Importance Sampling Applied to Value at Risk

by

Douglas Glass

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Mathematics
Courant Institute of Mathematical Sciences
New York University
January, 1999

Jonathan Goodman

1. Introduction

This thesis discusses using Monte Carlo Methods for evaluating calculations related to Value at Risk (VaR). This is a problem estimating the probability of a very unlikely event. We show that an Importance Sampling method motivated by the Theory of Large Deviations leads to significant improvements in the accuracy of the Monte Carlo Estimator. Implementation of this strategy leads to a constrained optimization problem, which may have multiple local minima. These local minima lead to new difficulties in the importance sampling.

Value at Risk is a tool for risk management in financial institutions [11]. It is about finding an amount that could be lost with probability α. Let V_t represent the value of a portfolio of (risky) assets at time t. Since V_t is a stochastic process it has a distribution at time t. The Value at Risk at time t at level α is the α^{th} quantile of the distribution of V_t. In the following equation,

$$P(V_t \leq v_\alpha) = \alpha \tag{1}$$

v_α is the VaR. For example, the 30-day 5% VaR would be $v_{.05}$ for $V(\frac{30}{365})$.

This paper focuses on the converse of the Value at Risk problem. Rather than determining the the αth quantile, we will compute α for a specified v_α.

There are a variety of ways to compute Value at Risk [11]. At the very minimum, one must be able to evaluate the portfolio at time, t. Monte Carlo simulation of the portfolio will provide an empirical distribution of V_t from which an estimator for α can be computed.

First we will give a short overview of Monte Carlo methods and Importance Sampling. Next, a portfolio model will be developed. Finally, the Importance Sampling method will be applied to several sample portfolios demonstrating its efficiency.

2. Overview of Monte Carlo Methods and Importance Sampling

For a brief overview of Monte Carlo integration, let X be a random variable in R^d with multi-variate distribution density $f(x)$. To find find a probability, p, of the event $X \in (A_\alpha \subseteq R^d)$ one must evaluate

$$p = \int_{R^d} I_{A_\alpha}(x) f(x) dx \tag{2}$$

where I_{A_α} is an indicator function defined as follows:

$$I_{A_\alpha}(x) = \begin{cases} 1 & \text{if } x \in A_\alpha \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

To integrate this function by Monte Carlo one takes advantage of the fact that

$$p = \int I_{A_\alpha}(x) f(x) dx = E_f[I_{A_\alpha}(x)] \approx \frac{1}{N} \sum_{i=1}^{N} I_{A_\alpha}(X_i) = \hat{p} \tag{4}$$

where $E_f[g(x)]$ is the mathematical expectation of $g(x)$ with x distributed with density f.

We will call $\hat{p}$ the Monte Carlo estimator and $\hat{\sigma}^2$ the variance of $I_{A_\alpha}(x)$[1]. The central limit theorem says that $\hat{p} \sim N(p, \frac{\hat{\sigma}^2}{N})$. Define the statistical error, $\epsilon = \frac{2\hat{\sigma}}{\sqrt{N}}$ Then $\hat{p} \pm \epsilon$ sets up approximately a 95% confidence interval for p. The relative error of $\hat{p}$ is $\frac{\epsilon}{\hat{p}}$. This can be thought of as the size of the statistical error relative to the estimator. For Monte Carlo calculations, this ideally should be no more than 5%.

If one can find another estimator for p with a smaller variance, then the relative error will be smaller for the same sample size, N. Conversely one can achieve a given relative error with a smaller sample. The computations will be more efficient since better numbers are obtained with improved computer time. Importance sampling can sometimes be used to find such an estimator.

Importance sampling is based on the following set of equalities:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} I_{A_\alpha}(X_i) \approx E_f[I_{A_\alpha}(x)] = \int I_{A_\alpha}(x) f(x) dx = \int I_{A_\alpha}(x) \frac{f(x)}{\tilde{f}(x)} \tilde{f}(x) dx = \tag{5}$$

$$E_{\tilde{f}}[I_{A_\alpha}(y) \frac{f(y)}{\tilde{f}(y)}] \approx \frac{1}{N} \sum_{i=1}^{N} I_{A_\alpha}(Y_i) \frac{f(Y_i)}{\tilde{f}(Y_i)} = \tilde{p}$$

where the Y_i are randomly drawn from density $\tilde{f}(y)$. The density $\tilde{f}$ is sometimes called the twisted density and $\phi(x) = \frac{f(x)}{\tilde{f}(x)}$ is called importance function. It should be noted that both f and $\tilde{f}$ need to be defined in the same domain.

To keep the estimators distinct, let $\tilde{p}$ come from the Y_i taken from $\tilde{F}$ using the importance function and $\hat{p}$ will refer to the estimator obtained using straight Monte Carlo. The central limit theorem applies to both $\hat{p}$ and $\tilde{p}$. The mean is p in either case. However, the corresponding variances should be different. Ideally, $\tilde{\sigma}^2$, the variance of $I_{A_\alpha}(x)\frac{f(x)}{\tilde{f}(x)}$ with respect to $\tilde{f}$, is much less than $\hat{\sigma}^2$. That makes $\tilde{\epsilon} = \frac{2\tilde{\sigma}}{\sqrt{N}}$ less than ϵ, creating a much tighter interval in which one has a 95% chance of finding p, the true mean. The corresponding relative error will be appropriately less.

[1] This is the variance of $I_{A_\alpha}(x)$ sampled from the density, $f(x)$, not the variance of x with respect to $f(x)$.

As an example, consider the prospect of finding $p = P(Z < -3)$ for $Z \sim N(0,1)$. In a computational experiment we generate, N random variables from the Gaussian density, $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$. The Monte Carlo estimator $\hat{p}$ was to estimate p for the event $A_\alpha = \{Z_i \mid Z_i < -3\}$. In hopes of finding an importance function that gave the estimator $\tilde{p}$ a smaller variance, we used $\tilde{f}(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y+3)^2}{2}}$, a Gaussian with mean -3 and variance 1. For the Y_i drawn from $\tilde{f}$, the set A_α was the same: $\{Y \mid Y < -3\}$. The two estimators are

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} I_{\{Z_i < -3\}}(Z_i)$$

and

$$\tilde{p} = \frac{1}{N} \sum_{i=1}^{N} I_{\{Y_i < -3\}}(Y_i) \frac{f(Y_i)}{\tilde{f}(Y_i)} = \frac{1}{N} \sum_{i=1}^{N} I_{\{Y_i < -3\}}(Y_i) e^{3Y_i + \frac{9}{2}}.$$

Trials were done for different values of N. However, the same sequences of random numbers were used for both estimators. For $i = 1, \ldots, N$, the random numbers U_{2i-1} and U_{2i} were sampled independently from a uniform distribution. By the Box-Muller method [12], the pair $Z_{2i-1} = \sqrt{-2\log(U_{2i-1})}\cos(2\pi U_{2i})$ and $Z_{2i} = \sqrt{-2\log(U_{2i})}\sin(2\pi U_{2i-1})$ are independent Gaussian random variables. The Z_i are used for the standard normal X_i. Then the $Y_i = X_i - 3$ were created and were independent with common distribution $N(-3,1)$. A new sequence of random numbers were used for each value of N. The results were as follows:

N	$\hat{p}$	$\epsilon = \frac{2\hat{\sigma}}{\sqrt{N}}$	rel. error ($\frac{\epsilon}{\text{est.}}$)	Samples
1000	0	0	N/A	0
10000	0.0016	0.00079936	49.9%	16
50000	0.00124	0.000314765	25.4%	62
100000	0.00133	0.000230498	17.3%	133
1000000	0.00132	7.26156e-05	5.5%	1320

TABLE 1. Regular Monte Carlo - estimator $\hat{p}$

N	$\tilde{p}$	$\epsilon = \frac{2\tilde{\sigma}}{\sqrt{N}}$	rel. error ($\frac{\epsilon}{\text{est.}}$)	Samples
1000	0.00134758	0.000153425	11.39%	510
10000	0.00133979	4.96594e-05	3.7%	4996
50000	0.00137612	2.25461e-05	1.6%	24978
100000	0.00133566	1.56792e-05	1.2%	49256
1000000	0.00134561	4.96111e-06	0.36%	499312

TABLE 2. Importance Sampling - estimator $\tilde{p}$

Clearly, using samples from $\tilde{f}$ and the importance function ϕ to weigh the contribution of each random variable is much more efficient than using random variables drawn from f. The number of samples needed with the importance sampling method is reduced by a factor of over 100. Figure 1 shows that one is far more likely to find a sample from the twisted distribution, $\tilde{f}$ than from the original. Roughly half will be in set A_α. And since there is far less fluctuation in the weight assigned to each each sample, the variance is much lower.

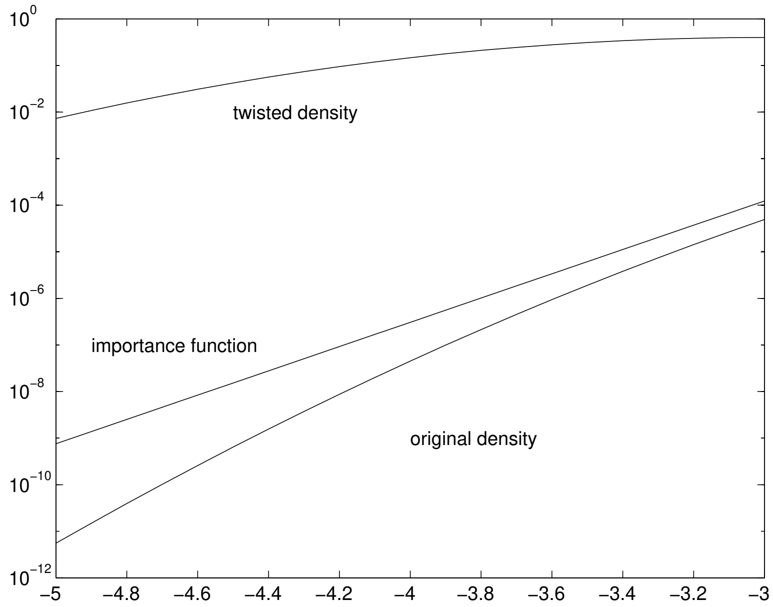

FIGURE 1. Log scale of the two densities and the importance function

Looking at the Figure 1, one wonders if it wouldn't be even more efficient to use a twisted distribution that more closely resembles the tail of the original or perhaps adjusting the variance parameter of $\tilde{f}$ to get samples that are assigned greater value (i.e. closer to -3). In fact, using an exponential distribution with the proper parameters is even more efficient than using a Gaussian with different parameters. In many cases, Large Deviation Theory can be used to find more optimal twisted distributions for a variety of problems [3] [17] .

