May 2015

Extensions of the Cross-Entropy Method with Applications to Diffusion Processes and Portfolio Losses

Alexandre Scott
The University of Western Ontario

Supervisor
Adam Metzler
The University of Western Ontario

Graduate Program in Applied Mathematics

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Extensions of the Cross-Entropy Method with Applications to Diffusion Processes and Portfolio Losses

(Thesis Format: Integrated-Article)

by

Alexandre Scott

Graduate Program in Applied Mathematics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
Western University
London, Ontario, Canada

Abstract

Rare event simulation is a crucial part of simulations. In financial mathematics, the study of rare events appear naturally when we consider risk measures such as the conditional value at risk. This thesis is composed of three related papers treating the rare event simulations subject: the first paper addresses rare event simulations for diffusion processes, the second paper addresses rare event simulations for the normal and the Student t-copula model while the last paper addresses rare event simulations for a portfolio model where there is a correlation structure between the loss-given-default and the probability of default.

Keywords: cross-entropy, Kullback-Leibler divergence, rare event simulations, loss estimation, Radon-Nykodym derivative, Importance Sampling.

The Co-Authorship Statement

My papers have all been written with Adam Metzler. I, Alexandre Scott, am the principal author.

Acknowledgments

I would like to thank the staff at the University of Western Ontario for their support. I would like to thank Adam Metzler, Mark Reesor, Matt Davison and Greg Reid for their advices and help when needed. I would also like to thank my parents, Marc and Sylvie and my brothers Maxime and Dominick who have always been very supportive. Finally, I would like to thank NSERC and FQRNT for their financial support.

Table of Contents

Abstract ii

The Co-Authorship Statement iii

Acknowledgments iv

List of Tables vii

List of Figures viii

List of Appendices x

List of Abbreviations, Symbols, Nomenclature xi

1 Introduction 1
1.1 The Cross-Entropy Method 6
Bibliography 8

2 Rare Event Simulation for Diffusion Processes 9
2.1 Introduction 9
2.1.1 Notation and Assumptions 11
2.2 Proposed Algorithm 12
2.2.1 First Stage Application - Discretization Error 12
2.2.2 Second Stage Application - Rare Event Problem 14
2.3 Rare Event IS for Brownian Motion 16
2.3.1 The Guasoni and Robertson (GR) Criteria 17
2.3.2 Entropy Minimization 18
2.3.3 Comparison Between Minimum Entropy and GR 22
2.4 Functions of Extremes 25
2.4.1 Example - CIR Process With Large Maximum 28
2.4.2 Example - OU Process With Large Maximum and Low Minimum 31
2.5 Concluding Remarks 34
Bibliography 35

3 Rare Event Simulation for Portfolio Losses 36
3.1 Introduction 36
3.1.1 Summary of Proposed Algorithm 37

3.1.2 Brief Literature Review . . . 38
3.1.3 Outline of the Chapter . . . 39
3.2 Financial Setting . . . 40
3.3 Statistical Setting . . . 41
3.3.1 Sequential Exponential Tilts . . . 43
3.3.2 Minimum Divergence with Sequential Exponential Tilts . . . 44
3.3.3 General Algorithm . . . 45
3.3.4 Important Special Case . . . 47
3.4 Application to the Portfolio Problem . . . 47
3.4.1 Estimating $\widetilde{\mathbb{E}}[f(\mathbf{Z})]$ and Computing $\hat{\lambda}$ . . . 48
3.4.2 Estimating $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}]$ . . . 49
3.5 Normal Copula Model . . . 51
3.5.1 Computing $\hat{\mu}$ . . . 53
3.5.2 Computing $\hat{\theta}_\bullet$ . . . 54
3.5.3 Numerical Results . . . 54
3.6 t Copula Model . . . 56
3.6.1 Multivariate t Distribution . . . 58
3.6.2 Candidate Family . . . 58
3.6.3 Computing $\hat{\mu}$ and $\hat{\alpha}$ . . . 60
3.6.4 Computing $\hat{\theta}_{\bullet,\bullet}$ . . . 62
3.6.5 Numerical Results . . . 62
3.7 Concluding Remarks . . . 64
Bibliography . . . 65

4 Variance Reduction for Models with PD-LGD correlation 67
4.1 Introduction . . . 67
4.2 The Model . . . 68
4.3 Appropriate IS Densities . . . 70
4.3.1 New Form for the Likelihood Ratio . . . 74
4.3.2 First Stage . . . 75
4.3.3 Second Stage . . . 76
4.3.4 Alternative for the Second Stage . . . 78
4.4 Numerics . . . 80
4.4.1 Performance . . . 82
4.5 Concluding Remarks . . . 85
Bibliography . . . 86

5 Conclusion 88
Bibliography . . . 90

Appendix 91

Curriculum Vitae 114

List of Tables

Table 3.1 This table compares the performance of the proposed algorithm to the performance of the Glasserman and Li (2005) algorithm. We assume there are ten groups and fifteen risk factors. Factor loadings, marginal default probabilities and relative exposures can be found in Appendix C. Note that factor loadings are chosen to ensure that $|a_g|^2 \leq 0.11$, meaning a low correlation environment. In all cases obligors are divided equally across groups. i.e. $N_g = N/G$. 57

Table 3.2 This table compares the performance of the proposed algorithm to the performance of the algorithm proposed by Chan and Kroese (2010). Base parameters are $N = 2000$, $\nu = 15$, $a_i = 0.3$ and $PD_i = 0.029$ for all i, $x = 0.4$. CPU Time is given in seconds. . 63

Table 3.3 This table compares the performance of the IS estimator (3.17) to that of the crude estimator, when estimating the conditional tail expectation $\mathbb{E}[L_N | L_N \geq x]$ in the t copula model. Portfolio consists of $N = 500$ obligors, evenly distributed among $G = 10$ groups (i.e. $N_g = 50$ for each g), and there are $D = 15$ Gaussian risk factors. Degrees of freedom are $\nu = 5$. Default probabilities, factor loadings and exposures are given in Appendix C, and loss given defaults are all set to 100%. For each value of x in the table, 20 realizations of each estimator were simulated using $M = 10,000$ in each case. Reported mean and standard deviation are the sample mean and sample standard deviation of the 20 realizations. A reported value of NA means that at least one of the 20 samples of 10,000 did not contain a single observation for which $L_N \geq x$, in which case both numerator and denominator of the crude estimator is zero. The reported value of $\mathbb{P}(L_N \geq x)$ is based on an IS estimate using a sample of size $200,000$. . . . 64

Table 4.1 Performance of the first-stage estimator and the two-stage estimator when $PD = 0.029$, $a = 0.63$, $b = 0.975$, $\rho = 0.2$, $x = 0.1$, $\beta = 0.25$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$ and 200 obligors. We used $1,000,000$ simulations for each estimator. 81

