanger. Monte carlo (importance) sampling within a benders decomposition algorithm for stochastic linear programs. *Annals of Operations Research*, 39:41–67, 1992.
- [33] G. Infanger. *Planning Under Uncertainty: Solving Large-Scale Stochastic Linear Programms*. Boyd and fraser publishing company, Danvers, MA, 1994.
- [34] J.D. Irwin. *Basic engineering circuit analysis, 3rd edition*. Macmillan Publishing Company, NY, 1990.
- [35] R.J. Vanderbei J.M. Mulvey and S.A. Zenios. Robust optimization of large-scale systems. *Operations Research*, 43(2):264–281, 1995.

- [36] P. Kall. Approximation to optimization problems: An elementary review. *Mathematics of Operations Research*, 11:9–18, 1986.
- [37] P. Kall and S.W. Wallace. *Stochastic Programming*. John Wiley and Sons Ltd, Chichester, UK, 1994.
- [38] S.S. Lavenberg and P.D. Welch. A perspective on the use of control variables to increase the efficiency of monte carlo simulations. *Management Science*, 27:322–335, 1981.
- [39] S.S. Lavenberg and P.D. Welch. A perspective on the use of control variates to increase the efficiency of monte carlo simulation. *Management Science*, 27:322–335, 1981.
- [40] A.M. Law and W.D. Kelton. *Simulation Modeling and Analysis, 2nd edition*. McGraw–Hill, Inc., New York, 1991.
- [41] D.G. Luenberger. *Linear and Nonlinear Programming, 2nd edition*. Addison–Wesley Publishing Company, Inc., 1973.
- [42] O.L. Mangasarian. Nonlinear programming problems with stochastic objective functions. *Management Science*, 10:353–395, 1964.
- [43] J. Mitra and C. Singh. Incorporating the dc load flow model in the decomposition–simulation method of multi–area reliability evaluation. *IEEE Transactions on Power Systems*, 11(3):1245–1254, August 1996.
- [44] G.C. Oliveira M.V.F. Pereira, L.M.V.G. Pinto and S.H.F. Cunha. A technique for solving lp problems with stochastic right hand sides. Technical report, CEPEL, Centro del Pesquisas de Energia Electria, Rio de Janeiro, Brazil, 1989.
- [45] G.C. Oliveira M.V.F. Pereira, M.E.P. Maceira and L.M.V.G. Pinto. Combining analytical models and monte carlo techniques in probabilistic power system analysis. *Transactions on Power Systems*, 7(1):265–272, February 1992.
- [46] M.V.F. Pereira and L.M.V.G. Pinto. A new computational tool for composite reliability evaluation. *Transactions on Power Systems*, 17(7):258–263, February 1992.
- [47] D. Schaeffer P.J. Fleming and B. Simon. Efficient monte–carlo simulation

of a product –form model for a cellular system with dynamic resource sharing. *ACM Transactions on Modeling and Computer Simulation*, 5:3–21, 1995.

- [48] IEEE Committee Report. Ieee reliability test system. *IEEE Transactions on Power Apparatus and Systems*, 98:2047–2054, 1979.
- [49] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of mathematical statistics*, 22:400–407, 1951.
- [50] R.Y. Rubstein and R. Marcus. Efficiency of multivariate control variates in monte carlo simulation. *Operations Research*, 33:661–677, 1985.
- [51] A. Sankarakrishnan and R. Billinton. Sequential monte carlo simulation for composite power system reliability analysis with time varying loads. *IEEE Transactions on Power Systems*, 10(3):1540–1545, August 1995.
- [52] L.M.V.G. Pinto S.H.F. Cunha, M.V.F. Pereira and G.C. Oliveira. Composite generation and transmission reliability evaluation in large hydroelectric systems. *IEEE Transactions on Power Apparatus and Systems*, 104(10):2657–2663, October 1985.
- [53] R. Van Slyke and R.J-B Wets. L-shaped linear programs with application to optimalcontrol and stochastic programming. *SIAM Journal on Applied Mathematics*, 17:638–663, 1969.
- [54] T.L. Moeller S.S. Lavenberg and C.H. Saucer. Concomitant control variables applied to the regenerative simulation of queueing systems. *Operations Research*, 27(1):134–160, 1979.
- [55] G.J. Stigler. The cost of subsistence. *Journal of Farm Economics*, 27:303–314, 1945.
- [56] G. Tintner. A note on stochastic linear programming. *Econometrica*, pages 490–495, 1960.
- [57] H. Vladimirov and S.A. Zenios. Stochastic programming and robust optimization. In T. Gal and H.J. Greenberg, editors, *Advances in Sensitivity Analysis and Parametric Programming*, chapter 12. Kluwer Academic Publishers, Boston, 1997.

- [58] C.L. Wadhwa. *Electric Power Systems*. John Wiley & Sons, New York, New York, 1983.
- [59] K. G. Waguespack and J. F. Healey. Managing crude oil quality for profitability in refining. *Hydrocarbon Processing*, 77(9), 1998.
- [60] J.R. Wilson. Variance reduction techniques for digital simulation. *Am. J. Math. Manag. Sci.*, 4:277–312, 1984.
- [61] R.W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1989.
- [62] A.J. Wood and B.F. Wollenberg. *Power Generation Operation and Control*. John Wiley & Sons, New York, New York, 1984.
- [63] N. WU and R. Coppins. *Linear Programming and Extensions*. McGraw-Hill, New York, New York, 1981.