However, stock prices are often modeled with a log-normal distribution. A log-normal distribution has no moment generating function [4]. Finding the moment generating function is the first step in the Large Deviations technique. But as we will show, this method of mean shifting will give us a significant reduction in the variance of the Monte Carlo estimator.

The next section develops a model for stock prices with correlated returns and a multi-variate log-normal joint distribution.

3. Portfolio, Stock Price Model and Joint Distribution

In order to use the importance sampling method described in the last section, one must have a density from which to draw samples. This section develops the process that we use to model stock prices and defines the function, V_t that represents the value of the portfolio at time, t.

A portfolio, Π, contains securities whose values are dependent on d underlying random (risky) variables. These are denoted $S_i, i \in \{1, \ldots, d\}$. The vector $S(t) = [S_1(t), \ldots, S_d(t)]^T$ is the state of the values of the underlying variables at time, t. The random variables, S_i, will be assumed to represent stock prices. The assets can be the underlying variables or derivative securities whose value is based on the underlying variables, such as options or futures. Since path-dependent options do not have a unique price for a given state-space, $S(T)$, they will not be considered in this paper.

The value of the portfolio is a function $V(S(t)) : R^d \to R^1$. This may be written as $V_t(S)$ or $V(S(t))$, depending on notational requirements. If V or S appear without a t, the time dependency will be implied. For the purposes of this paper, the portfolio will consist of B_0 of risk free money (bonds, money market accounts, etc.) acquired at time, $t = 0$, accruing continuously compounded interest at rate, r, d stocks and European puts and calls on those stocks. The value of the portfolio is

$$V(S(t)) = B_0 e^{rt} + \sum_{i=1}^{d} s_i S_i(t) + c_i C(S_i(t), \sigma_i, r, K_{c,i}, T_{c,i}) + p_i P(S_i(t), \sigma_i, r, K_{p,i}, T_{p,i}). \tag{6}$$

Here, s_i, c_i, and p_i are the amount of shares, calls, and puts on the underlying asset, S_i. $C(\cdot)$ and $P(\cdot)$ are the Black-Scholes prices of calls and puts. The parameters σ_i, r, $K_{(\cdot),i}$, $T_{(\cdot),i}$ are the volatility, risk-free rate, strike price, and expiration time for the options on S_i. Negative values of B_0, s_i, c_i, and p_i correspond to borrowing money or short-selling the assets. For the purposes of this paper, S will typically mean a collection of stocks and will be referred to as such. However, one could easily substitute any underlying variable that is assumed to follow the same geometric Brownian motion process described below, such as foreign exchange rates.

In one dimension, the equations are as follows [1]. Itô's theorem states that for a stochastic process $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$ where W_t is a Weiner process, and a function $U(X_t, t)$:

$$dU = \left(\frac{\partial U}{\partial t} + \frac{\partial U}{\partial X} a(X, t) + \frac{1}{2} \frac{\partial^2 U}{\partial X^2} b(X, t)^2 \right) dt + \frac{\partial U}{\partial X} dW. \tag{7}$$

The random variables, S_i are each assumed to follow the following Itô process:

$$dS_i = S_i \mu_i dt + S_i \sigma_i dW_t. \tag{8}$$

Here, μ_i is the stock's expected growth rate, and σ_i is the volatility.

This SDE in equation 8 can be solved using Itô's theorem (equation 7) to transform it into an ODE. Let $dX = dS_i$, $a(S_i, t) = S_i \mu_i$, $b(S_i, t) = S_i \sigma_i$, and $U(X, t) = \log(S_i)$. Then (using short-hand notation for the partials) $U_t = 0$, $U_S = \frac{1}{S_i}$, and $U_{SS} = -\frac{1}{S_i^2}$. Using Itô's theorem:

$$d(\log(S_i)) = (\mu_i - \frac{1}{2}\sigma_i^2)dt + \sigma_i dW_t.$$

Integrating both sides with respect to t yields:

$$\log(S_i(t) = (\mu_i - \frac{1}{2}\sigma_i^2)t + \sigma_i W_t + C.$$

Let $t = 0$ to solve for C (recalling that $W_0 = 0$):

$$\log(S_i(0)) = C.$$

Taking exponentials of both sides,

$$S_i(t) = S_i(0)e^{(\mu_i - \frac{1}{2}\sigma_i^2)t + \sigma_i W_t}. \tag{9}$$

This process is called Geometric Brownian Motion. With this solution, one can obtain the density for $S_i(t)$.

Let G be a standard normal random variable, i.e. $N(0,1)$. Then $X = bG + a$ has distribution $N(a, b^2)$. The random variable $Y = e^{bX+a}$ has a log-normal distribution with parameters a and b^2. This is denoted $Y \sim \Lambda(a, b^2)$ [4]. It has the probability density function

$$f(x) = \frac{1}{xb\sqrt{2\pi}}\, e^{-\frac{(\log x - a)^2}{2b^2}}. \tag{10}$$

Applying it to the above equation for $S_i(t)$ We find that $S_i(t) \sim \Lambda\Big((\mu_i - \frac{1}{2})t, \sigma_i^2 t\Big)$.

Since we are interested in modeling a portfolio, we must consider the multi-dimensional case. Let, G be a d-dimensional column vector of independent $N(0,1)$ random variables. One must proceed as before to use G to create $X \sim N(a, B)$, where a is the mean vector and B is the covariance matrix with elements $(B)_{ij} = \rho_{ij} b_i b_j$ where ρ_{ij}, b_i, and b_j are the corresponding covariance and variances of X_i and X_j. First, a "square root" of the covariance matrix, B, is needed to act as the standard deviation, b, in the previous example. This is the Cholesky decomposition. If any matrix, B is positive definite, then a lower triangular matrix, L, can be found such that $B = LL^T$. Since the covariance matrix of a multi-variate normal distribution is always positive definite, this is possible. Then $X = LG + a \sim N(a, B)$.

If $X \sim N(a, B)$ and Y is a random vector such that $y_i = e^{x_i}$, where x_i and y_i are the i^{th} element of X and Y, then Y has a multivariate log-normal distribution, $\Lambda(a, B)$ [10]. The density function is analogous to the scalar case in equation 10:

$$f(Y) = f(y_1, \ldots, y_d) = \frac{1}{y_1 \ldots y_i \sqrt{\det(B)}(2\pi)^{\frac{d}{2}}}\ \exp\Big[\frac{1}{2}(Y-a)^T \Sigma^{-1}(Y-a)\Big]. \tag{11}$$

The next step is constructing a d-dimensional diffusion process that reasonably simulates correlated stock prices. Then we will solve the multi-variate equation using Itô's Theorem to determine if it is log-normally distributed component-wise. This would imply portfolio has a multi-variate log-normal distribution. If it does, we would then determine the proper parameters.

Clearly, if the S_i were all independent, the collection of stocks would have a $\Lambda(\mu, I)$ distribution where μ is the vector of expected growth rates for the assets and I is the identity matrix. But instead of using d independent Weiner processes, W_i, to represent the returns, the model requires correlated underlying processes.

Let $C_{d\times d}$ be the correlation matrix with elements, $(C)_{ij} = \rho_{ij}$, $V_{d\times d}$ be a diagonal matrix whose elements are $(V)_{ii} = \sigma_i$, and $\hat{S}_{d\times d}$ be a diagonal matrix whose elements are $\hat{S}_{ii} = S_i$, the i^{th} entry in S. Since C is symmetric and positive definite, by virtue of being a correlation matrix, it has a Cholesky decomposition, $C = LL^T$ [7]. Then the covariance matrix $\Sigma = VLL^TV = (VL)(VL)^T$. However, if one wants to use the previous, one dimensional process as a guide to the general case, the vector of Brownian Motions, $Z = LW$ will give a vector of $N(0,1)$ random variables with correlation structure, C. Component-wise $Z_i = \sum_{j=1}^{i} l_{ij}W_j$. Keeping $Var[Z_i] = 1$ allows a diffusion process to take the form

$$dS = \hat{S}\mu dt + \hat{S}VLdW,$$

with the individual stocks following,

$$dS_i = S_i\mu_i dt + S_i\sigma_i dZ_i = S_i\mu_i dt + S_i\sigma_i \sum_{j=1}^{i} l_{ij}dW_j$$

As with the one dimensional model, the factor $\sigma_i dZ_i(t)$ makes the returns have a $N(0, \sigma^2 t)$ distribution. Individually, the stocks behave as usual, but as a whole, the trends will be apparent, since the Z_i are correlated according to C (with covariance Σ).

To solve the system of SDEs and determine the parameter for Λ, one must use the multi-dimensional Itô theorem. Let $U(X,t) \in R^k$, $X, A(X,t) \in R^d$, $B(X,t) \in R^{d\times m}$, and $W_t \in R^m$. Then for a d-dimensional diffusion driven by m Weiner processes, $dX = Adt + BdW$,

$$dU = \Big(U_t + U_X A + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} U_{X_iX_j}(BB^T)_{ij}\Big)dt + U_X BdW. \tag{12}$$

Here, U_X is a $k\times d$ matrix with elements $(U_X)_{ij} = \frac{\partial U_i}{\partial X_j}$ where U_i is the i^{th} element of U and X_j is the j^{th} element of X. Also, $U_{X_iX_j}$ is a k-dimensional column vector such that $U_{X_iX_j} = \big(\frac{\partial^2 U_1}{\partial X_i \partial X_j}, \frac{\partial^2 U_2}{\partial X_i \partial X_j} \dots, \frac{\partial^2 U_k}{\partial X_i \partial X_j}\big)^T$ and U_t is a k-dimensional column vector whose entries $(U_t)_i = \frac{\partial U_i}{\partial t}$.