List of Figures

Figure 1.1 Coefficient of variation for estimating $P(W_1 > 4)$ against μ. . . 6

Figure 2.1 Average path against time 21
Figure 2.2 Comparison between $\hat{\alpha}$ and $\widetilde{\alpha}$ 24
Figure 2.3 Quadratic fit for $v(\alpha)$ 24
Figure 2.4 Estimated coefficient of variation against the probability 30
Figure 2.5 Estimated coefficient of variation against the probability estimating a non-trivial function 31
Figure 2.6 Estimated coefficient of variation versus the estimated probability for the First-Stage estimator, Two-Stage estimator with constant drift and the Two-Stage estimator with piecewise constant drift. 33

Figure 3.1 100 simulated pairs $(L_N, \mathbb{E}[L_N|\mathbf{Z}])$ in the single-factor (i.e. $d = 1$) homogeneous Gaussian copula model. Parameters are $N = 1000$, $PD_i = PD = 0.02$, and $a_i = 0.2$ for all i. The empirical correlation between the two series exceeds 90%. 52
Figure 3.2 Average absolute error between $\widetilde{\mathbb{E}}[L_N|\mathbf{Z} = \mathbf{z}]$ and $\max\{\mathbb{E}[L_N|\mathbf{Z} = \mathbf{z}], x\}$ in the single-factor (i.e. $d = 1$) Gaussian copula model. Parameters are $N = 200$, $PD_i = 0.02$, $a_i = 0.2$ for all i. For each value of x, we simulated $10,000$ values of $\mathbf{Z}$ and for each value of $\mathbf{Z}$ we computed $\widetilde{\mathbb{E}}[L_N|\mathbf{Z} = \mathbf{z}]$ using the fact that $N \cdot L_N|\mathbf{Z} = \mathbf{z} \sim \text{Bin}(N, p(\mathbf{z}))$. The error $\widetilde{E}[L_N|\mathbf{Z}] - \max(x, \ell_N(\mathbf{z}))$ was then averaged across all simulated values and the solid blue line is a cubic spline fitted to that averaged data. The maximum error is located near $x = PD$ (point in red), and we recall that in the rare event setting one is typically interested in $x \gg PD$. 55
Figure 3.3 This figure illustrates the near-optimal performance of the proposed algorithm in the single-factor homogeneous Gaussian model. The horizontal axis is the tilt parameter used in the first stage (in all cases (3.14) is used in the second stage) and the vertical axis is the log of the estimated coefficient of variation (estimated using $1,000,000$ simulations). Other parameters are $N = 1000$, $x = 0.2$, $a_i = 0.2$, $PD_i = 0.02$ for all i. 56

Figure 3.4 This figure illustrates the near-optimal performance of the proposed algorithm in the two-factor inhomogeneous Gaussian model. Contour plot of $CV(\mu)$ where $CV(\mu)$ is the coefficient of variation assuming the first stage tilt parameter is $\mu \in \mathbb{R}^2$ and the second stage tilt parameter is $\hat{\theta}_\bullet$. $CV(\mu)$ is estimated by Monte Carlo using $1,000,000$ simulations. Parameters are $a_1 = [0.1 \quad 0.1]^T$, $a_2 = [0.1 \quad 0.2]^T$, $PD = [0.025 \quad 0.0145]^T$, $x = 0.2$ with 500 assets distributed equally among each group (i.e. $N_1 = N_2 = 250$). . . 57
Figure 3.5 100 simulated pairs $(L_N, \mathbb{E}[L_N|\mathbf{Z}])$ in the t copula model. Parameters are $N = 1000$, $\nu = 15$, $PD_i = PD = 0.02$ and $a_i = 0.3$ for all i. The empirical correlation between the two series exceeds 98%. 59

Figure 4.1 Order of magnitude of the estimated probability $\mathbb{P}(L_N > x)$ against the threshold. The parameters are $a = 0.63$, $b = 0.975$, $PD = 0.029$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$, $\rho = 0.2$, $\beta = 0.25$ with 200 obligors. We used $1,000,000$ simulations per point. 83
Figure 4.2 Coefficient of variation of the estimator for $\mathbb{P}(L_N > x)$ against the threshold. The parameters are $a = 0.63$, $b = 0.975$, $PD = 0.029$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$, $\rho = 0.2$, $\beta = 0.25$ with 200 obligors. We used $1,000,000$ simulations per point. 84
Figure 4.3 Average iterations required for two-stage entropy algorithm against the threshold. The parameters are $a = 0.63$, $b = 0.975$, $PD = 0.029$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$, $\rho = 0.2$, $\beta = 0.25$ with 200 obligors. We used $1,000,000$ simulations per point. 85
Figure 4.4 Order of magnitude of the coefficient of variation of the estimation of $\mathbb{P}(L_N|L_N > x)$ for the One-Stage entropy criteria and the two-stage entropy criteria. The parameters are $a = 0.63$, $b = 0.975$, $PD = 0.029$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$, $\rho = 0.2$, $\beta = 0.25$ with 200 obligors. We used $100,000$ simulations per point. 86

List of Appendices

A Supplementary Results and Proofs **91**

B Useful Conditional Expectations for Standard Brownian Motion **96**
B.1 Large Maximum 96
B.2 Large Maximum and Small Minimum 97

C Parameters **100**

D Equations **102**

E Algorithms **104**

Bibliography **113**

List of Abbreviations, Symbols, Nomenclature

$C_\phi(W, N)$ Poisson kernel estimator

$\mathbf{F}$ Lamperti Transform

$\ell_N(\mathbf{Z})$ Asymptotic approximation of L_N given $\mathbf{Z}$

L_N Loss on a portfolio of N obligors

PD_i Marginal default probability for obligor i

$\Psi(\lambda)$ Cumulant generating function

$\psi_{\mathbf{r}}(\theta)$ Cumulant generating function conditioned on $\mathbf{R} = \mathbf{r}$

$\widetilde{\mathbb{P}}$ Conditional measure of $\mathbb{P}$

W_t Standard Brownian Motion

W^θ Brownian Motion with deterministic drift θ_t

$\overline{X}$ Running maximum of the stochastic process X

$\underline{X}$ Running minimum of the stochastic process X

Chapter 1

Introduction

Monte Carlo simulation is an incredibly powerful and versatile computational method for stochastic models. A crude Monte Carlo estimator can easily and rapidly be applied to a variety of problems, however there are many cases where the crude Monte Carlo estimator fails to provide satisfactory estimates. For example, let us suppose that a bank has set aside five million capital reserves against losses on a corporate portfolio. The bank needs to estimate (i) the probability that losses exceed this reserve level and (ii) the expected shortfall (i.e. expected loss given that losses exceed 5 million). If the true probability is 0.1%, then only 1 simulation out of $1,000$ is informative for (ii) and this is very problematic since computational budgets are limited in practice.

