Fast Computation of Loss Distributions for Credit Portfolios

Quantitative Analytics Research Group
Standard & Poor’s

William Morokoff* and Liming Yang†
55 Water Street, 44th Floor
New York, NY 10041
* william_morokoff@sandp.com, 212 438 4828
† liming_yang@sandp.com, 212 438 1870

Mathematical Problems in Industry Workshop
Department of Mathematics
New Jersey Institute of Technology
June 2011

Copyright © 2011 by Standard & Poors Financial Services LLC (S&P), a subsidiary of The McGraw-Hill Companies, Inc. All rights reserved.
Abstract

We pose here an optimization problem related to determining the best set of parameters for minimizing the variance of a Monte Carlo estimator of the probability of a rare event associated with potentially large losses on a portfolio of credit-risky investments. The parameters are associated with a specific importance sampling framework that seeks to sample scenarios with large losses by increasing the default correlation among the portfolio’s assets.

Disclaimer: The mathematical framework and questions posed here represent a research project intended to study improvements in computational speed for a general class of Monte Carlo simulations. No inferences should be made with regard to Standard & Poor’s credit ratings or any current or future criteria or models used in the ratings process for credit portfolios or any type of financial security.

1 Introduction

Assessing the potential for loss on portfolios of financial assets is an important part of the risk management operations for banks, asset managers, insurance companies, etc. Portfolios with substantial credit risk exposure from holdings such as bonds, loans and credit derivatives face particular challenges due to skewed nature of returns on credit instruments arising from the high probability of relative low positive returns and small probability of relatively high negative returns associated with defaults. Systematic exposure to macro-economic, regional and sector factors across the portfolio must also be accounted for. The result is the distribution of potential losses on a credit portfolio is often heavily skewed with quite significant losses occurring at the 0.1% probability level. Banks often use such modeling to help establish capital requirements to survive periods of extreme credit stress.

One common approach to credit portfolio loss modeling is to combine a process for changes in individual obligor credit quality (i.e. modeling credit transitions including default) with a dependence structure across obligors that captures the joint credit quality evolution. In some models, dependence structure is represented by a normalized Gaussian distribution (or other multivariate distribution) used to model joint changes in the underlying value of the obligors; such distributions are characterized by an explicit correlation matrix or an associated factor model that specifies the obligor correlations. More details on credit portfolio modeling can be found in [Bohn and Stein 2009] and [Blum et al. 2010].

When the dependence structure of a credit portfolio is complex (e.g. joint credit evolution cannot be characterized by exposure to a single factor), Monte Carlo simulation is often employed to compute credit portfolio loss distributions. While this allows for detailed modeling at the individual exposure level, it comes with a substantial computational cost. This is particularly true as we are often interested in the far tail of the loss distribution corresponding to rare events, as well as sensitivity of tail loss size to portfolio parameters such as individual exposure size. For this reason there has been a significant amount of work in developing importance sampling methods for improving the computational performance of the Monte Carlo simulations. Methods such as those described in [Glasserman et al. 2008], [Glasserman and Li 2005] and [Morokoff 2004] focus on increasing the number of samples corresponding to the loss level of interest. Significant work also continues in related areas of analytical approximations for the credit portfolio problem [Voropaev 2011].

One promising area of research in developing importance sampling methods concerns creating a new sampling measure based on modifying the dependence structure to increase the number of correlated defaults. While
such methods have been used successfully for certain credit portfolio problems, it is of interest to develop a general optimal method for modifying the dependence structure. The optimization should involve the dependence structure as well as the portfolio characteristics (obligor exposure size, etc.) and the level of loss of interest, with the goal being to determine an importance sampling measure that minimizes the variance of the estimator of the portfolio loss at the specified loss level (e.g. losses occurring with probability less than 0.1%). Ideas here include determining an optimal basis for the dependence structure such that increasing the volatility of samples in key basis directions minimizes the simulation error. Adaptive or staged importance sampling, in which information from previous simulation trials is used to update the optimization, may also be of interest.

2 Credit Portfolio Modeling Framework

Credit risk is broadly defined as the risk of not receiving timely payment of contractually promised interest or principle payments. Credit portfolio risk assessment often begins with describing the credit state of each obligor (i.e. a borrower with an obligation to pay interest and principal). This may be done through classification into discrete credit categories typically known as ratings (whether assessed through a bank’s internal rating system or by an independent credit rating agency such as Standard & Poor’s), or through other continuous metrics (combined with an absorbing default state) such as probability of default (over a specified time horizon), distance to default, or default intensity. A transition to the default state is accompanied by a loss in promised interest and principal. However, mark-to-market losses may also result when credit quality deteriorates (i.e. transitions to a less credit-worthy state).