Now we solve the multi-dimensional diffusion equation (12) for the vector, S, with $k = d = m$. Let

$$A = \hat{S}\mu = (S_1\mu_1, \dots, S_d\mu_d)^T,$$

$$B = \hat{S}VL$$

so that

$$BB^T = \hat{S}\Sigma\hat{S}, \qquad (BB^T)_{ij} = S_iS_j\sigma_i\sigma_j\rho_{ij}$$

and

$$U = \log(S) = (\log(S_1), \dots, \log(S_d))^T$$

and apply Itô's general theorem (equation 12). Then the vector

$$U_t = 0,$$

with

$$(U_X)_{ii} = \begin{cases} \frac{1}{S_i} & i = j \\ 0 & i \neq j \end{cases},$$

and

$$U_{X_iX_j} = \begin{cases} (0, \ldots, -\frac{1}{S_i^2}, \ldots, 0)^T & i = j \\ 0 & i \neq j \end{cases}.$$

Substituting the values into the full equation yields:

$$\begin{pmatrix} d(\log(S_1)) \\ \vdots \\ d(\log(S_d)) \end{pmatrix} = \left[\begin{pmatrix} \frac{1}{S_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{S_d} \end{pmatrix} \begin{pmatrix} \mu_1 S_1 \\ \vdots \\ \mu_d S_d \end{pmatrix} + \frac{1}{2} \left[\begin{pmatrix} -\frac{1}{S_1^2} \\ \vdots \\ 0 \end{pmatrix} S_1^2 \sigma_1^2 + \ldots + \begin{pmatrix} 0 \\ \vdots \\ -\frac{1}{S_d^2} \end{pmatrix} S_d^2 \sigma_d^2 \right] \right] dt$$

$$+ \begin{pmatrix} \frac{1}{S_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{S_d} \end{pmatrix} \begin{pmatrix} S_1\sigma_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & S_d\sigma_d \end{pmatrix} \begin{pmatrix} l_{11} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ l_{d1} & \ldots & l_{dd} \end{pmatrix} \begin{pmatrix} dW_1 \\ \vdots \\ dW_d \end{pmatrix}$$

$$= \left[\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} - \frac{1}{2} \begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_d^2 \end{pmatrix} \right] dt + \begin{pmatrix} \sigma_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_d \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{1} l_{1j} dW_j \\ \vdots \\ \sum_{j=1}^{d} l_{dj} dW_j \end{pmatrix}$$

$$= \left[\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} - \frac{1}{2} \begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_d^2 \end{pmatrix} \right] dt + \begin{pmatrix} \sigma_1 dZ_j \\ \vdots \\ \sigma_d dZ_j \end{pmatrix}.$$

Looking at the components,

$$d(\log(S_i)) = (\mu_i - \frac{1}{2}\sigma_i^2)dt + \sigma_i dZ_i.$$

This has already been solved for dW_i. Since taken individually, Z_i and W_i are indistinguishable, the solution,

$$S_i(t) = S_i(0) \exp\left((\mu_i - \frac{1}{2}\sigma_i^i)t + \sigma_i Z_i(t) \right)$$

still holds.

For a fixed, t, the process $Z_i(t)$ is a random variable with distribution $N(0,t)$. Since, $S_i(t)$ is the exponential of a $N\left((\mu_i - \frac{1}{2}\sigma_i^2)t, \sigma_i^2 t\right)$ random variable for $i = 1, \ldots, d$, of which Σ is the correlation (and covariance) matrix, $S \sim \Lambda\left(\log S_0 + \left(\mu - \frac{1}{2}\sigma^2\right)t, \Sigma\right)$, where σ^2 is the vector whose elements are the variances of their correspondent stock.

It should be noted that for $X \sim \Lambda(a, B)$, the moments are not as direct as with the $N(a, B)$ case. $E[X_i] = \exp\left(a_i + \frac{1}{2}b_i^2\right)$ and $Var[X_i] = \exp(2a_i + b_i)\left(e^{b_i} - 1\right)$. The mode is $\exp\left(a_i - \frac{1}{2}b_i^2\right)$. So with the stock model having log-normal distribution with parameters $a_i = \log S_0 + \left(\mu_i - \frac{1}{2}\sigma_i^2\right) t$ and $b_i^2 = \sigma_i^2 t$, we have $E[S_i(t)] = S_i(0)e^{\mu_i t}$, which is what they would be valued at if they were risk-free.

It is a bit reassuring that the parameters for an individual stock do not have a dependency on other stocks. Otherwise we couldn't evaluate a single asset without knowing the state of the whole market. This model of course assumes that one can obtain the correlation and volatility matrices for each of the d stocks. However, if one were to use a different model, this would not necessarily be the case. For example, if we created a model that uses an index such as the Dow Jones as the driving random process of the portfolio, another application of Itô's Theorem would be needed to see what the stock process looks like and whether or not it has a well-known distribution. It is required that $S(t)$ have an explicitly defined density to employ the importance sampling explored in the next section.

4. Importance Sampling on Portfolios

Now that it has been established that $S(t) \sim \Lambda\left((\log S(0) + \mu - \frac{1}{2}\sigma^2)t, \Sigma t\right)$, a multi-dimensional log-normal distribution, it is time to examine the set $A_\alpha = \{S \,|\, V(S) \leq v_\alpha\}$ where v_α is a prescribed lower tolerance level for the value of the portfolio at time t. Importance sampling will be used to find a more efficient estimator for $p = P\left(S(t) \in A_\alpha\right) = P\left(V(S(t)) \leq v_\alpha\right) = \alpha$.

To motivate the specific method, we will work through several examples, applying each phase of the method as we progress. The first is a simple portfolio, Π_1, a linear combination of two assets. This allows us to visualize the problem in two dimensions. The second portfolio, Π_2, has a value function dependent on eight stocks. The portfolio consists of an array of both the stocks and European puts and calls whose values are contingent on the underlying assets. This will show that the method is still effective in higher dimensions with non-linear value functions. The first two portfolios are not too different from the Monte Carlo example in Section 2. However, the final portfolio, Π_3 is different. While it is only dependent on one stock, there are two values of S where $V(S) = v_\alpha$. So two twisted densities must be used to use importance sampling, one at each value of S. This demonstrates that method's effectiveness in a more general sense.

First we will consider Π_1, the two stock portfolio with a linear value function. Then we will examine the results for a larger, eight-stock portfolio, Π_2, with a non-linear value function. Both of these portfolios and corresponding value functions will have one minimum rate point. A third portfolio, Π_3, will have one stock, but two minimum rate points. The importance sampling method will then be expanded and generalized to efficiently estimate probabilities for the multiple-minima scenario. The definition of Minimum Rate Point (MRP) will be introduced as the discussion continues.

For the purposes of visualization, we first consider a simple portfolio, Π_1. It contains 150 shares of one stock, S_1, and 100 of another, S_2. The initial values, $S_1(0) = 18$ and $S_2(0) = 24$. The expected growth rates, μ_1 and μ_2 are .09 and .12 respectively[2]. The volatilities are $\sigma_1 = .2$ and $\sigma_2 = .18$ with correlation, $\rho_{12} = .25$. The time frame in question will be 120 days, $t = .3288$. Then

$$\mu = \begin{pmatrix} \log 18 \\ \log 24 \end{pmatrix} + \begin{pmatrix} .09 - \frac{1}{2}.2^2 \\ .12 - \frac{1}{2}.18^2 \end{pmatrix} .3288,$$

$$\Sigma = \begin{pmatrix} .2^2 & (.2)(.18)(.25) \\ (.18)(.2)(.25) & .18^2 \end{pmatrix} .3288,$$

and

$$V(S(.3288)) = 150S_1(3288) + 100S_2(.3288).$$

[2]It should be noted that for these calculations the risk free rate, r, is used only with option pricing. The future values of the stocks should be modeled using μ. The parameter μ goes away in risk-neutral valuation of options, but we are not trying to create a martingale, but model the future behavior of a portfolio. This is true for other "projective" models such as Value at Risk, where one is interested in the future state of a portfolio, not the present value.

The initial value, $V_0 = V(S(0)) = \$5100$. If the stocks grew exponentially at their expected rates, $V(S(.3288)) = \$5298.92$. In this example, let v_α be \$4300. Figure 2 shows the level sets of the density, f of Π_1. The line represents $V(S(.3288)) = v_\alpha$ and the are below is $A_\alpha = \{V(S(.3288)) \mid V(S(.3288)) \leq v_\alpha\}$.

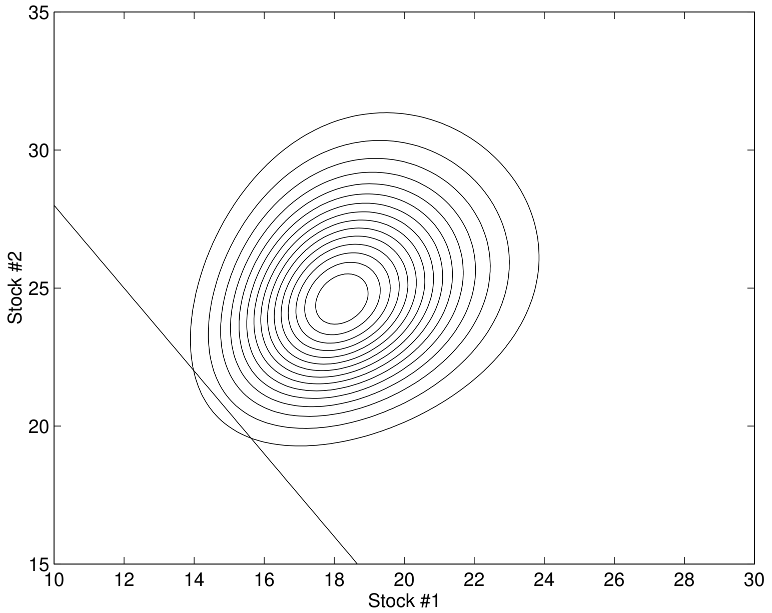

FIGURE 2. Joint Density of S_1 and S_2 at time, $t = .3288$

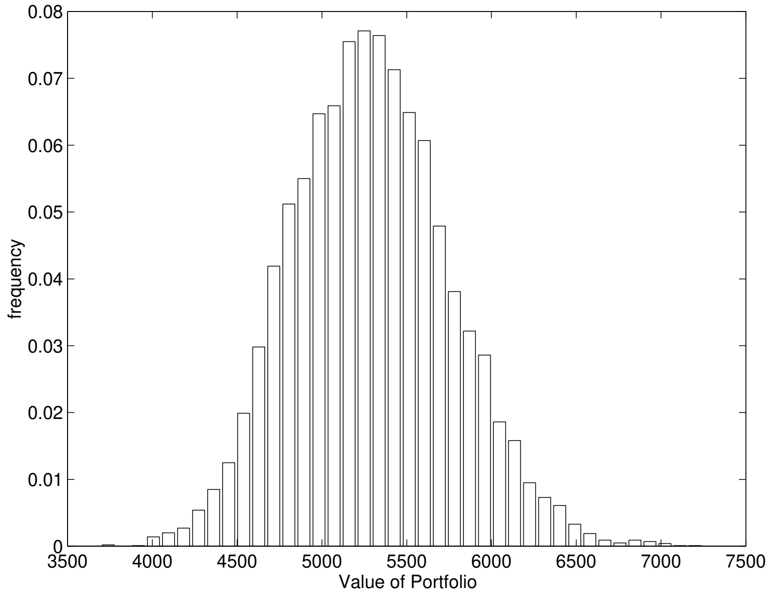

FIGURE 3. Empirical Density of $V(S(.3288))$ for portfolio Π_1

Figure 4 shows the contour plot of the density, f, along the curve $V(S) = v_\alpha$. This reveals that the majority of the mass of $f(S_1, S_2)$ is located near one point. This is due to the rapid decay of f. This is the main point of the theory of Large Deviations [3]. That is, rare events happen in a predictable manner. To find this point, one must maximize the density $f(S)$ over the set A_α.