To accelerate the rate at which informative scenarios are generated, one can use Importance Sampling in conjunction with Monte Carlo simulations. In this thesis, we consider problems of the form

$$\mathbb{E}\left[f(X)\cdot\mathbf{1}_{\{X\in A\}}\right], \tag{1.1}$$

which can easily be extended to problems of the form $\mathbb{E}\left[f(X)|X\in A\right]$ via the relation:

$$\mathbb{E}\left[f(X)|X \in A\right] = \frac{\mathbb{E}\left[f(X) \cdot \mathbf{1}_{\{X \in A\}}\right]}{\mathbb{E}\left[\mathbf{1}_{\{X \in A\}}\right]}.$$

Let

$$\mathbf{1}_{\{X \in A\}} = \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{otherwise.} \end{cases}$$

and where f is a real-valued function, X is a random element mapping to an arbitrary space G and A is a subset of G. In practice, it is often impossible to compute (1.1) directly, thus Monte Carlo is the only feasible option. A naive scheme can easily be implemented as follows

1. Generate N independent copies of the random element X. We shall refer to $X^{(i)}$ as the i^{th} simulated copy.

2. Return

$$\frac{1}{N}\sum_{i=1}^{N} f(X^{(i)}) \cdot Y_i,$$

   where $Y_i = \mathbf{1}_{\{X^{(i)} \in A\}}$.

We are interested in the behaviour of $f(X)$ over A but if $\mathbb{P}(X \in A)$ is small then very few simulated values are informative. This manifests itself in an unacceptably large coefficient of variation (CV), defined as the ratio of an estimator's standard deviation to its expected value. The standard error can be controlled by using a very large number of simulations, but in practice, this can be prohibitively expensive.

To improve the quality of the simulations, one can implement variance reduction techniques such as exponential tilts, control variates or Importance Sampling. However, exponential tilts are quite difficult to implement in practice because we do not always know the moment generating function of the function we want to estimate. We would

then need to implement an acceptance rejection algorithm, but when if the tilted distribution does not bear enough resemblance to the original distribution, we might end up with a very low acceptance probability as discussed in Chapter 4. For control variates, it is not exactly clear how to compute the needed parameters outside of the homogeneous model without having to resort to using pre-simulations to approximate our optimal parameters. In this thesis, we propose Importance Sampling estimators to improve the performance of the estimators for two reasons: firstly, we are able to exploit Girsanov in Chapter 2 and bring back our problem to simulating Brownian Motions with deterministic drifts and secondly, we are able to use the asymptotic behaviour of large portfolios in Chapter 3 and Chapter 4 to rapidly compute the parameters for our IS estimators.

Recall that, using IS estimators, (1.1) can be rewritten as

$$\mathbb{E}\left[f(X)\cdot\mathbf{1}_{\{X\in A\}}\right]=\mathbb{E}^{\mathbb{Q}}\left[f(X)\cdot\mathbf{1}_{\{X\in A\}}\cdot\frac{d\mathbb{P}}{d\mathbb{Q}}\right], \tag{1.2}$$

where $\mathbb{E}^{\mathbb{Q}}[\cdot]$ denotes the expected value under measure $\mathbb{Q}$, $\frac{d\mathbb{P}}{d\mathbb{Q}}$ denotes the Radon-Nykodym derivative (RND) and where $\mathbb{P}$ is a measure dominated by $\mathbb{Q}$ (if $\mathbb{Q}(x\in C)=0$ then $\mathbb{P}(x\in C)=0$). Using the right-hand side of (1.2), we can improve the estimation of (1.1). But, choosing $\mathbb{Q}$ is not a trivial problem. On the one hand, we would like to choose $\mathbb{Q}$ such that most of our simulations are informative. On the other hand, if $\mathbb{Q}$ is too different from $\mathbb{P}$, then the simulated RNDs might be very large, resulting in a worse estimator than the crude Monte Carlo estimator. To show how complex this problem can become, we will consider the example of estimating the probability that a standard Brownian Motion exceeds a fixed threshold $a>0$ at a given time time T. Although we can obtain this probability in closed-form, this problem really highlights the difficulty that one faces when choosing an importance density.

Example 1.0.1. *Let W_t be a standard Brownian Motion and consider the problem of estimating*

$$\mathbb{P}(W_1 > a),$$

where $a > 0$. By the properties of the Brownian Motion, we know that $\mathbb{P}(W_1 > a) = \Phi(-a)$, where $\Phi(\cdot)$ denotes the cdf of a standard normal distribution. If $a = 4$, then $\mathbb{P}(W_1 > a) \approx 3.17 \times 10^{-5}$. Using the naive estimator and $1,000,000$ simulations, we obtain the following estimate for $\mathbb{P}(W_1 > a)$:

$$\hat{p} = 2.70 \times 10^{-5}, \quad \hat{\sigma} = 5.20 \times 10^{-3},$$

which leads to an approximate CV of 192.60. Let us consider an IS measure, $\mathbb{Q}^\mu$, where W_t is a Brownian Motion with drift μ, i.e $\mathbb{Q}^\mu(W_t \in dz) = \phi(z, \mu, \sqrt{t})dz$, where $\phi(\cdot, \mu, \sqrt{t})$ denotes a normal pdf with mean μ and variance t. Intuitively, choosing large positive μ should steer more trajectories upwards, and this should make the simulations more informative. Using the right-hand side of (1.2) as the IS estimator with $\mu = a$, we obtain

$$\widetilde{p} = 3.16 \times 10^{-5}, \quad \widetilde{\sigma} = 6.72 \times 10^{-5},$$

which leads to a coefficient of variation of 2.13. The above estimation is very accurate, and one could suspect that increasing μ would only result in a better estimator. Using the same IS estimator when $\mu = 8$, we obtain

$$\widetilde{\widetilde{p}} = 3.08 \times 10^{-5}, \quad \widetilde{\widetilde{\sigma}} = 3.30 \times 10^{-3}.$$

This leads to a coefficient of variation of 107.14, which is almost as inaccurate as the crude Monte Carlo estimator. Even though more simulations are informative, the

RND is becoming increasingly large, and the trade-off between the two can be measured by looking at the coefficient of variation $c(a, \mu)$ *defined as*

$$c(a,\mu) = \frac{\sqrt{\mathrm{Var}^{\mathbb{Q}^\mu}\left[\mathbf{1}_{\{W_T>a\}} \cdot \frac{d\mathbb{P}}{d\mathbb{Q}^\mu}\right]}}{\mathbb{E}^{\mathbb{Q}^\mu}\left[\mathbf{1}_{\{W_T>a\}} \cdot \frac{d\mathbb{P}}{d\mathbb{Q}^\mu}\right]} = \sqrt{\frac{e^{\mu^2}\Phi(-a-\mu)}{\Phi(-a)^2} - 1}. \tag{1.3}$$