When viewed from today, $T_0$, the goal of a credit portfolio model is to describe the probability distribution of potential losses on the portfolio, arising from either defaults or credit deterioration, that may occur up to a horizon $T_H$, typically one year for bank loan portfolios. The portfolio value at time $t$ may be expressed as

$$\pi(t) = \sum_{i=1}^{N} \omega_i V_i(t)$$ (2.1)

where the portfolio has $N$ positions, each of notional $\omega_i$ and value (relative to par) at time $t$ of $V_i(t)$. One definition of loss to horizon for the portfolio is then

$$L_H = \pi(T_0)(1 + r) - \pi(T_H).$$ (2.2)

Here $r$ is the risk-free rate to horizon, and the loss is defined as relative to a risk-free investment. We are often interested in computing the loss level $L$ such that $P(L_H > L) = \alpha$ for a specified confidence level $\alpha$.

If we assume that the notional exposures are fixed, then the uncertainty in the loss comes from the uncertainty in the value of the assets, which in turn derives from the credit state of the associated obligor. Thus we are concerned with modeling $V_i(t) = V_i(z_i(t))$, where $z_i(t)$ is related to the credit state of the obligor associated with the $i^{th}$ asset. Whatever credit state description and associated transition dynamics are employed, it is usually possible to map the distribution of possible credit states at the horizon $T_H$ to a standard Normal distribution. For example transitioning at horizon to discrete credit state $j$ may be mapped to sampling a standard Normal random variate $z$ in the interval $z_j < z < z_{j+1}$ with the corresponding probability of the transition given by $\Phi(z_{j+1}) - \Phi(z_j)$ where $\Phi(z)$ is the cumulative standard Normal distribution.
Large credit portfolio losses occur during economic downturns when macro-economic and sector factors drive joint credit deterioration and default of multiple obligors. Thus it is key to capture this credit dependence across all obligors. In a portfolio model where the credit state of an obligor at horizon is described by a standard Normal random variable \( z_i \), it is the joint probability distribution of \( z = (z_1, \ldots, z_N) \) that determines the dependence and likelihood of joint defaults. In one of the most widely employed portfolio modeling frameworks, the distribution of \( z \) is assumed to be multi-variate Normal with mean zero, unit variance and a correlation matrix \( C \). Of course other joint distributions are possible. In other modeling frameworks, continuous time or multi-period evolution of credit quality and the dependence structure may be modeled.

Assuming the Gaussian framework, the standard single-step Monte Carlo simulation proceeds by sampling \( z \sim N(0, C) \) and evaluating the portfolio loss \( L_H(z) \) associated with horizon credit states implied by the simulation draw \( z \). The probability distribution of losses is then built up based on many samples (often hundreds of thousands). The probability of exceeding a given loss level is estimated by

\[
P(L_H > L) = E(X_{[L, \infty]}(L_H))
\]

\[
\approx \frac{1}{M} \sum_{j=1}^{M} X_{[L, \infty]}(L_H(z_j))
\]

Here \( X_{[a, b]} \) is the characteristic function of the interval \([a, b]\), and \( M \) is the number of simulation runs. In a large portfolio with tens or hundreds of thousands of obligors, it is impractical to directly specify and store a huge correlation matrix. Generally correlations are specified through either a block correlation structure with groupings according to sector and geography, or a factor model by which the credit state is specified as a weighted combination of a relatively small number of systematic factors (usually related to sector, geography, etc.) and an independent, obligor-specific idiosyncratic factor. In the factor model setting with \( N_F \) systematic factors, the credit state of the portfolio \( z \) with \( N \) obligors may be specified as

\[
z = \Gamma^{1/2} B \epsilon_F + (I - \Gamma)^{1/2} \epsilon_I
\]

where \( \Gamma \) is an \( N \times N \) diagonal matrix specifying the percentage of variance attributed to systematic factors versus the idiosyncratic factor, \( B \) is an \( N \times N_F \) matrix of factor loadings, normalized such that \((BB^T)_{ii} = 1\), and the vectors \( \epsilon_F \) and \( \epsilon_I \) are the independent standard Normal systematic and idiosyncratic draws. The corresponding correlation matrix is then

\[
C = \Gamma^{1/2} BB^T \Gamma^{1/2} + I - \Gamma.
\]