In many cases, this property leads to a twisted density for importance sampling [3] [17]. But, the log-normal distribution doesn't decay rapidly enough to support a large deviation-based importance sampling method since $\int e^{t \cdot x} f(x)dx$ is divergent. However, the majority of the mass of A_α still collects at the the point which is the maximum of $f(S)$ subject the constraint $V(S) = v_\alpha$.

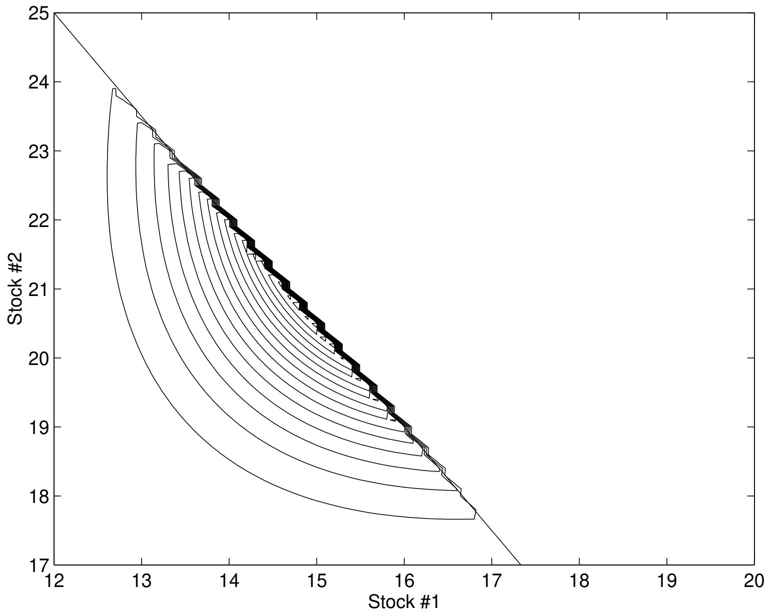

FIGURE 4. Close up of f near $V \leq v_\alpha$

In the case where $V(S)$ is a linear function, solution to the constrained optimization problem, $S_0 = \min -f(S)$ subject to $V(S) = v_\alpha$ is unique[3]. We call this point, S_m, the minimum rate point. This must be found numerically and is usually a non-trivial computational problem. In the case of a non-linear value function, such as portfolio with options, there may be more than one minimum rate point. That is, there are more than one local minima in the constrained optimization problem. If there is only one, it is sometimes referred to as a dominating point.

Finding the minimum rate points requires the implementation of a non-linear constrained optimization algorithm. There are a variety of methods that can be used, some more efficient or more effective than others. For this project the exterior penalty method was used. It is a fairly general method that is not too difficult to implement, but does need some fine-tuning to make it broad and robust. Its main advantage is that it works well with non-linear constraints. The details of this method can be found in Appendix A.

[3]Minimizing -$f(S)$ is the same as maximizing $f(S)$. Finding a maximum is sometimes more difficult than finding a minimum, so we invert the problem and seek the minima.

Returning to portfolio Π_1, the minimum rate point, S_m, is the only minimum rate point (and as such is a dominating point). This is the key to this method of importance sampling. As shown in Figure 4, if $V(S)$ is be less than v_α, it has a far greater probability of being near S_m than away from it. Now, one must construct a titled distribution, $\tilde{F}$, that covers A_α and has most of its mass clustered near x_m.

There is no "correct" way to do this. Some methods are more efficient than others. Sadowsky [17] and Bucklew [3] describe a way that uses large deviations, Wagner's [19] approach makes use of the transition density of a diffusion, and Newton [14] uses sophisticated functional analysis. The heuristic method we will use is to change the parameters of $f(S)$ so the mean is at S_m.

For estimating the probability of a loss in a portfolio, instead of using $\Lambda(a, \Sigma)$ random samples, a new $\tilde{a}$ is needed so that the tilted distribution is $\Lambda(\tilde{a}, \Sigma)$ and has mean S_m. This new $\tilde{a}$ can be found component-wise by solving the following equation for $\tilde{a}_i$:

$$\exp\left(\tilde{a}_i + \frac{1}{2}\sigma_i^2\right) = (S_m)_i \tag{13}$$

which is

$$\tilde{a}_i = \log(S_m)_i - \frac{1}{2}\sigma_i^2.$$

This tilted distribution $\tilde{F}$ has the majority of its mass near S_m. This is apparent in Figure 4.

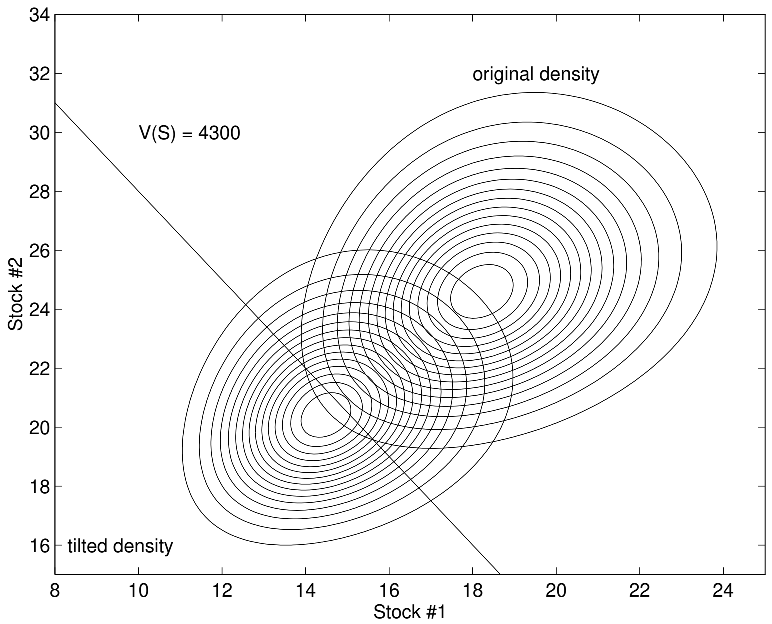

FIGURE 5. Original density, f, and tilted density, $\tilde{f}$.

Using $\tilde{f}$ and the importance sampling method described above, a much better estimate for $p = P(V(S) \leq v_\alpha)$ can be obtained. Taking N samples from $\tilde{F}$, the importance function $\phi(S) = \frac{f(S)}{\tilde{f}(S)}$ weighs them appropriately to give an estimator, $\tilde{p}$, that will converge to p at the rate $O(n^{-\frac{1}{2}})$. While this is true for standard Monte Carlo as well, the variance of the estimator, $\tilde{\sigma}^2$, is much smaller with importance sampling. This is because the samples that are used in the computations are much more frequent and closer in value; all key aspects of having small variance (and corresponding small error bars).

Table 3 shows the results of the ordinary Monte Carlo estimation. Table 4 is the results of the same data (i.e. the same sequence of random numbers) only with Importance Sampling.

N	$\hat{p}$	$\epsilon = \frac{2 \hat{sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	$\hat{\sigma}^2$
10,000	.0075	.0017	23.0%	75	.00744
50,000	.00908	.0008484	9.34%	454	.00899
100,000	.00901	.0005976	6.63%	901	.00892
500,000	.00883	.0002645	3.00%	4413	.00874

TABLE 3. Regular Monte Carlo for Π_1 with $v_\alpha = 4300$

N	$\tilde{p}$	$\epsilon = \frac{2\tilde{\sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	$\tilde{\sigma}^2$
10,000	.008907	.000292	3.28%	5126	.000213
50,000	.008989	.000132	1.47%	26013	.000218
100,000	.008932	9.292e-05	1.04%	51831	.000216
500,000	.008939	4.156e-05	0.45%	258740	.00216

TABLE 4. Importance Sampling for Π_1 with $v_\alpha = 4300$

This example shows just how effective this method of Importance Sampling can be. It took 50 times as many samples using regular Monte Carlo to get an estimator with a better relative error than the one obtained with Importance Sampling.

For an even more extreme value for v_α the results are far more apparent. Table 5 shows what happens if we set v_α to 3300. For $N = 10,000$ the relative error of $\tilde{p}$ was only 6% using Importance Sampling. As with any Monte Carlo calculation, the error decreases as N increases. There is no table for the regular Monte Carlo results because there wasn't a single sample of S for which $V(S) \leq 3300$ using the same data.

Events that rare are normally out of the scope of Value at Risk. In fact, many users of VaR like to have their portfolios exceed their VaR levels α% of the time to validate their computations. Having a 2.79e-08 probability for an event one would actually like to see happen on a periodic basis is counter-productive.

The probabilities associated with the first portfolio, Π_1, could be been calculated analytically since the value function, $V(S)$ is linear. The next example, Π_2 shows how this works for a very non-linear portfolio requiring a more difficult constrained optimization. Furthermore, the dimension is increased from $d = 2$ to $d = 8$. The error bars (confidence intervals) for the estimators, $\tilde{p}$ and $\hat{p}$ should still be $O(n^{-\frac{1}{2}})$ since it is still Monte Carlo. But the higher dimension means each sample takes longer to compute. Here, the added efficiency of the Importance Sampling method is much more noticeable, especially as one waits for the results. If a sample of size 10,000 (or even 5,000 or 1,000) gives a small relative error, it is far more efficient when

N	$\tilde{p}$	$\epsilon = \frac{2\tilde{\sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	$\tilde{\sigma}^2$
10,000	2.78e-08	1.688e-09	6.09%	5118	7.130e-15
50,000	2.73e-08	7.001e-10	2.57%	25859	6.129e-15
100,000	2.81e-08	4.742e-10	1.69%	51799	5.622e-15
500,000	2.79e-08	2.153e-10	0.77%	258641	5.7946e-15

TABLE 5. Importance Sampling for Π_1 with $v_\alpha = 3300$

one takes computational resources into consideration. Even the added time for the constrained optimization is far less than the time needed to compute an accurate estimator.