Minimizing (1.3) *with respect to* μ *(assuming* a *fixed) is equivalent to solving the non-linear equation below*

$$2\mu = \frac{\phi(-a-\mu)}{\Phi(-a-\mu)}. \tag{1.4}$$

If we assume that μ *increases as* a *increases, then as* $a + \mu \to \infty$*, the right-hand side of* (1.4) *is asymptotically equivalent to*

$$\frac{\phi(-a-\mu)}{\Phi(-a-\mu)} \sim a + \mu.$$

Therefore, it follows that (1.4) *is asymptotically equivalent to*

$$\mu \sim a,$$

which explains why our first estimator performs so well. Therefore, the optimal trade-off (for this example) between IS measures too similar and not similar enough is to have about 50% *of the simulations to lie in the region where* $\{W_1 > a\}$*.*

For this particular problem, we know exactly the quantity we want to estimate so it was possible to minimize the coefficient of variation directly. However, for more complex problems we do not have a closed-form solution for the probability we want to estimate, thus we need a criteria that will allow us to choose an IS measure efficiently.

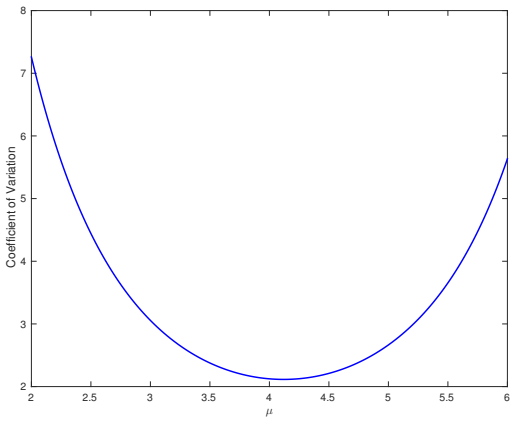

Figure 1.1: *Coefficient of variation for estimating* $P(W_1 > 4)$ *against* μ.

1.1 The Cross-Entropy Method

The cross-entropy methodology is a well-established method to determine how similar two measures are (see Asmussen and Glynn (2007) and references therein for a comprehensive treatment of the methodology). It has been introduced by Rubinstein (1997) and although it has been applied to a wide variety of contexts, it has not been applied to either diffusion processes or the portfolio problem. Applying the methodology as-is for either of these problems is not straightforward, and the contribution of this thesis is to extend the cross-entropy method to areas where it has not been applied yet. Moreover, we find that combining the cross-entropy method and exponential families leads to optimal IS parameters that solve very intuitive moment-matching criteria. However, the relevant moments are almost never available in closed-form, thus we need to develop tractable approximations in order to apply the method.

The cross-entropy method cannot really be applied as developed in Asmussen and Glynn (2007) for diffusion processes or for the portfolio problem considered in Chapter 3. In Chapter 2, we show how the cross-entropy method should be applied to Brownian Motion processes and compare minimum entropy's performance to the Large Deviation criteria introduced by Glasserman et al. (1999) and later improved

by Guasoni and Robertson (2008). Then, we extend our algorithm to more general diffusion processes using a two-stage IS estimator where the first stage uses the Lamperti transform and the so-called Poisson-Kernel estimator introduced by Beskos and Roberts (2005) and Chen and Huang (2012) to transform the diffusion process into a standard Brownian Motion and where the second stage uses cross-entropy to improve the accuracy of the simulations.

In Chapter 3, we apply the cross-entropy method on the portfolio problem. We extend the idea of exponential tilts to the idea of sequential tilts. We show that when the cross-entropy method is combined with sequential exponential tilts, we still get the intuitive moment-matching conditions. Then, we show how tractable approximations of these moment-matching conditions are of paramount importance in order to obtain a fast algorithm and compared our results to those of Glasserman and Li (2005). We also show how this methodology can be extended to the t copula, and compare the performance of our estimator to the estimator found in Chan and Kroese (2010). We find that our algorithm performs similarly, however, their methodology requires sorting high dimensional random vector, which is very slow in practice. Thus, as the number of obligors grows, their methodology becomes prohibitively slow which is not the case with our algorithm.

In Chapter 4, we extend the cross-entropy method to the model introduced in Miu and Ozdemir (2006). As opposed to previous models, this model exhibits correlation between the probability of default (PD) and the loss given default (LGD). To the best of our knowledge, there has been no work done on IS in the presence of PD-LGD correlation despite the fact that the PD-LGD correlation is an empirical fact and ignoring it can dramatically underestimate risk measures. We propose to combine cross-entropy minimization with rejection sampling to obtain an estimator that performs extremely well with probabilities as small as 10^{-37}.

Bibliography

Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 1st edition, 2007.

Alexandros Beskos and Gareth O. Roberts. Exact Simulation of Diffusions. *Ann. Appl. Probab.*, 15(4):2422–2444, 2005.

Joshua C.C. Chan and Dirk P. Kroese. Efficient Estimation of Large Portfolio Loss Probabilities in t-Copula Models. *European J. of Oper. Res.*, 205(2):361–367, September 2010.

Nan Chen and Zhengyu Huang. Brownian Meanders, Importance Sampling and Unbiased Simulation of Diffusion Extremes. *Oper. Res. Lett.*, 40(6):554–563, 2012.

Paul Glasserman and Jingyi Li. Importance Sampling for Portfolio Credit Risk. *Management Science*, 51(11):1643–1656, November 2005.

Paul Glasserman, Philip Heidelberger, and Perwez Shahabuddin. Asymptotically Optimal Importance Sampling and Stratification for Pricing Path-Dependent Options. *Math. Finance*, 9:117–152, 1999.

Paolo Guasoni and Scott Robertson. Optimal Importance Sampling With Explicit Formulas in Continuous Time. *Finance Stoch.*, 12:1–19, 2008.

Peter Miu and Bogie Ozdemir. Basel Requirement of Downturn LGD: Modeling and Estimating PD & LGD Correlations. *Journal of Credit Risk*, 2(2):43–68, 2006.

Reuven Y. Rubinstein. Optimization of Computer Simulation Models With Rare Events. *European J. Oper. Res.*, 99(1):89–112, May 1997.

Chapter 2

Rare Event Simulation for Diffusion Processes

2.1 Introduction

The general problem that we address in this chapter is the following[1]

> **Main Problem.** Estimate $\mathbb{E}[\mathbf{h}(S) \cdot \mathbf{1}_{\{S \in B\}}]$ via Monte Carlo, where S is a real-valued diffusion, $\mathbf{h}$ is a path functional and B is a set of continuous functions such that $\mathbb{P}(S \in B)$ is small.