### 3 Importance Sampling Method

Importance Sampling is a Monte Carlo simulation variance reduction technique that works by replacing the original sampling measure with a new measure that concentrates the samples in a region of interest. For the purpose of assessing the probability of exceeding a large loss level in a credit portfolio, one approach is to increase correlations among the obligors, thereby leading to more high-default and therefore large loss samples. However, this must be done carefully to avoid introducing so many extreme loss samples that the accuracy of the estimator decreases.
In general, let \( \phi(\theta) \) be a probability density depending on a parameters \( \theta \) of a vector valued random variable \( z \), where the parameters \( \theta \) take their values in a subset of \( \mathbb{R}^d \) for a \( d \) dimensional parameter space. Define the expectation of the function \( h(z) \) with respect to the measure \( \phi(\theta) \) as

\[
E_{\theta}[h] = \int h(z) \phi(\theta) dz
\]  

(3.1)

Let \( \phi_{\theta_0}(z) \) be the original reference or sample density and

\[
\mu_h = E_{\theta_0}[h]
\]

(3.2)

be the quantity of interest. For the credit portfolio model, \( h(z) \) would be the indicator function for the interval of losses exceeding a given level, while \( \phi_{\theta_0}(z) \) would be the specified Gaussian distribution with correlation matrix \( C \). In this case, \( \mu_h \) would be the probability of exceeding the given loss level. To implement the importance sampling method, define

\[
W_{\theta}(z) = \frac{\phi_{\theta_0}(z)}{\phi_{\theta}(z)}
\]

(3.3)

to be the likelihood ratio or Radon-Nikodym derivative (weight function) associated with the change of measure (note that the support for \( \phi_{\theta_0}(z) \) should be contained in the support for \( \phi_{\theta}(z) \)). It follows that

\[
\mu = E_{\theta}[W_{\theta}h] = E_{\theta_0}[h].
\]

(3.4)

Define the second weighted moment as

\[
m_h(\theta) = E_{\theta}[(W_{\theta}h)^2] = E_{\theta_0}[W_{\theta}h^2]
\]

(3.5)

and the weighted variance as

\[
\sigma_h^2(\theta) = m_h(\theta) - \mu^2.
\]

(3.6)

Define the estimator

\[
\widehat{\mu}_h(M, \theta) = \frac{1}{M} \sum_{i=1}^{M} W_{\theta}(z_i) h(z_i)
\]

(3.7)

where \( z_i \) are i.i.d. samples drawn from the density \( \phi_{\theta}(z) \). If the weighted variance \( \sigma_h^2(\theta) \) is finite, then \( \widehat{\mu}_h(M, \theta) \) is an unbiased estimator of \( \mu_h \) that converges as \( M \to \infty \) with the estimation error proportional to \( \sigma_h(\theta) \). The objective is to find \( \theta^{opt} \) so that the variance of the new estimator \( \sigma_h^2(\theta) \) is minimized.

In addition to minimizing the variance of the estimator, there are two implementation issues that must be considered when selecting a new sample density for use in an importance sampling method. First, it must be relatively easy to sample from the new density, and second, it must be relatively easy to compute the value of the weight function for each sample. If either of these calculations imposes a significant increase in computation time, the benefits of reduced estimator variance (and therefore fewer required samples) may be outweighed by the additional computation time per sample. With this in mind, we propose the following importance sampling framework for credit portfolio problems in which the dependence structure is specified through a Gaussian copula (or more generally an elliptical copula specified by a positive definite covariance matrix).
Consider a credit portfolio with \( N \) distinct obligors for which the credit state at the horizon of obligor \( i \) can be derived from a standard Normal draw \( z_i \). Let the distribution of the column vector \( z = (z_1, \ldots, z_N)^T \) be specified as a multi-variate Normal distribution with zero mean, unit variance for all variables and pair-wise correlations of the variables given by the positive definite matrix \( C \), i.e. \( z \sim N(0, C) \). The standard Monte Carlo simulation would consist of sampling \( z \) from this distribution and evaluating the portfolio loss at this sampled value, then determining whether the loss exceeds the specified level in the process of estimating the probability of exceeding that level.