The new portfolio, Π_2, consists of 8 stocks with correlated returns and a variety of European puts and calls (with long and short positions). The risk free interest rate[4] will be $r = .07$. The time frame will be 30 days, $t = .0822$. The correlation matrix, C, whose elements are $(C)_{ij} = \rho_{ij}$ is

$$C = \begin{pmatrix} 1.0000 & 0.0497 & 0.1579 & 0.0648 & 0.0744 & 0.0498 & 0.0507 & 0.0583 \\ 0.0497 & 1.0000 & -0.0843 & -0.1134 & -0.1916 & -0.4140 & 0.4857 & -0.2857 \\ 0.1579 & -0.0843 & 1.0000 & 0.1474 & 0.4641 & -0.0192 & -0.0889 & 0.5782 \\ 0.0648 & -0.1134 & 0.1474 & 1.0000 & -0.2782 & 0.3582 & -0.3612 & 0.0268 \\ 0.0744 & -0.1916 & 0.4641 & -0.2782 & 1.0000 & -0.1920 & -0.0101 & 0.5875 \\ 0.0498 & -0.4140 & -0.0192 & 0.3582 & -0.1920 & 1.0000 & -0.0715 & -0.0865 \\ 0.0507 & 0.4857 & -0.0889 & -0.3612 & -0.0101 & -0.0715 & 1.0000 & -0.2561 \\ 0.0583 & -0.2857 & 0.5782 & 0.0268 & 0.5875 & -0.0865 & -0.2561 & 1.0000 \end{pmatrix}.$$

The parameters and positions of each asset can be seen in Table 6 and Table 7, respectively.

i	1	2	3	4	5	6	7	8
$S_i(0)$ (initial price)	35	45	10	32	70	30	48	21
μ_i (growth rate)	.15	.09	.12	.08	.04	.1	.085	.09
σ_i (volatility)	.2	.23	.3	.2	.14	.11	.16	.21

TABLE 6. Parameters of Portfolio Π_2

This sample portfolio was created to have a distinctly non-normal distribution of values at the terminal time[5]. The empirical distribution of $V(S(T)) = V(S(.0822))$ can be seen in Figure 6. Π_2 has no risk-less bonds. Since there are short positions, it would normally be assumed that the cash from the sales would have been invested. But since this paper focuses on random events and not trading strategies, the wisdom of the portfolio should be ignored.

[4]Used for computing the prices of the portfolios' options.

[5]The correlation matrix was created with the aid of semi-random (some numbers were modified) matrix R. Then RR^T was guaranteed to be symmetric and positive definite with all values between -1 and 1, yielding a suitable correlation matrix.

i	1	2	3	4	5	6	7	8
Shares	-30	13	100	21	100	300	230	49
p_i (puts)	0	0	200	0	-100	0	0	0
K_{p_i} (strike price)	–	–	12	–	68	–	–	–
T_{p_i} (expiration time)	–	–	0.9	–	1.0	–	–	–
c_i (calls)	0	-400	0	10	0	0	-300	0
K_{c_i} (strike price)	–	49	–	30	–	–	50	–
T_{c_i} (expiration time)	–	.8	–	.52	–	–	.6	–

TABLE 7. Positions of Portfolio Π_2

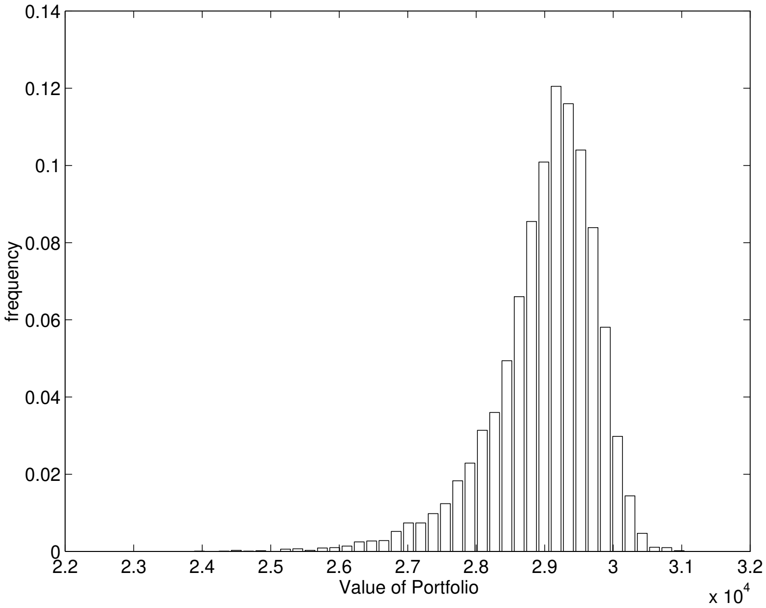

FIGURE 6. Empirical Density for $V_{\Pi_2}(S(.0822))$.

The initial value is $V_{\Pi_2}(S(0)) = 29581.30$. The lower limit is $v_\alpha = 26500$. The probability of being at or below this level 30 days into the future is $p = P\left(V\left(S(.0822)\right) \leq 26500\right)$. This was estimated with $\hat{p}$ and $\tilde{p}$ using, respectively, regular Monte Carlo and Importance Sampling as before. The sole minimum rate point in for this value of v_α is $S_m \approx (35, 53, 10, 31.5, 69, 29, 51, 20)^T$ with $f(S_m) \approx$ 1.2e-06. The results of the Monte Carlo calculations were as follows.

Looking at the results for Π_2 with $v_\alpha = 26500$, it is clear that a sample of $N = 5,000$ with importance sampling is better than a sample of $N = 50,000$ with regular Monte Carlo sampling.

Now that we've seen how effective Importance Sampling can be for a single Minimum Rate Point in multiple dimensions, we will develop the general case for multiple MRPs.

N	$\hat{p}$	$\epsilon = \frac{2\hat{\sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	$\hat{\sigma}^2$
1,000	.009	.005973	66.4%	9	.008919
5,000	.008	.002520	31.5%	40	.007936
10,000	.0094	.001930	20.5%	94	.009316
30,000	.009567	.001124	11.7%	286	.009442
50,000	.009420	.000864	9.17%	471	.009931

TABLE 8. Regular Monte Carlo for Π_2 with $v_\alpha = 26500$

N	$\tilde{p}$	$\epsilon = \frac{2\tilde{\sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	$\tilde{\sigma}^2$
1,000	.009404	.00131	13.9%	492	.0004281
5,000	.008989	.000576	6.41%	2588	.0004154
10,000	.009244	.000322	3.49%	5074	.0002599
30,000	.009551	.000242	2.54%	15207	.0004408
50,000	.00937	.000170	1.81%	25297	.0003612

TABLE 9. Importance Sampling for Π_3 with $v_\alpha = 26500$

With one MRP, the density, f was shifted so that it's mean corresponded with x_m. Let $x_{m_1}, \ldots, x_{m_M}$ be a set of M Minimum Rate Points and N sample size for a Monte Carlo estimator. To compute an accurate estimator, $\tilde{p}$, we must perform Importance Sampling near each of the x_{m_j}.

We first divide the N random samples into M parts so that

$$N = N_1 + \cdots + N_M. \tag{14}$$

To make the notation simpler, take the set of n_k where $n_j = \sum_{k=1}^{j} N_k$. This enumerates the samples into the following sequence[6]:

$$n_0 = 0, 1, \ldots, n_1 = N_1, \ldots, n_2 = N_1 + N_2, \ldots, n_j = \sum_{k=1}^{j} N_k, \ldots, n_M = N. \tag{15}$$

Clearly the zeroth sample doesn't exist but n_0 is needed for notation used below in equation 17 .

Recall that the set A_α is $\{x \mid V(x) < v_\alpha\}$, the portfolio states that have a value less than v_α. For the general case, A_α is subdivided into M pairwise disjoint subsets $A_{\alpha_1}, \ldots, A_{\alpha_M}$, with each A_{α_i} corresponding to an x_{m_j}. The subsets, A_{α_j} contain all points in set A_α that are nearest to x_{m_j}. That is, $A_{\alpha_j} = \{x \text{ s.t. } |x_{m_j} - x| < |x_{m_k} - x| \,\forall\, k \neq j\}$. The indicator function, I_{A_α}, becomes a set of functions defined as:

$$I_{A_{\alpha_j}} = \begin{cases} 1 & \text{if } x \in A_{\alpha_j} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

[6]For example, if we partition 100 samples into $M = 3$ parts with $N_1 = 60, N_2 = 30$, and $N_3 = 10$, we have $n_0 = 0, n_1 = 60, n_2 = 90$, and $n_3 = 100$.

Then equation 5 then becomes:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} I_{A_\alpha}(X_i) \approx E_f\left[I_{A_\alpha}(x)\right] = \int I_{A_\alpha}(x) f(x)dx = \sum_{j=1}^{M}\int I_{A_{\alpha_j}} f(x)dx \tag{17}$$

$$= \sum_{j=1}^{M}\int \left(I_{A_{\alpha_j}}\frac{f(x)}{\tilde{f}_j(x)}\right)\tilde{f}_j(x)dx = \sum_{j=1}^{M} E_{\tilde{f}_j}\left[I_{A_{\alpha_j}}(y)\frac{f(y)}{\tilde{f}_j(y)}\right]$$

$$\approx \sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=n_{j-1}+1}^{n_j} I_{A_{\alpha_j}}(Y_i)\frac{f(Y_i)}{\tilde{f}_j(Y_i)} = \sum_{j=1}^{M}\tilde{p_j} = \tilde{p}.$$

Recall that $E_{\tilde{f}}$ is expectation with respect to the shifted density $\tilde{f}$.

The value $\tilde{p}$ is the estimator we seek. Since the indicator functions, $I_{A_{\alpha_j}}$, divide the domain into disjoint sets, we have integrated over the entire domain and fully computed $P(x \in A_\alpha)$.

If we do not use the disjoint sets, it violates the basic equations of Monte Carlo estimators and the computations produce erroneous results. In other forms of Monte Carlo, it is necessary to account for bad samples by simply ignoring them in both the summations for the integrand and in accumulating the sample size [12]. Here, N does not need to be decreased when a sample from one MRP shift falls into a valid region for another MRP. Partitioning the sample space as we have says that even though a point may give a portfolio value less than v_α, it may not be important to the overall computation unless it does so only for the MRP from which it came.

As with any Monte Carlo estimator, we need to find its error bar, ϵ. This requires the Central Limit Theorem and properties of the Normal Distribution [13]. Each estimator, $\tilde{p}_j$, is $N(p_j, \frac{\tilde{\sigma}_j^2}{N_j})$ where $\tilde{\sigma}_j^2$ is the sample variance of the estimators[7]. Since $\tilde{p}$ is the sum of Normal random variables, $\tilde{p} \sim N(\sum_{j=1}^{M}\tilde{p}_j, \sum_{j=1}^{M}\frac{\tilde{\sigma}_j^2}{N_j})$. The error-bar, ϵ, we use is two standard deviations of the distribution of the estimator. Therefore,

$$P(x \in A_\alpha) = p \approx \tilde{p} \pm e = \sum_{j=1}^{M}\tilde{p_j} \pm 2\sqrt{\sum_{j=1}^{M}\frac{\tilde{\sigma}_j^2}{N_j}}$$

with a roughly 95% level of confidence.

The only question remaining is how to choose the N_i, the number of samples near each MRP. One could use using number of samples from the $\tilde{f}_j$ should be proportional to the value of the original density at the MRP, $f(x_{m_j})$. That is for a sample of size N, the j^{th} minimum rate point could be sampled

$$N_j = \frac{f(S_{m_j})}{\sum_{k=1}^{M} f(S_{m_k})}\cdot N \tag{18}$$

times. This is a purely heuristic method. Clearly one needs to sample more heavily near the more likely MRPs. If there are two points and one is more likely to happen, the majority of the Monte Carlo integration will occur near that one. But the other should not be ignored since you would not be integrating over the whole domain. This approach is validated by the results of computations on the third portfolio, Π_3.