Computationally this problem is difficult for two reasons. First, one can rarely simulate trajectories of S exactly and second, very few simulated trajectories are informative (we say that a simulated trajectory s is informative if it lies in the region of interest, i.e. if $s \in B$). These issues can be mitigated by using a very fine time discretization and simulating a very large number of trajectories but the associated computational cost of achieving an acceptable level of precision can be prohibitive. In this paper we propose a two-stage importance sampling (IS) procedure with each

[1] This chapter was based on a published paper (see Metzler and Scott (2014)).

stage designed to address one of the aforementioned issues.

Section 2.2 describes the algorithm in general terms. The first stage application is reasonably well-developed in the literature (see Beskos and Roberts (2005), DiCesare (2006), DiCesare and McLeish (2008) or Chen and Huang (2012), for example, as well as Giesecke and Smelov (2013) for an extension to jump diffusions) and restates the problem in terms of standard Brownian motion. In a sense this can be considered a non-standard application of IS as it is designed to eliminate bias as opposed to reduce variance. The second stage application is a more standard application of IS and is designed to address the rare event problem by steering trajectories of the standard Brownian motion towards the region of interest via a time-dependent deterministic drift.

The problem of identifying effective importance measures for standard Brownian motion, i.e. the problem faced in the second stage, is underdeveloped in the literature and is the subject of Section 2.3. We propose choosing the drift so that the average path under the importance measure coincides with the average path of the standard Brownian motion over the region of interest. We characterize this choice as the solution to an entropy minimization problem and provide numerical evidence that the resulting estimator compares favourably to an alternative, proposed by Guasoni and Robertson (2008) and based on an elegant appeal to large deviations, that is known to be asymptotically optimal but appears limited to non-negative functionals and is difficult to implement in complex situations.

In Section 2.4 we return to the main problem and implement our proposed two-stage algorithm in the context of more general diffusions. We restrict ourselves to the situation where the functional and region of interest depend only on the extremes and terminal values of the diffusion (the first stage application of IS in this context was considered, outside the rare event regime, by Chen and Huang (2012)) and find

that our proposed algorithm performs well in cases where the region of interest is so rare that crude estimators (i.e. those based only on the first stage application of IS or two-stage IS estimators based on ineffective importance drifts) fail completely.

2.1.1 Notation and Assumptions

Throughout the paper we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and time horizon $T > 0$. W will denote a standard Brownian motion (beginning at zero) on this space and for $\theta \in \mathcal{L}^2[0,T]$ we define W^θ via $W_t^\theta = W_t + \int_0^t \theta_s \, ds$. Thus W^θ is a standard Brownian motion under the measure $\mathbb{P}^\theta$ defined via the Radon-Nikodym derivative

$$\begin{aligned} \Lambda_\theta :=& \tfrac{d\mathbb{P}^\theta}{d\mathbb{P}} \\ =& \exp(-\textstyle\int_0^T \theta_t dW_t - \frac{1}{2}\int_0^T \theta_t^2 dt) \\ =& \exp(-\textstyle\int_0^T \theta_t dW_t^\theta + \frac{1}{2}\int_0^T \theta_t^2 \, dt) \; . \end{aligned} \tag{2.1}$$

S will denote a diffusion constructed as a solution to the SDE

$$dS_t = \mu(S_t)\, dt + \sigma(S_t)\, dW_t \; , \qquad S_0 = s_0 \; . \tag{2.2}$$

We let $\ell \geq -\infty$ denote the left endpoint of the state space of S and $r \leq \infty$ the right endpoint. We assume that σ is strictly positive and twice continuously differentiable on (ℓ, r). For simplicity we also assume that μ is continuous on (ℓ, r), a condition that is satisfied in most cases of practical interest.

The volatility function σ and the initial point $s_0 \in (\ell, r)$ define a Lamperti transform which we denote by $\mathbf{F} : \mathcal{C}[0,T] \to \mathcal{C}[0,T]$ and define as follows. If $Y = \mathbf{F}(S)$ then $Y_t = F(S_t)$, where $F(s) = \int_{s_0}^s [\sigma(u)]^{-1} \, du$. The Lamperti transform of S has unit volatility; it is readily verified that if $Y = \mathbf{F}(S)$ then

$$dY_t = b\,(Y_t)\,dt + dW_t\;, \qquad Y_0 = 0\;, \tag{2.3}$$

where

$$b\,(y) = \left[\mu(F^{-1}(y))/\sigma(F^{-1}(y))\right] - \left[\sigma'(F^{-1}(y))/2\right]\;. \tag{2.4}$$

Our assumptions on μ and σ ensure that b is well-behaved on $(F(\ell), F(r))$, though it may have asymptotes at the endpoints of this interval, and that $\mathbf{F}$ is invertible.

Given a process X, $\overline{X}$ and $\underline{X}$ will denote its running maximum and minimum, respectively. That is $\overline{X}_t = \max_{0\le s\le t} X_s$ and $\underline{X}_t = \min_{0\le s\le t} X_s$. Bold lower case letters $\mathbf{f}$, $\mathbf{g}$, $\mathbf{h}$ will denote real-valued path functionals, i.e. mappings from $\mathcal{C}[0,T]$ to $\mathbb{R}$.

2.2 Proposed Algorithm

In this section we outline our proposed algorithm. In order to avoid delicate issues concerning the nature of the boundaries ℓ and r we restrict ourselves to path functionals $\mathbf{h}(S)$ that can be put in the form

$$\mathbf{h}(S) = \mathbf{f}(S)\cdot \mathbf{1}_{\{\underline{S}_T>\ell,\ \overline{S}_T<r\}}\;, \tag{2.5}$$

for some path functional $\mathbf{f}$. This restriction, together with our assumptions on the drift and volatility functions of S, paves the way for an application of the generalized Girsanov result given in Theorem A.0.1.

2.2.1 First Stage Application - Discretization Error

Recall that our ultimate goal is to estimate $\mathbb{E}[\mathbf{h}(S)\cdot \mathbf{1}_{\{S\in B\}}]$ under the conditions noted in the introduction. In general solutions to (2.2) cannot be simulated exactly,

and this introduces discretization error. This can be mitigated (but never completely eliminated) in an Euler or Milstein scheme by choosing a very fine time step, however the associated computational cost can be substantial; this is especially true in the rare event setting where S lies outside of B with high probability.