Now consider a class of probability density functions indexed by a parameter vector \( \theta = (\theta_1, \ldots, \theta_N) \) with \( -\infty < \theta_i < 1 \) and a linear basis for \( \mathbb{R}^N Q = [q_1, \ldots, q_N] \) where the \( q_i \) are \( N \times 1 \) vectors that are orthonormal with respect to the matrix \( C^{-1} \), i.e. \( q_i^T C^{-1} q_j = 0 \) for \( i \neq j \) and \( q_i^T C^{-1} q_i = 1 \) for all \( i, j \) between 1 and \( N \). Define the covariance matrix \( C(\theta, Q) \) as

\[
C(\theta, Q) = C + \sum_{i=1}^N \frac{\theta_i}{1-\theta_i} q_i q_i^T \tag{3.8}
\]

and let \( \phi_{C(\theta,Q)}(z) \) be the density function of the \( N \) dimensional normal distribution with mean zero and covariance matrix \( C(\theta, Q) \). Define

\[
W(z, \theta, Q) = \frac{\phi_C(z)}{\phi_{C(\theta,Q)}(z)}. \tag{3.9}
\]

Here \( \phi_C \) is the original sample density, where \( C = C(\theta_0, Q) \) with \( \theta_0 = (0, 0, \ldots, 0) \). Note that \( \theta \) may be restricted to a smaller dimensional space of size \( k < N \) by setting the remaining \( \theta_i = 0 \) and therefore only adding \( k \) additional terms in Equation [3.8]. This method extends the special case described in Morokoff [2004] in which \( k = 1 \) and \( q_1 \) was taken to be the eigenvector of \( C \) corresponding to the largest eigenvalue.

The weight function is the ratio of two zero mean multi-variate Normal density functions with covariances \( C \) and \( C(\theta, Q) \) respectively. It can be shown that the weight function can be written as

\[
W(z, \theta, Q) = \prod_{i=1}^N f_i(z, \theta_i, q_i), \tag{3.10}
\]

where

\[
f_i(z, \theta_i, q_i) = \frac{1}{\sqrt{1-\theta_i}} \exp\left(-\frac{\theta_i}{2} (q_i^T C^{-1} z)^2\right). \tag{3.11}
\]

We now address the feasibility of sampling from \( \phi_{C(\theta,Q)} \) and computing the weight function \( W(\theta, Q) \). It can be shown that if \( z \) is a sample from \( \phi_C \), then

\[
z_\theta = z + \sum_{i=1}^N \frac{1-\sqrt{1-\theta_i}}{\sqrt{1-\theta_i}} (q_i^T C^{-1} z) q_i \tag{3.12}
\]

is a sample from \( \phi_{C(\theta,Q)} \). Thus assuming the vectors \( C^{-1} q_i \) can be precomputed before the simulation, it is possible to sample \( z_\theta \) by sampling \( z \) from the original distribution with order \( kN \) additional work for each sample (where again \( k \) is the number of non-zero \( \theta_i \)). It also follows that

\[
(q_i^T C^{-1} z_\theta)^2 = \frac{1}{1-\theta_i} (q_i^T C^{-1} z)^2. \tag{3.13}
\]
so there is little additional computation required to calculate the sample weight function. This is an important observation because for most large portfolio with tens or hundreds of thousands of obligors, the original covariance matrix \( C \) would never be formed, but as described above would be specified through either a block correlation structure or a factor model. Efficient methods are known for sampling from the original distribution without forming the matrix \( C \) or related decompositions (Cholesky, etc.). See [Huang and Yang 2010] for details on the block correlation case. Thus the importance sampling framework described here can leverage this efficient sampling of the original distribution to sample from the new distribution \( \phi_{C(\theta,Q)} \). Alternately, if the full basis representation of \( Q \) is available, then the original distribution can be sampled as \( z = Q \epsilon \), where \( \epsilon \) is a \( N \times 1 \) vector of independent standard Normal variates, and the importance sampling distribution can be sampled as \( z_\theta = QD\epsilon \), where \( D \) is a diagonal matrix with \( D_{ii} = 1/\sqrt{1-\theta_i} \). Note that using the full basis and the full length \( N \) parameter vector \( \theta \) requires order \( N^2 \) work for each simulation sample. Therefore using a full basis may impose considerable extra computation time per sample for a large portfolio compared with sampling from the original distribution, which typically requires only order \( N \) calculations. Thus the best performance improvement may come from using a small number \( k \) of well-chosen directions. An alternate approach in the factor model setting is to consider importance sampling only for the independent standard Normal draws that are used to sample the correlated credit states. In this case, the original covariance matrix is the identity matrix, the basis \( Q \) can be taken to be the identity matrix, and the question is how much to scale the volatility of each of the \( N = N_F + N_O \) variables, where \( N_F \) is the number of factors and \( N_O \) is the number of obligors. The advantage here is that the full \( N \) dimensional vector \( \theta \) may be used while still only requiring order \( N \) calculations per sample.