[7] The variance $\tilde{\sigma}_j^2$ is analogous to the way the variance of the importance sampled estimator is defined in Section 2 for equation 5.

This portfolio, Π_3 consists of 150 shares of stock, S_1, and is short 400 puts on S_1. The current price of the stock is 20, the expected rate of returns, μ_1 is .09 and its volatility is .2. The options have the following parameters: Expiration Time, $T = .56$, Strike Price, $K = 20$, Risk Free Interest Rate, $r = .07$. The time-frame is .3288 years. There is only one underlying asset and the positions create a sort of straddle that will lose money if the stock goes too far up or down. Whether or not this is a good investment strategy is for another paper. But it does create a scenario with two Minimum Rate Points. The frequency of S_1 and corresponding value function can be seen in Figure 7.

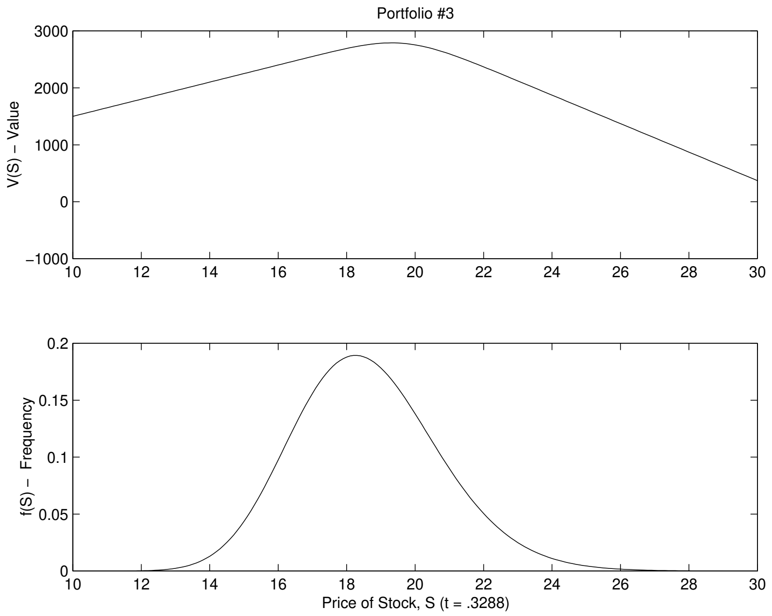

FIGURE 7. Frequency and Value of Π_3

To illustrate the effectiveness of this method with regards to multiple MRPs, we will begin with low values for v_α, which will have MRPs that are far away from each other. The two shifted densities will be probabilistically distinct with very little crossover.

The lower limit, v_α, will then be increased and the MRPs will be drawn closer to each other. More samples will be rejected for falling outside of their MPR's portion of the sample space despite yielding portfolio values, $V(S)$, that are less than v_α. As v_α increases, so does the accuracy of the regular Monte Carlo estimators. This decreases the need for importance sampling, but it provides us a way to check the accuracy of the full Importance Sampling method, which are still have greater accuracy for smaller sample sizes.

Table 9 shows values for $v_\alpha = 1800$ with regular Monte Carlo and Importance Sampling. The MPRs are very far apart, with one for $S_{m_1} = 12.0$ and the other at $S_{m_2} = 24.3$. Using the notation of equation 14, $N_1 = .026N$ and $N_2 = .974N$. When Importance Sampling is used, the majority of the samples clearly come from the region around S_{m_1}. The estimators, $\hat{p}$ and $\tilde{p}$ are within each others confidence intervals $(\pm\epsilon)$ at each value of N, demonstrating the validity of equation 17. The effectiveness of importance sampling is obvious since the relative error (the true measure of a Monte Carlo estimator) is less for $\tilde{p}$ with $N = 5000$ than for $\hat{p}$ with 100,000 samples.

The "rejected" samples are the X_n that would have been in A_α if not for the condition that they be closest to the minimum rate point, S_{m_i}, about which their twisted density, $\tilde{f}_j$, is created. In other words, they are not part of subset, A_{α_j}, from equation 16. Since portfolio Π_1 uses two twisted densities that are nearly disjoint, there are not too many rejections (around .6%). But they still must be rejected otherwise the algorithm implied by the general importance sampling equation (17) would not be followed.

method	N	estimator	$\epsilon = \frac{2\tilde{\sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	rejected
Reg. M.C. - $\hat{p}$	5,000	.010000	.002814	28.14%	50	N/A
	10,000	.009200	.001909	20.76%	92	N/A
	20,000	.009100	.001343	14.76%	182	N/A
	50,000	.009540	.000869	9.11%	477	N/A
	100,000	.009050	.000599	6.61%	905	N/A
Imp. Samp. - $\tilde{p}$	5,000	.008727	.000486	5.57%	2357	26
	10,000	.008977	.000349	3.89%	4778	63
	20,000	.008921	.000247	2.77%	9558	130
	50,000	.008880	.000156	1.75%	23820	325
	100,000	.008861	.000110	1.24%	47690	662

TABLE 10. Estimators for Π_3 with $v_\alpha = 1800$

The next computation uses $v_\alpha = 2000$. This has minimum rate points of $S_{m_1} = 13.33$ and $S_{m_2} = 23.49$. These are closer together and as Figure 7 suggests, the proportions for each MRP are more balanced than in the previous example. Equation 18 gives $N_1 = .206N$ and $N_2 = .794N$. Again, the estimators are within each others confidence intervals as N increases and ϵ shrinks, implying that they both will converge to the true value, p. But even ten times as many samples for regular Monte Carlo are not as accurate as Importance Sampling.

The final example, in Table 12, shows that even in a situation where Importance Sampling isn't really necessary, it still works and is still more accurate. With $v_\alpha = 2300$, we have $S_{m_1} = 15.33$, $S_{m_2} = 22.28$, $N_1 = .414N$, and $N_2 = .586N$. The twisted densities are much more equally used than before and about 5% of the samples are rejected.

The regular Monte Carlo estimator, $\hat{p}$ has a relative error of 5.75% for 10,000 samples. That is often a target, so importance sampling is not really called for. But $\tilde{p}$ is more than twice as efficient as $\hat{p}$, attaining greater accuracy with half as many samples. The sequences of both estimators appear to be converging to the same value, implying that the importance sampling equations (17) can be used in a general situation and will provide more accurate results while requiring fewer samples.

method	N	estimator	$\epsilon = \frac{2\tilde{\sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	rejected
Reg. M.C. - $\hat{p}$	5,000	.022600	.004204	18.60%	113	N/A
	10,000	.021600	.002907	13.46%	216	N/A
	20,000	.021500	.002051	9.54%	430	N/A
	50,000	.020880	.001279	6.12%	1044	N/A
Imp. Samp. - $\tilde{p}$	5,000	.020409	.001124	5.51%	2394	78
	10,000	.021089	.000812	3.85%	4871	155
	20,000	.021013	.000573	2.73%	9790	314
	50,000	.020917	.000362	1.73%	24358	752

TABLE 11. Estimators for Π_3 with $v_\alpha = 2000$

method	N	estimator	$\epsilon = \frac{2\tilde{\sigma}}{\sqrt{N}}$	rel. error($\frac{\epsilon}{\text{est.}}$)	samples	rejected
Reg. M.C. - $\hat{p}$	5,000	.105400	.008585	8.24%	527	N/A
	10,000	.107900	.006205	5.75%	1079	N/A
	20,000	.104700	.004330	4.14%	2094	N/A
	50,000	.104160	.002723	2.62%	5208	N/A
Imp. Samp. - $\tilde{p}$	5,000	.102885	.005170	5.03%	2509	262
	10,000	.102358	.003616	3.53%	5083	509
	20,000	.103240	.002578	2.50%	10055	1077
	50,000	.103157	.001627	1.58%	25231	2565

TABLE 12. Estimators for Π_3 with $v_\alpha = 2300$

5. Conclusion

As we've shown, Importance Sampling can greatly increase the efficiency and accuracy of Value at Risk computations. However, it should be reiterated that this was only valid for a a portfolio whose value at time t has an explicitly defined density. Portfolios with more exotic, time-dependent options, and those based on indices rather than d underlying variables do not necessarily lead to well-known distributions.

The general formula for Importance Sampling (17) was obtained through the analysis and expansion of known equations and the partitioning of the sample space (14) was obtained heuristically. Perhaps this method can be a starting point to computing VaR for more complex portfolios and other problems.

Appendix A. Non-Linear Optimization

In the general sense, constrained optimization with equality is about finding

$$\min_x M(x) \text{ subject to } \quad g_i(x) \leq 0, i = 1, \ldots, j.$$

$M(x)$ is referred to as the objective function and the $g_i(x)$ are the constraints. Since this appendix is about constrained optimization, separate from Value at Risk and Importance Sampling, the functions $M(x)$ and $g(x)$ will be used. The method described in this section was pooled from various techniques described in Bertsekas [2], Gill [6], Jacoby [9], and Pierre [15].

For this project, $M(x)$ is the opposite of the density, $-f(S)$ (since it is easier to recognize a local minimum than a local maximum), and the sole equality constraint $g(x)$ is taken to be $V(S)-v_\alpha$. The equality constraint can be used since the log-normal distribution is decreasing as it goes away from the global maximum and the minimum rate point will always be on the constraint. The method described in this section can be modified to account for inequality constraints.

The penalty method turns a constrained optimization problem into a global optimization problem of finding

$$\min_x \Phi(x, c) = M(x) + c_k \sum_{i=1}^{j} [g_i(x)]^2$$

where $\{c_k\}$ is an increasing sequence, typically where $c_{k+1} = mc_k$ with m between 4 and 10. The local minima of $\Phi(x)$ converge (as k increases) to the constrained minima of $M(x)$. Increasing the c_k is referred to as further penalizing $M(x)$. The fine-tuning referred to earlier is setting up the conditions under which c_k must be increased. Too little penalization creates a very "flat" $\Phi(x, c)$ which and too much creates one that is very steep away from the $g_i(x)$ and a "valley" along them. This can cause ill conditioning in matrices used to find the minima of $\Phi(x)$ and other numerical instabilities.

This is referred to as sequential unconstrained minimization, from the sequence of global optimizations that must be performed. However, the global minimizations aren't very exhausting since $x_m(c_k)$ is usually an excellent starting point to minimize $\Phi(x, c_{k+1})$. We are already close to the constrained minima, so we increase the penalty and begin the global minimization of $\Phi(x)$ from the previous point. The increased penalizing of $M(x)$ by going from c_k to c_{k+1} makes $x_m(c_{k+1})$ closer to satisfying one of the constraints.