In recent years several authors have observed that importance sampling can be used to restate the expectation of interest in terms of processes that *can* be simulated without discretization error. This idea appears to have originated from Beskos and Roberts (2005) and DiCesare (2006) (see Chen and Huang (2012) for a recent, and very clear, description of the methodology in the case where $\mathbf{h}$ depends on the extremes and terminal values of the diffusion), who make the ingenious observation that we can write

$$\begin{aligned}
\mathbb{E}[\mathbf{h}(S)\cdot\mathbf{1}_{\{S\in B\}}] &= \mathbb{E}[\mathbf{g}(Y)\cdot\mathbf{1}_{\{Y\in\mathbf{F}(B)\}}] \\
&= \mathbb{E}[C_\phi(W)\cdot\exp(A(W_T))\cdot\mathbf{g}(W)\cdot\mathbf{1}_{\{W\in\mathbf{F}(B)\}}] \qquad (2.6)\\
&= \mathbb{E}[\hat{C}_\phi(W,N)\cdot\exp(A(W_T))\cdot\mathbf{g}(W)\cdot\mathbf{1}_{\{W\in\mathbf{F}(B)\}}]\,. \qquad (2.7)
\end{aligned}$$

where $\mathbf{g}=\mathbf{h}\circ\mathbf{F}^{-1}$, $\mathbf{F}(B)$ is the image of B under $\mathbf{F}$, $C_\phi(W)=\exp(-\int_0^T\phi(W_t)\,dt)$, $\phi(w)=[b^2(w)+b'(w))]/2$, $A(w)=\int_0^w b(u)du$, N is a homogeneous Poisson process (under $\mathbb{P}$) independent of W with intensity λ and event times $\tau_1<\tau_2<\ldots$, and

$$\hat{C}_\phi(W,N):=\prod_{i=1}^{N_T}\frac{\lambda-\phi(W_{\tau_i})}{\lambda}\,. \qquad (2.8)$$

We call (2.8) the Poisson kernel estimator, terminology which appears to have originated from Chen and Huang (2012).

The identity (2.6) follows from the generalized version of Girsanov's Theorem given in Theorem A.0.1[2] and the identity (2.7) follows from the fact that

[2]In order to apply the theorem first apply Itô's Lemma to the process $A(W_t)$ to find that

$$\mathbb{E}[\hat{C}_\phi(W,N)|\mathcal{F}_T^W] = C_\phi(W)\ ,$$

where $\{\mathcal{F}_t^W : t \in [0,T]\}$ is the filtration generated by W; see Chen and Huang (2012) for more details. In order for the right-hand side of (2.7) to provide the basis for an estimator that is free from discretization error one must be able to simulate the vector

$$(\mathbf{g}(W), W_{t_1}, \ldots, W_{t_n}, W_T) \tag{2.9}$$

for an arbitrary set of times $0 < t_1 < \ldots < t_n < T$ (which in practice will be the simulated Poisson event times), and determined on the basis of this vector whether or not a simulated trajectory lies in $\mathbf{F}(B)$. Depending on the complexity of $\mathbf{h}$, $\mathbf{F}$ and B this can be a formidable task. When the value of $\mathbf{h}$ and membership in B depend only on the extreme values of a given trajectory Chen and Huang (2012) describe an elegant sequential algorithm for simulating (2.9) using Brownian meanders; see Appendix E for more details.

2.2.2 Second Stage Application - Rare Event Problem

Although the identity

$$\mathbb{E}[\mathbf{h}(S)\cdot\mathbf{1}_{\{S\in B\}}] = \mathbb{E}[\hat{C}_\phi(W,N)\cdot\exp(A(W_T))\cdot\mathbf{g}(W)\cdot\mathbf{1}_{\{W\in\mathbf{F}(B)\}}] \tag{2.10}$$

provides a means for eliminating discretization error, and therefore produces an unbiased estimator, it typically does not address the rare event problem. Indeed as we

---

$$C_\phi(W)\cdot\exp(A(W_T)) = \exp(\textstyle\int_0^T b(W_t)dW_t - \frac{1}{2}\int_0^T [b(W_t)]^2 dt)\ ,$$

which is effectively the Radon-Nikodym derivative between the measures induced by W and Y. Also note that $\mathbf{g}(W)$ has the special form $\mathbf{g}(W) = \mathbf{h}(\mathbf{F}^{-1}(W))\cdot\mathbf{1}_{\{\underline{W}_T > F(\ell),\, \overline{W}_T < F(r)\}}$.

will see in Section 2.4 if the event $\{S \in B\}$ is sufficiently rare then estimators based on (2.10) can fail completely.

To this end we propose a second application of IS based on the identities

$$\begin{aligned}\mathbb{E}[\mathbf{h}(S) \cdot \mathbf{1}_{\{S \in B\}}] &= \mathbb{E}[\hat{C}_\phi(W, N) \cdot \exp(A(W_T)) \cdot \mathbf{g}(W) \cdot \mathbf{1}_{\{W \in \mathbf{F}(B)\}}] \\ &= \mathbb{E}_\theta[\hat{C}_\phi(W, N) \cdot \Lambda_\theta^{-1} \cdot \exp(A(W_T)) \cdot \mathbf{g}(W) \cdot \mathbf{1}_{\{W \in \mathbf{F}(B)\}}] \\ &= \mathbb{E}[\hat{C}_\phi(W^\theta, N) \cdot \Lambda_\theta^{-1} \cdot \exp(A(W_T^\theta)) \cdot \mathbf{g}(W^\theta) \cdot \mathbf{1}_{\{W^\theta \in \mathbf{F}(B)\}}] \,,\end{aligned}$$

which follow directly from Girsanov's Theorem. We restrict ourselves to deterministic importance drifts because in that case Radon-Nikodym derivatives can be simulated exactly[3].

In order for this change of measure to be effective the second moment of

$$\hat{C}_\phi(W^\theta, N) \cdot \Lambda_\theta^{-1} \cdot \exp(A(W_T^\theta)) \cdot \mathbf{g}(W^\theta) \cdot \mathbf{1}_{\{W^\theta \in \mathbf{F}(B)\}} \tag{2.11}$$

must be small. An explicit expression for the variance-minimizing θ is clearly beyond reach here and so alternative criteria are necessary. To begin first note that the variability of $\hat{C}_\phi(W^\theta, N)$ can be controlled by choosing a sufficiently large λ (though increasing λ does come at a cost of increased computational time, as noted by Chen and Huang (2012)). Thus it seems reasonable to focus efforts on minimizing the second moment of

$$C_\phi(W) \cdot \Lambda_\theta^{-1} \cdot \exp(A(W_T^\theta)) \cdot \mathbf{g}(W^\theta) \cdot \mathbf{1}_{\{W^\theta \in \mathbf{F}(B)\}} \,. \tag{2.12}$$

Choosing θ in such a way as to reduce the variability of (2.12) is effectively the main problem specialized to the case where S is standard Brownian motion, and in Section 2.3 we consider this problem in detail. Our ultimate proposal will be to set

[3]Recall that if $\theta \in \mathcal{L}^2[0, T]$ then $\int_0^T \theta_t dW_t$ is normal with mean zero and variance $\int_0^T \theta_t^2 \, dt$.

$\theta_t = \frac{d}{dt}\mathbb{E}[W_t|W \in \mathbf{F}(B)]$, a choice that is motivated, justified and explored in the following section.