We now consider the question of finding an optimal \( \phi_{C(\theta,Q)} \) is the sense of minimizing the variance \( \sigma^2_h(\theta) \) of the importance sampling estimate of \( \mu_h \). For the moment, take the basis \( Q \) to be given; a natural choice might be the eigenvector basis for \( C \) scaled so that \( \| q_i \|^2 = \lambda_i \) for eigenvalue \( \lambda_i \). In this case it can be observed that the effect of adding the \( i^{th} \) term to the original correlation matrix in constructing \( C(\theta,Q) \) is to scale the variance of the distribution in the \( q_i \) direction. We seek to find the optimal choice of \( \theta = (\theta_1, \ldots, \theta_k) \).

A key result for this framework is that with respect to the parameter vector \( \theta \), the Hessian matrix of the weight function \( \Delta_\theta W(z;\theta,Q) \), is positive definite for all finite \( z \). Therefore the quantity \( m_h(\theta) \) of Equation 3.5 must also have a positive definite Hessian with respect to \( \theta \). As we can also show that

\[
\frac{\partial m_h}{\partial \theta_i} < 0 \quad \text{as} \quad \theta_i \to -\infty \\
\frac{\partial m_h}{\partial \theta_i} > 0 \quad \text{as} \quad \theta_i \to 1
\]

it follows that there must be a unique solution to

\[
\nabla_\theta m_h(\theta_{opt}) = 0
\]

for which the variance of the importance sampling estimator is minimized. Note that this minimum depends on the basis \( Q \) chosen.

We now pose several questions regarding this importance sampling framework:

**Question 1.** For a given \( k \) and \( Q \), how can one best determine the optimal \( \theta \), remembering that the ultimate goal is to most efficiently estimate \( \mu \) and related quantities. One could estimate \( m_h(\theta) \) and its derivatives through simulation, then use a Newton iteration to find a new \( \theta \), then repeat the simulation. The
difficultly is to avoid spending so much time on the search for the optimal $\theta$ that it exceeds the time required to simulation to sufficient accuracy under the original measure. Another approach could be to adaptively update $\theta$ as the simulation proceeds. If we replace the single parameter $\theta$ by a sequence of parameters $\{\theta_k\}$ and use the estimator

$$e_{\{\theta_k\}}(M, h) = \frac{1}{M} \sum_{i=1}^{M} W_{\theta_{k-1}}(z_i) h(z_i).$$

(3.17)

where $z_i$ is drawn from $\phi_{\theta_{k-1}}(z)$ that is independent of $(z_j, \theta_j | j = 0, 1, i-1)$, then the estimator is called adaptive importance sampling. Suppose that the sequence $\theta_k$ converges to $\theta^{opt}$, then under some integrability conditions on $W_{\theta^{opt}}(z) h(z)$, the estimator converges to $E(h(z))$.

In order to generate a sequence $\{\theta_k\}$, we can consider stochastic optimization methods. Since the Hessian is positive definite, we could apply the generalized Robbins-Monro algorithm (see Egloff and Leippold [2010] and Arouna [2003] for more details) as follows:

$$\theta_{k+1} = \theta_k - a_k \nabla_{\theta} m_{h}(\theta_k),$$

(3.18)

with $a_k$ being a decreasing sequence that satisfies the condition $\sum a_k = \infty$, and $\sum a_k^2 < \infty$. The difficulty of the algorithm is to choose suitable $a_k$ so that the convergent speed is not too slow. The choice of $a_k$ will depend on the Hessian. How can we choose $a_k$ to get the fastest convergent speed in our case?

**Question 2.** For $k = 1$, is there an optimal direction $q_1$ such that $\nabla_{q_1, \theta_1} m_{h}(q_1, \theta_1) = 0$? We know that for a fixed $q_1$ we can find an optimal $\theta$. In the one dimensional case, the $\Delta_{q_1}(m_{h}(q))$ is no long positive definite. Thus, we might have many local minimal points and a numerical algorithm may only produce a local minimal value. How much can the variance of the importance sampling estimator be reduced if only one direction is chosen? In a number of test cases we have shown that adding additional directions will decrease the estimator variance dramatically. What conditions are required for this to be true? For $k = 1$, will adding a suitable second direction reduce the estimator variance a significant amount? If so, how do you effectively solve for the two directions best directions? As adding more directions requires additional computational work, can an approach for determining the optimal number of directions be developed? Is it better to use a small number (perhaps one) of optimal directions or a larger number (perhaps a full basis) of non-optimized directions?