Begin with

$$c_0 = \frac{\|\nabla M(x)\|}{\left\|\nabla \left(\sum_{i=1}^{j} [g_i(x)]^2\right)\right\|}$$

where ∇ the gradient operator. Use a global optimization method to find $x_m(c_0)$, a minimum of $\Phi(x, c_0)$. If $\forall\ i,\ g_i(x_m(c_0)) > \epsilon$, where ϵ is a preset tolerance level, increase the penalty to c_1. Continue increasing the c_k until $x_m(c_k)$ is found that will both minimize $\Phi(x, c)$ and satisfy the the constraints within a certain level of tolerance. Of course with equality constraints, many times only one will be satisfied. But it will be the one constraint that yields the infimum of the constrained minima if the problem were to be done for each individual constraint. It cannot be overstated how important it is to have good values for the $\{c_k\}$.

Using other values may work in some cases, but proceed with caution. For the simple portfolio, Π_1, c_0 was approximately 10^{-5}.

A modified Newton's method was used for the global optimization technique. For $x \in R^d$, Newton's methods for the minimization of a function $U(x)$ will immediately find the minimum of a quadratic approximation of $U(x)$ at a, the current guess. Let x_i, a_i be the i^{th} element of the vectors x and a. The polynomial is

$$Q(x,a) = U(a) + \sum_{i=1}^{d} \frac{\partial U(a)}{\partial x_i}(x_i - a_i) + \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial U(a)}{\partial x_i \partial x_j}(x_i - a_i)(x_j - a_j)$$

If that isn't the actual minimum, the routine is performed again with another quadratic approximation of $U(x)$.

Let $x^{(k)}$ be the k^{th} guess for x_m. Compute, $g^{(k)} = \nabla U(x^{(k)})$ and $H^{(k)} = \nabla^2 U(x^{(k)})$, the gradient and Hessian matrices of $U(x)$. Solve the linear system

$$H^{(k)} q^{(k)} = -g^{(k)}.$$

The solution, $p^{(k)}$ is the minimum of a quadratic approximation of $U(x^{(k)})$. Then set $x^{(k+1)} = x^{(k)} + q^{(k)}$. If $g^{(k+1)} = \nabla U(x^{(k+1)}) \approx 0$ and $H = \nabla^2 U(x^{(k+1)})$ is positive definite, then $x^{(k+1)} = x_m$, a local minimum of $U(x)$. If not, repeat the process for $x^{(k+1)}$. One is hoping $U(x^{(k+1)}) < U(x^{(k)})$.

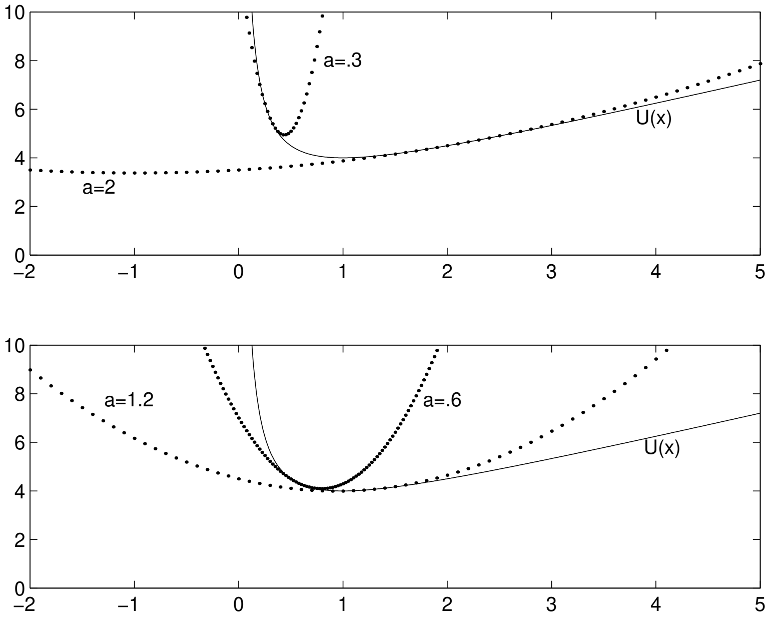

FIGURE 8. Quadratic approximations of $U(x) = \frac{(x+1)^2}{x}$ at $x = a$

Newton's methods can run into problems, so modifications must be made to insure convergence to x_m. If the quadratic approximation isn't very accurate, the method can return an $x^{(k+1)}$ that is very far from the minimum. An example would be $U(x) = \frac{(x+1)^2}{x}$. The minimum is at $U(1) = 4$. As seen in Figure 8, a quadratic approximation might be good only near the minimum. The function $\Phi(x, c)$ can often resemble this. So, care must be taken that one does not allow $x^{(k+1)}$ to be based on a bad approximation.

This can be done with a line search in the direction of q. Since q points to the minimum of $Q(x, x^{(k)})$, it is reasonable that it is in the right direction, but it may be too far or too close. So starting with $t = 1$, let

$$a = U(x^{(k)}), \quad b = U(x^{(k)} + tq^{(k)}), \quad c = U(x^{(k)} + 2tq^{(k)}).$$

If $a > b$ and $b \leq c$ then $x^{(k+1)} = b$. If $a \leq$, the approximation goes too far, so let $t = .5$. If $a \geq b > c$, it hasn't gone far enough, so let $t = 2t$ and try again.

Also, if the domain of $U(x)$ has any restrictions care should be taken that the line search does not attempt to evaluate the function in that area. If necessary, let $t = .5t$ and start the search over. In the case of the log-normal density, negative values are not permitted. And if the example in Figure 1 were to be used, assuming negative values are off limits, the line search at $x^{(k)} = 2$ would have to cut t in half several times before proceeding. It would return a value much closer to $x = 2$ than $x = 1$. After a few steps of small values of t, it would get to a point where the initial quadratic approximation is very accurate, like $x^{(k)} = 1.2$.

Another place where Newton's method might fail is if one is in a region where $H(x)$ is not positive definite. When $H(x)$ positive definite, it is the multi-dimensional equivalence the second derivative being positive. In this case, $x^{(k+1)}$ is the minimum of $Q(x, x^{(k)})$. If not, it could be a maximum since the method only finds a point where the gradient of $Q(x, x^{(k)})$ is zero. Performing a line search will often return $x^{(k+1)}$ so that $U(x^{(k+1)}) \geq U(x^{(k)})$. To counter this problem, Jacoby recommends the following: if $q^{(k)} \cdot g(x^{(k)}) \geq 0$, set $q^{(k)} = -q^{(k)}$. Then proceed with the line search for the best magnitude of $q^{(k)}$.

This is a fairly robust minimization method. The most computationally intensive parts are evaluating the functions and solving the linear system. There are ways to speed this up which are mentioned in [7]. And this is obviously not the only non-linear global optimization method.

A safeguard should be in place to guard against in ill conditioned Hessian matrix. If the process of solving the linear system, $Hq = -g$, has zero pivots, division by zero can occur and ruin the computations. To see how this is possible, return to the example of $U(x) = \frac{(x+1)^2}{x}$. In one dimension, the algorithm is $x^{(k+1)} = x^{(k)} - \frac{U'(x^{(k)})}{U''(x^{(k)})}$. Since $U''(x) = O(x^{-1})$, evaluating it very close to $x = 0$ could cause problems in floating point arithmetic.

Barring a failure of this form, the full constrained optimization method will return a local minimum, or at least a point that satisfies the terminating conditions. When searching for the minimum rate point(s), it is possible that different local minima are found for different starting points, $x^{(0)}$. Some may be valid minimum rate points, in the case of non-linear constraints, $g_i(x)$. In some cases, $\Phi(x, c)$ may be very flat so that $H(x)$ is positive definite, the gradient is suitable small and a constraint is sufficiently satisfied. But the minimum rate point, x_{m1} might be insignificant compared to another, x_{m0}, i.e. $f(x_{m1}) \ll f(x_{m0})$ for the density, f. To find as many potential minimum rate points as possible, a method that works that one samples the initial guess, $x^{(0)}$, from F and then implement the constrained optimization and do this several times. It could be that a particular $x^{(0)}$ leads the optimization routine to a bad point. For example, when trying to find the most likely way that $V(S) = v_\alpha$, it might go in a direction where the constraint would be satisfied if an S_i is negative. This would should cause the algorithm to terminate, returning a bad value for S_m. This is more likely to happen in the case of a v_α that represents a large loss or in a portfolio with options. And looking at Figure 9, a graph of of $\Phi(S,c)$, for Π_1, one can see that the objective function being optimized forms a basin along the constraint. If $x^{(k)}$ is the first step that enters this basin and $\|\nabla\Phi(x^k, c_k)\|$ is sufficiently small, the algorithm might terminate, assuming it has found a minimum when in fact it needs to take another step or two along the path of the basin.

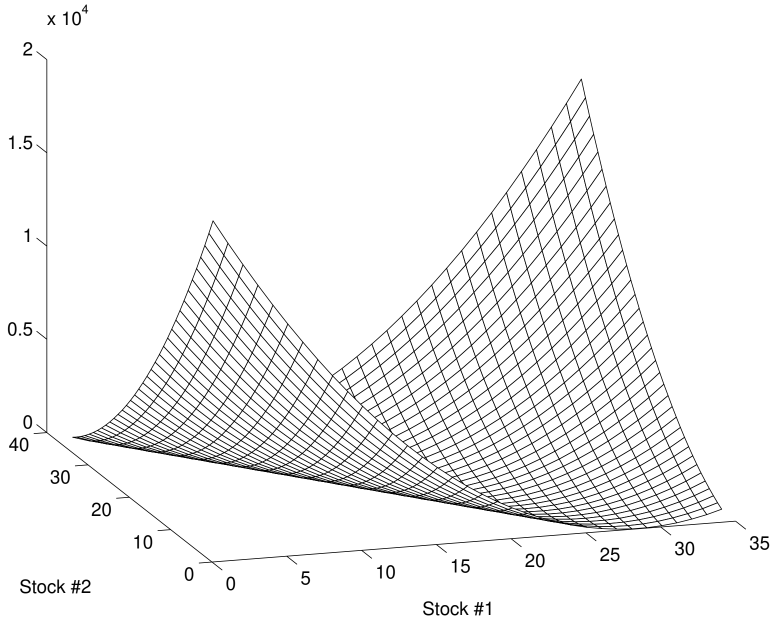

FIGURE 9. Penalty function, $\Phi(S,c)$ for Π_1 with $c = .1, v_\alpha = 4300$

To do the constrained optimization with a portfolio with options, recall that the probability density function of $S(t)$ is for $a_i = \log(S_i(0)) + (\mu_i - \frac{1}{2}\sigma^2)t$ and $(\Sigma)_{ij} = \rho_{ij}\sigma_i\sigma_j t$,

$$f(S) = \frac{1}{S_1, \ldots, S_d\sqrt{det(\Sigma)}(2\pi)^{\frac{d}{2}}} \exp\left[\frac{1}{2}(S-a)^T \Sigma^{-1}(S-a)\right]$$

is that which must be maximized subject to the constraint of the value function,

$$V(S(t)) = B_0 e^{rt} + \sum_{i=1}^{d} s_i S_i(t) + c_i C(S_i(t), \sigma_i, r, K_{c,i}, T_{c,i}) + p_i P(S_i(t), \sigma_i, r, K_{p,i}, T_{p,i}).$$

Using the penalty method and changing the sign of f to make it a minimization problem,

$$\Phi(S,c) = -f(S(t)) + c\left(V(S(t)) - v_\alpha\right)^2$$

is the objective function to which Newton's method for global optimization must be applied.