Before moving on we note that Section 2.3 considers the optimal choice of $\theta \in \mathcal{L}^2[0,T]$. In order to implement estimators based on (2.11) without discretization error, one must be able to simulate the vector

$$(\mathbf{g}(W^\theta), W^\theta_{t_1}, \ldots, W^\theta_{t_n}, W^\theta_T, \Lambda_\theta) \tag{2.13}$$

for an arbitrary set of times $0 < t_1 < \ldots < t_n < T$, and determine on the basis of this vector whether or not a simulated trajectory lies in $\mathbf{F}(B)$. As noted in the previous section this is in general quite formidable and in many cases one must insist that θ lie in some feasible subset of $\mathcal{L}^2[0,T]$ such as constant or piecewise constant functions.

2.3 Rare Event IS for Brownian Motion

In this section we consider the problem of selecting an effective importance measure for estimating $\mathbb{E}[\mathbf{h}(W)\cdot\mathbf{1}_{\{W\in B\}}]$, where $\mathbf{h}:\mathcal{C}[0,T]\to\mathbb{R}$ and $B\subset\mathcal{C}[0,T]$ is rare in the sense that $\mathbb{P}(W\in B)$ is small. In the interest of tractability we restrict attention to importance measures from the family $\{\mathbb{P}^\theta :\ \theta\in\mathcal{L}^2[0,T]\}$, since in this case Λ_θ can be simulated exactly. Thus our IS estimators will be based on the identities

$$\mathbb{E}[\mathbf{h}(W)\cdot\mathbf{1}_{\{W\in B\}}] = \mathbb{E}_\theta[\mathbf{h}(W)\cdot\Lambda_\theta^{-1}\cdot\mathbf{1}_{\{W\in B\}}] = \mathbb{E}[\mathbf{h}(W^\theta)\cdot\Lambda_\theta^{-1}\cdot\mathbf{1}_{\{W^\theta\in B\}}]\,.$$

The basic problem is then to choose θ so as to minimize the second moment of $\mathbf{h}(W^\theta)\cdot\Lambda_\theta^{-1}\cdot\mathbf{1}_{\{W^\theta\in B\}}$.

Despite its apparent simplicity this measure-selection problem is underdeveloped in the literature and to the best of our knowledge Guasoni and Robertson (2008) is the

only paper that explicitly considers a similar problem (though it does receive some attention in Asmussen and Glynn (2007)). In Guasoni and Robertson (2008) the authors consider non-negative functionals $\mathbf{h}$ and develop a criteria for selecting θ, henceforth referred to as the GR criteria, that is based on an elegant application of large deviations and is known to be asymptotically optimal in a certain sense. Unfortunately the criteria does not appear amenable to the problem described in Section 2.2 since it appears limited to non-negative functionals and, even when applicable, it appears difficult to determine explicit expressions for the optimal θ.

In Section 2.3.1 we provide a brief overview of the GR criteria and in Section 2.3.2 we propose a simpler and apparently more general criteria that selects θ in order to ensure that the averages trajectory under the importance measure coincides with that under the conditional law of W, given that $W \in B$, and characterize this choice as the solution to an entropy-minimization problem. In Section 2.3.3 we compare the two criteria using an example from Guasoni and Robertson (2008), finding that the two criteria are asymptotically equivalent and provide near-optimal variance reduction.

2.3.1 The Guasoni and Robertson (GR) Criteria

The criteria proposed by Guasoni and Robertson (2008) deals with non-negative $\mathbf{h}$ and is to select the importance drift θ so as to maximize, if possible, the quantity

$$2\log\left(\mathbf{h}(\Theta)\cdot\mathbf{1}_{\{\Theta\in B\}}\right)-\int_0^T[\Theta'_t]^2\,dt\;, \tag{2.14}$$

where $\Theta_t=\int_0^t\theta_s ds$ (see Guasoni and Robertson (2008) for justification and intuition about (2.14)). The authors provide sufficient conditions for this problem to have a well-defined solution and for that solution to be characterized as a solution to the following variational problem

$$D[\log \mathbf{h}(\Theta)] + \Theta'' = 0 \,, \qquad \Theta \in B \,, \tag{2.15}$$

where $D[\mathbf{g}]$ denotes the Frechet derivative of $\mathbf{g}$.

In general the indicated conditions can be quite difficult to verify and, even when they can be verified, (2.15) can be quite difficult to solve. In Guasoni and Robertson (2008) the authors do provide one example that admits an explicit solution, which we will eventually use as a benchmark in Section 2.3.3.

Example 2.3.1. *In Example 4.1 of Guasoni and Robertson (2008) the authors consider the pricing of geometric Asian options in the Black-Scholes model. This leads to a functional of the form* $\mathbf{h}(x) = b[\exp(a\int_0^T x_t\,dt) - c]$ *and a region of interest of the form* $B = \{x \in \mathcal{C}[0,T] : \int_0^T x_t dt \geq \log(c)/a\}$*, for positive constants* a, b, c*. As discussed in Guasoni and Robertson (2008) the corresponding Euler-Lagrange equation is*

$$\Theta_t'' = -\beta \,, \qquad \beta = \frac{a\exp(a\int_0^T \Theta_t dt)}{\exp(a\int_0^T \Theta_t dt) - c} \,,$$

which has quadratic solution $\Theta_t = -\hat{\beta}t^2/2 + \hat{\beta}Tt$*, where* $\hat{\beta} > a$ *is the unique solution in* β *to* $a\beta T^3 + 3\log[(\beta - a)/c\hat{\beta}] = 0$ *and must in general be determined numerically. Thus the optimal drift according to the GR criteria is of the linear form* $\theta_t = \hat{\beta}(T - t)$*.*

2.3.2 Entropy Minimization

In this section we propose an alternative measure-selection criteria. Our proposal is motivated by the success of Asmussen et al. (2005), who applied a similar idea in the context of using IS to estimate tail probabilities of sums of heavy-tailed random variables.

Let $\mathbb{P}(W \in B) > 0$, and let $\widetilde{\mathbb{P}}$ be the conditional measure defined as

$$\widetilde{\mathbb{P}}(A) := \mathbb{P}(A|W \in B).$$

Here we note that the measure induced by $\widetilde{\mathbb{P}}$ on $\mathcal{C}[0,T]$ will be the conditional law of W, given that its trajectory lies in the region of interest. Our proposal is to choose that member of $\mathcal{L}^2[0,T]$ that leads to an importance measure that is "as close" to $\widetilde{\mathbb{P}}$ as possible in the sense of the Kullback-Leibler divergence[4]

$$d(\theta) := \widetilde{\mathbb{E}}[\log(d\widetilde{\mathbb{P}}/d\mathbb{P}^\theta)] \ .$$

In other words our proposal is to use $\mathbb{P}^{\widetilde{\theta}}$ as an importance measure, where $\widetilde{\theta} = \arg\min_{\theta \in \mathcal{L}^2[0,T]} \ d(\theta)$. This optimization problem is based on the well-known CE method, and admits an explicit and intuitive solution under very mild conditions is confirmed in Proposition 2.3.2.