**Question 3.** We consider now the case of looking for a large parameter vector $\theta$ with $k >> 1$ for a large portfolio with $N_O$ obligors. As the parameter dimension is large, we would likely be considering the factor model case and only applying importance sampling to the independent generating samples with $C = I$ and $Q = I$ to reduce the computational complexity of sampling and calculating the weight function. Assume that we run a sample Monte Carlo simulation based on the original measure with $M$ trials (say $M = 10000$) which results in $m = \alpha M$ tail samples (say $m = 100$ at the 1% confidence level). Suppose that $k >> m$. The following approach provides a simple algorithm to approximate optimal parameters $\theta^{opt}$.

Because the function $h(z)$ is an indicator function, its value is zero for all but the tail samples, for which it
has value one. Therefore we can estimate the gradient of $m_h$ with respect to $\theta$ from the initial simulation by

$$(\nabla m_h(\theta))_i \approx \frac{1}{M} \sum_{j=1}^{m} \frac{df_i(z_j, \theta_i)}{d\theta_i}$$

(3.19)

$$\approx \frac{1}{M} \sum_{j=1}^{m} \frac{1}{2} f_i(z_j, \theta_i)d_i(z_j)$$

(3.20)

where

$$d_i(z_j) = \frac{1}{1 - \theta_i - \left(q_i^T C^{-1} z_j\right)^2}. \quad \text{(3.21)}$$

The Hessian matrix can be approximated by

$$H(\theta) = \Delta_0(W) \approx D + VV^T. \quad \text{(3.22)}$$

Here $D$ is a $k \times k$ positive definite diagonal matrix with $D_{ii} = \hat{W}/(1 - \theta_i)^2$ and $V$ is an $k \times m$ matrix with $V_{ij} = W_j d_i(z_j)$. The quantity $\hat{W}$ is the average over the $m$ tail samples of the weight function $W_j$ observed for each tail sample $j$.

Step 1. Start with an initial value $\theta_0$, we want to find a direction $g_0$ such that

$$H(\theta_0)g_0 = -\nabla (m_h(\theta_0)). \quad \text{(3.23)}$$

$H(\theta_0)$ is $k$ dimensional square matrix. Solving for $g_0$ would in general be an order $k^2$ calculation, which could be time consuming as $k$ is large. However, we can calculate the inverse as the following

$$H^{-1} = D^{-1} - D^{-1}V (I + V^T D^{-1}V)^{-1} V^TD^{-1}$$

(3.24)

where

$$I + V^TD^{-1}V \quad \text{(3.25)}$$

is an $m$ dimension square matrix. Since $m$ is small, the inverse is computable, and the calculation of $g_0$ can be computed in order $mk$ calculations.

Step 2. Find the value of $c$ that minimizes

$$m_h(\theta_0 + cg_0) = \frac{1}{M} \sum_{j=1}^{m} W(z_j, \theta_0 + cg_0). \quad \text{(3.26)}$$

One can use standard one dimensional root finding methods to solve for $c_0$.

Step 3. Set $\theta_1 = \theta_0 + c_0g_0$ and repeat the iteration until $\theta_{i+1} = \theta_i + c_ig_i$ is sufficiently close to $\theta_i$.

For a given number of initial samples $M$, is it possible to improve this algorithm to produce a better estimate of the optimal parameter $\theta$? How does this approach compare with the importance sampling approach for the full covariance matrix using a small number of parameters?
4 A Simple Example

To examine the effectiveness of the importance sampling method described here, we consider here a simple example related to the first passage time of Brownian motion. Let \( Z(t) \) be Brownian motion with the properties that \( Z(0) = 0 \), \( \text{cov}(Z(t), Z(s)) = \text{min}(t, s) \), and \( Z(t + \Delta t) \sim \phi(Z(t), \Delta t) \). We define the first passage time of a barrier \( b > 0 \) as \( \tau = \text{min}(t : Z(t) \geq b) \). It is easy to show the well known result that the first passage time has the probability distribution given by \( P(\tau < T) = 2 \ast (1 - \Phi(b/\sqrt{T})) \).

We consider now the discrete version of the first passage time problem and ask, for the set of discrete times \( t_1 < t_2 < \ldots < t_N \), what is the probability that \( Z(t_i) \geq b \) for some \( 1 \leq i \leq N \)? If we take \( T = t_N \) and assume equally spaced times with \( \Delta t = T/N \) and \( t_i = i\Delta t \), then \( P(Z(t_i) \geq b) \) but converges to \( P(\tau < T) \) as \( N \to \infty \). For finite \( N \) there is no simple analytic formula for the probability of exceeding \( b \); however, for a test case example, brute force Monte Carlo simulation can be used to obtain the answer to the degree of accuracy required for testing the effectiveness of importance sampling.