To evaluate the gradient vector and Hessian matrix of $\Phi(S,c)$ one must use the sensitivies, or the "Greeks", for the option components of $V(S)$. The ones required for this model are

$$\Delta_{p_i} = \frac{\partial P(\cdot)}{\partial S_i},\ \Delta_{c_i} = \frac{\partial C(\cdot)}{\partial S_i},\ \Gamma_{p_i} = \frac{\partial^2 P(\cdot)}{\partial S_i^2},\ \text{and}\ \Gamma_{\mathrm{c_i}} = \frac{\partial^2 \mathrm{C}(\cdot)}{\partial \mathrm{S_i^2}}.$$

The exact formulae for European puts and calls can be found in texts such as Hull [8] or Wilmot [20]. If a closed form expression for the Greeks isn't obtainable, one can always use finite differences. It should be noted that for the expanded stock model that includes stochastic volatility and interest rates, the vegas and rhos of the options would be required.

Then

$$\nabla\Phi(S,c)_i = -f(S) f_{S_i} + 2c\left[V(S) - v_\alpha\right] V_{S_i}$$

where

$$V_{S_i} = \frac{\partial V(S)}{\partial S_i} = s_i + c_i \Delta_{c_i} + p_i \Delta_{p_i}.$$

And for the Hessian matrix,

$$(\nabla^2\Phi(S,c))_{ij} = -\left(f(S) f_{S_i S_j} + f_{S_i} f_{S_j}\right) + 2c\left(V(S) - v_\alpha\right) V_{S_i S_j} + V_{S_i} V_{S_j}$$

with

$$V_{S_i S_j} = \frac{\partial^2 V(S)}{\partial S_i \partial S_j} = \left\{ \begin{array}{cc} c_i \Gamma_{c_i}(S) + p_i \Gamma_{p_i}(S) & i = j \\ 0 & i \neq j \end{array} \right. .$$

Of course f_{S_i} and $f_{S_i S_j}$ are the partial derivatives of f evaluated at S and can be (carefully) derived with calculus.

Returning to the portfolio, Π_1 with $v_\alpha = 4300$, the constrained optimization of the $f(S)$, the multi-variate log-normal density, subject to $V(S) - v_\alpha = 0$ returns

$$S_m = \begin{pmatrix} 14.8076 \\ 20.7886 \end{pmatrix}.$$

This is the global minimum of $\Phi(S,c) = -f(S) + c(V(S) - v_\alpha)^2$. A graph of this can be seen in Figure 9.

Appendix B. Quantile Estimation

The importance sampling method presented here is in the spirit of Value at Risk (VaR), it is about estimating $p = P(V \le v_\alpha)$ for a given v_α. VaR is concerned with the other perspective, given α, what is v_α such that $P(V \le v_\alpha) = \alpha$. If one is to make no assumptions about the distribution of V, and with portfolios of options, that is often the case, there are a few methods of quantile estimation related to Monte Carlo for finding an estimator, $\tilde{v}_\alpha$ for v_α.

The first is order statistics [5]. This is taking a sample of random variables, $X_1, X_2, \ldots, X_N$ and sorting them into $X_{1:N}, X_{2:N}, \ldots, X_{N:N}$. The obvious choice for $\tilde{v}_\alpha$ is $X_{[100\alpha N]:N}$ where $[\cdot]$ is the largest integer operator. However, like all statistics, this needs a confidence interval that will ideally shrink as N increases. Many writings on VaR do not mention this when discussing Monte Carlo. They just assume that for the valuations, $V^{(1)}, V^{(2)}, \ldots, V^{(N)}$, $V_{[100\alpha N]:N}$ will work perfectly well. In many cases it does since N is usually quite large. But it is a statistical estimator and does require a confidence interval.

Information on the distribution of order statistics can be found in [5]. For a random variable, X, with distribution $F(x)$ let x_α be the value of x such that $P(X < x_\alpha) = \alpha$. If $X_1, X_2, \ldots, X_N$ are a sample from F, and $X_{1:N}, X_{2:N}, \ldots, X_{N:N}$ are the corresponding order statistics then the interval $(X_{r:N}, X_{s:N})$ covers x_α with probabilty

$$I_\alpha(r, N - r + 1) - I_\alpha(s, N - s + 1).$$

Here $I_\alpha(\cdot)$ is the incomplete Beta function, defined by

$$I_a(b, c) = \frac{\int_0^a t^{b-1}(1-t)^{c-1}dt}{\int_0^1 t^{b-1}(1-t)^{c-1}dt}.$$

Fortunately, this can be calculated in Matlab. With $N = 10,000$, and $\alpha = .05$, the probability that $x_{.05}$ is in the interval $(X_{455:10,000}, X_{545:10,000})$ is .9610. If $\alpha = .01$, then $(X_{80:10,000}, X_{120:10,000})$ forms a 95.54% confidence interval for $x_{.01}$.

VaR is usually given for the first or fifth percentile. For $V_1, V_2, \ldots, V_{10,000}$ sampled from Π_2 in the last section, the eight dimensional portfolio. Figure 1 shows the empirical distribution function of the V_i. Using the 100th and 500th order statistic for the respective levels of Value at Risk gives $v_{.01} = 26,654$ and $v_{.05} = 27,510$. The confidence intervals associated with $v_{.01}$ is $(27437, 27560)$ for a relative error of .26%. In this case the relative error was taken to be $\frac{X_{s:N} - X_{r:N}}{X_{[100\alpha N]:N}}$. For $v_{.01}$, the confidence interval is $(26410, 26660)$ for a relative error of .47%.

Another way of estimating quantiles is the Robbins-Monro stochastic approximation method [16]. Without going into details, this method is not very suitable for the problems in this thesis. It is better for problems where data is scarce which is not our case. It is has problems with extreme quantiles. However, in Tierney [18] a method is laid out that does work for quantiles in the tails of a density as well as providing error bars.

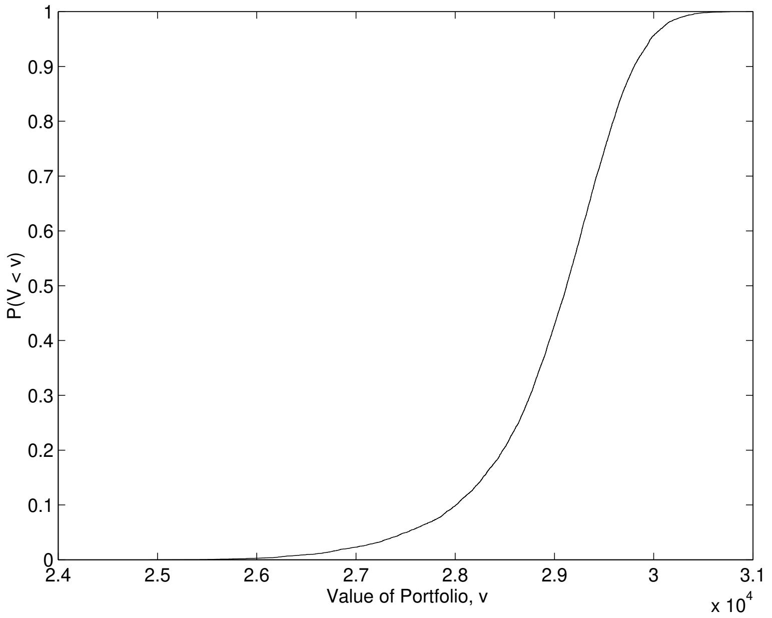

FIGURE 10. Empirical Distribution of of $V_{\Pi_2}(S(.0822))$

REFERENCES

[1] Ludwig Arnold. *Stochastic Differential Equations: Theory and Applications.* Kreieger, Malabar, Florida, 1992.

[2] Dimitri P. Bersekas. *Constrained Optimization and Lagrange Mulitplier Methods.* Academic Press, New York, 1982.

[3] James Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley, New York, 1990.

[4] Edwin L. Crow and Kunio Shimizu, editors. *Lognormal Distribution: Theory and Applications.* M. Dekker, New York, 1988.

[5] H. A. David. *Order Statistics.* Wiley, New York, 1970.

[6] P. E. Gill and W. Murry. *Numerical Metthods for Constrained Optimization.* Academic Press, London, New York, 1974.

[7] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins Press, Balitimore, Maryland, 1983.

[8] John C. Hull. *Options, Futures and other Derivative Securities.* Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[9] S. L. S. Jacoby, J. S. Kowalik, and J. T. Pizzo. *Iterative Methods for Nonlinear Optmization Problems.* Prentice Hall, Englewood Cliffs, New Jersey, 1972.

[10] Norman Lloyd Johnson and Samuel Kotz. *Continuous Multivariate Distributions.* Wiley, New York, 1972.

[11] Philippe Jorion. *Value at Risk: The New Benchmark in Controlling Market Risk.* Irwin, Chicago, 1997.

[12] Malvin Kalos and Paula A. Whitlock. *Monte Carlo Methods Volume 1: The Basics.* Wiley, New York, 1986.

[13] Alan F. Karr. *Probability.* Springer-Verlag, New York, 1993.

[14] Nigel J. Newton. Variance reduction for simulated diffusions. *SIAM Journal of Applied Mathematics*, 54(6):1780–1805, 1994.

[15] Donald Pierre. *Optimization Theory with Applications.* Wiley, New York, 1969.

[16] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[17] John Sadowsky and James Bucklew. On large deviations and asymptotically efficient monte carlo estimation. *IEEE Transactions on Information Theory*, 36(3):159–171, 1990.

[18] Luke Tierney. A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal of Scientific Statistical Computing*, 4(4), Dec 1983.

[19] W. Wagner. Monte carlo techniques for stochastic differential equations. In Blagovest H. Sendov and Ivan Dimov, editors, *Internationl Youth Workshop on Monte Carlo Methods and Parallel Algorithms, Primorsko, Bulgaria, 24-30 September 1989*. World Scientific, Singapore and Teaneck, New Jersey, 1991.

[20] Paul Wilmot, Sam Whoison, and Jef Dewynn. *The Mathematics of Financial Derivatives*. Cambridge University Press, Cambridge, 1995.