The previously described importance sampling method can be applied in two ways here. We first look at the correlated set of Normal random variables \( Z = [Z(t_1), \ldots, Z(t_N)] \) with covariance matrix \( C(i,j) = \Delta t \min(i,j) \). We seek a single direction \( q \) and a scalar parameter \( -\infty < \theta < 1 \) to define a new covariance matrix \( C_\theta = C + \frac{\theta}{1-\theta} qq^T \).

Here \( q \) must satisfy \( q^T C^{-1} q = 1 \). If we define the function \( h(Z) \) as

\[
h(Z) = \begin{cases} 
1 & \text{if } \max(Z) \geq b \\
0 & \text{otherwise}
\end{cases}
\]

then

\[
P(\max(Z) \geq b) = \int h(Z)\phi(Z|0,C)dZ = \int W(Z,\theta)h(Z)\phi(Z|0,C_\theta)dZ
\]

where

\[
W(Z,\theta) = \frac{1}{\sqrt{1-\theta}} \exp\left(-\frac{\theta}{2}(q^T C^{-1} Z)^2\right).
\]

If we define \( \mu = P(\max(Z) \geq b) \) as the expectation of \( h \) under the orginal measure, because \( h^2 = h \), we have that for the orginal measure, the variance \( \sigma^2 = E((h - \mu)^2) \) is given by \( \mu - \mu^2 \). For the importance sampling measure, the variance is given by \( \sigma^2(\theta) = m_h - \mu^2 \), where

\[
m_h(\theta) = \int W^2(Z,\theta)h(Z)\phi(Z|0,C_\theta)dZ.
\]

We consider first the simplest case where \( N = 1 \) and \( T = 1 \). In this case, we have the analytic solution \( \mu = 1 - \Phi(b) \). Taking \( C = 1 \) and \( q = 1 \), it follows that \( C_\theta = 1/(1-\theta) \). It is easy to show that for this simple case

\[
m_h(\theta) = \frac{1}{\sqrt{1-\theta}} \left(1 - \Phi\left(b\sqrt{1+\theta}\right)\right).
\]
Asymptotically, the optimal $\theta$ that minimizes $m_b(\theta)$ is given by $\theta_{opt} = 1/2 + 1/2(1 - (1/b)^2)^2$ as $b \to \infty$. This is actually close to optimal for $b > 1$. With this choice of $\theta$ we plot on a log base 10 scale in Figure 4.1 the ratio of the variances of the importance sampling method over the original method for the range of barriers $1 \leq b \leq 5$. Note that the original method corresponds to $\theta = 0$. We observe that as the probability of crossing the barrier decreases to become more of a rare tail event, the power of the importance sampling method using the optimal $\theta$ increases. For example, the probability of exceeding the barrier $b = 5$ is around $2.87\times10^{-7}$. With standard Monte Carlo, it would require around one billion samples to estimate this number to 10% relative accuracy with reasonable confidence. Using the importance sampling method with the optimal $\theta$, the same accuracy can be achieved with around ten thousand samples. Asymptotically at the optimal $\theta$, the variance reduction is $0.5b\exp(-0.5b^2)$ for this simple case.

![First Passage Time Example, Single Step Variance Reduction](image)

Figure 4.1: Variance Reduction for First Passage Time Example with One Step

To evaluate the performance of the importance sampling method for a multi-variate problem, we consider the case with $N = 100$, $T = 1$ and $b = 3$. For this case, $P(\max(Z) \geq b) \approx 0.00225$. For reference, the first passage time probability for this barrier on the interval $[0,1]$ is $P(\tau < T) = 0.00270$, while for the single step case ($N = 1$), $P(Z_N \geq b) = 0.00135$. We first take the direction $q$ as the eigenvector corresponding to the largest eigenvalue, normalized so that $q^TCq = 1$. Figure 4.2 shows the variance reduction ratio as a function of $\theta$ for the barrier $b = 3$. In this case, the optimal $\theta$ is around 0.84, leading to an optimal variance reduction with this $q$ of around 90%.

This demonstrates that substantial variance reduction can be achieved using only knowledge of the covariance.
structure without considering properties of the integrand $h(Z)$. However, we should be able to achieve better results by taking $h(Z)$ into account. More precisely, we seek to minimize the weight function $W(Z)$ for those paths for which $\max(Z) > b$ where $h(Z) = 1$. For a given $\theta$ this corresponds to find a direction $q$ that tends to give large values of $q^T C^{-1} Z$ when $h(Z) = 1$, subject to $q^T C^{-1} q = 1$. With inspiration from the simple $N = 1$ case where good result were obtained by focusing only on the last point of the path, we consider the vector $q$ that gives $q^T C^{-1} Z = Z_N$. For paths that cross the barrier, it is likely that $Z_N$ will be relatively large, and thus $W(Z)$ will be small. If we define

$$v = C^{-1} q$$

then selecting $v = [00 \ldots 01]^T$ produces the desired result. The normalization constraint on $q$ corresponds to $v^T C v = 1$. It can be seen that for the covariance matrix $C$ defined for this problem and given that $T = N \Delta t = 1$, this choice of $v$ also satisfies the constraint. With this choice labeled the 'Last Point Direction’, the resulting variance reduction, relative to the standard case, is also shown in Figure 4.2. The optimal choice of $\theta$ for this direction is around 0.86, and the performance does indeed improve to a variance reduction of around 93%.

For a fixed $\theta$, on the ellipsoid defined by $v^T C v = 1$ there should be an optimal direction $v_{opt}$ (possibly not unique) where $m_h(v)$ obtains its minimum. The difficulty is that over this surface there are likely to be multiple local maxima and minima, so it may be difficult to find $v_{opt}$. However, for the Last Point Direction case we can compute the gradient of $m_h(v)$ with respect to $v$ at the optimal $\theta$ and use this gradient direction to update $v$ in hopes of improving performance. The gradient is given by

$$\nabla m_h(v) = -\theta \left( \int ZZ^T W(Z, v) h^2(Z) \phi(Z|0, C) dZ \right) v.$$
Using a single step update, starting with the Last Point Direction $v$ we can define

$$
\hat{v} = v - \nabla m_h(v) \tag{4.1}
$$

$$
v_{\text{opt}} = \frac{\hat{v}}{\sqrt{\hat{v}^T \hat{C} \hat{v}}}. \tag{4.2}
$$

Results for this direction are also shown in Figure 4.2 labeled ‘Optimal Direction’. There is in fact a further improvement in performance, with the optimal $\theta$ now at around 0.87 and the optimal variance reduction approaching 94.5%.

More generally, the constrained optimization problem can be written in a Lagrange multiplier form to optimize the unconstrained function

$$
F(v, \lambda) = m_h(v) + \lambda(v^T C v - 1)
$$

for a fixed $\theta$. Setting the gradient of this function equal to zero and substituting for $\lambda$ leads to the equation for $v$

$$
(v^T A v) C v = A v
$$

where the matrix $A$ is defined as

$$
A(v) = E(Z Z^T W(Z, v) h^2(Z))
$$

and the expectation is with respect to the original probability measure for $Z$. This suggests that an eigenvector of $C^{-1} A$ may be the optimal solution, although the dependence of $A$ on $v$ likely requires an iterative solution. This also suggests that there are likely to be many local maxima or minima solutions corresponding to the various eigenvectors. Once the optimal direction is found, the process would need to be repeated to determine the optimal $\theta$ and associated optimal direction. A further question is whether the combination of two or more directions leads to substantially better variance reduction.

An alternate approach is to consider the original probability distribution to be the independent identically distributed Normal samples used to generate the steps of the path as opposed to the correlated steps of the path itself. Again let $Z$ be the $N$ step path and $h(Z)$ be defined as above. Now, however, we generate $Z$ as

$$
Z_j = \sqrt{\Delta t} \sum_{i=1}^{j} \epsilon_j
$$

where the $\epsilon_j$ are iid standard Normal variates. We are then interested in evaluating

$$
P(\max(Z) \geq b) = \int h(Z(\epsilon)) \phi(\epsilon|0, 1) d\epsilon.
$$

In this case the original covariance matrix is the identity matrix $C = I$, and the importance sampling distribution is given by a diagonal matrix $C_\theta$ such that $C_\theta(i, i) = 1/(1 - \theta_i)$. The Hessian of the variance $m_h(\theta)$ with respect to this $\theta$ vector is positive definite, so an optimal choice of $\theta$ does exist. However, for the case of $N = 100$, computational experiments show that optimal $\theta$ is very close to zero vector (i.e., no importance sampling), and the optimal variance reduction is only around 3%. The difficulty is that the weight function $W(\theta)$ is the product of 100 very similar values, so that it is either very small (if the value is less than one) or very large (if the value is greater than one), unless the typical $\theta$ values are close to zero.
References