Proposition 2.3.2. *Let $h_t = \mathbb{E}\left[W_t \mid W \in B\right]$ under $\mathbb{P}$. If h_t is twice differentiable with square-integrable first derivative then $d(\theta)$ is minimized by setting $\theta = h'$.*

Proof. The Radon-Nykodym derivative of $\widetilde{\mathbb{P}}$, with respect to $\mathbb{P}$, is demonstrably

$$\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}} = \frac{\mathbf{1}_{\{W \in B\}}}{\mathbb{P}(W \in B)} \ .$$

Thus

$$\begin{aligned} d(\theta) &= \widetilde{\mathbb{E}}[\log(d\widetilde{\mathbb{P}}/d\mathbb{P}^\theta)] \\ &= \widetilde{\mathbb{E}}[\log(d\widetilde{\mathbb{P}}/d\mathbb{P}) + \log(d\mathbb{P}/d\mathbb{P}^\theta)] \\ &= -\log(\mathbb{P}(W \in B)) + \mathbb{E}[(d\widetilde{\mathbb{P}}/d\mathbb{P})\log(d\mathbb{P}/d\mathbb{P}^\theta)] \\ &= -\log(\mathbb{P}(W \in B)) - \mathbb{E}[\textstyle\int_0^T \theta_t \, dW_t | W \in B] + \frac{1}{2}\int_0^T \theta_t^2 \, dt \ . \end{aligned}$$

[4] $d(\theta)$ is often referred to as the Kullback-Leibler divergence of $\mathbb{P}^\theta$ from $\widetilde{\mathbb{P}}$ and is directly related to the cross entropy between $\widetilde{\mathbb{P}}$ and $\mathbb{P}^\theta$.

By Itô's Lemma we have $\theta_T W_T = \int_0^T \theta_t \, dW_t + \int_0^T \theta_t' \, W_t dt$, hence in order to minimize $d(\theta)$ it suffices to minimize

$$\begin{aligned} -\mathbb{E}[\textstyle\int_0^T \theta_t \, dW_t | W \in B] + \frac{1}{2}\int_0^T \theta_t^2 \, dt &= \int_0^T \theta_t' h_t \, dt - \theta_T h_T + \frac{1}{2}\int_0^T \theta_t^2 \, dt \\ &= -\int_0^T \theta_t h_t' \, dt + \frac{1}{2}\int_0^T \theta_t^2 \, dt \\ &= \frac{1}{2}\int_0^T [\theta_t - h_t']^2 \, dt - \frac{1}{2}\int_0^T [h_t']^2 \, dt \ , \end{aligned}$$

which clearly attains its minimum when $\theta = h'$. □

Remark 2.3.3. *The average trajectory of W under $\mathbb{P}^\theta$ is given by the function $t \mapsto \int_0^t \theta_s ds$. Thus our proposed criteria is equivalent to ensuring that the average trajectories under $\widetilde{\mathbb{P}}$ and $\mathbb{P}^\theta$ coincide.*

In many cases of practical interest the expectation $h_t = \mathbb{E}[W_t|B]$ can be computed explicitly.

Example 2.3.4. *If $B = \{x \in \mathcal{C}[0,T] : \int_0^T x_t \geq K\}$ then one can use the joint distribution of W_t and $\int_0^T W_t \, dt$, which is bivariate normal, to find that*

$$h_t = \sqrt{3/T^3}[\phi(-\sqrt{3/T^3}K)/\Phi(-\sqrt{3/T^3}K)]t(T - t/2) \ , \tag{2.16}$$

which is quadratic in t. Thus the optimal drift according to our criteria will in this case be linear. Using the well-known asymptotic approximation $\Phi(-x) \sim \phi(x)/x$ as $x \to \infty$ we see that when K is large we will have, for fixed t, $h_t \sim (3K/T^3)t(T - t/2)$ as $K \to \infty$. Thus we will have $\int_0^T h_t \, dt \approx K$ for large K and we see that our entropy criteria effectively places the average importance path on the boundary of B.

Example 2.3.5. *If $B = \{x \in \mathcal{C}[0,T] : \overline{x}_T \geq a\}$, where $\overline{x}_T := \max_{t\in[0,T]} x_t$ and $a > 0$, one can use the results in Appendix B.1 to compute h_t numerically. It is always the case that $h_T = a$ and if a is sufficiently large numerical evidence indicates that h is very nearly linear. Once again we see that our entropy criteria effectively places the*

average importance path on the boundary of B.

Example 2.3.6. *If $B = \{x \in \mathcal{C}[0,T] : \overline{x}_T \geq a,\ \underline{x}_T \leq -b\}$, where $\underline{x}_T := \min_{t\in[0,T]} x_t$ and $b > 0$, one can use the results in Appendix B.2 to compute h_t numerically. Figure 2.1 plots the function h for several (a, b) pairs and we see that the average path strikes the closer boundary first and terminates very close to the more remote barrier. In this case h is well-approximated by a piecewise linear function having two pieces.*

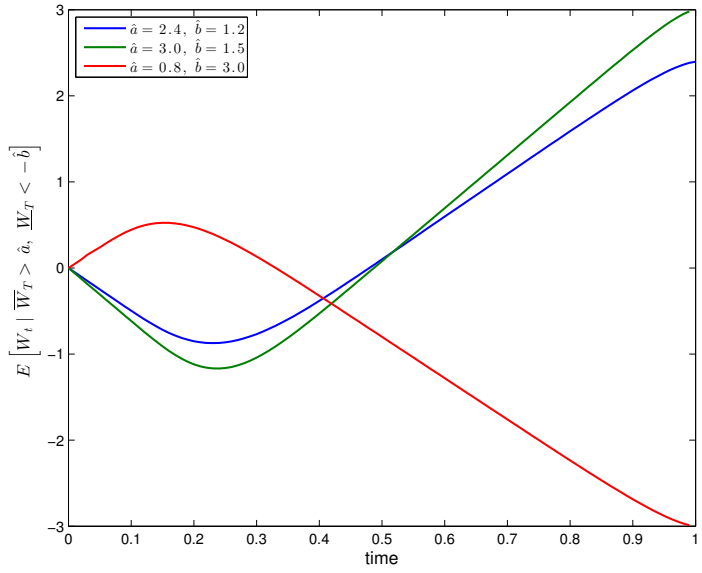

Figure 2.1: *Average path $h_t := \mathbb{E}[W_t|\overline{W}_T > a, \underline{W}_T < -b]$ against t. This expected value is computed based on the discussion in Appendix B.2. A piecewise linear approximation to this function will be used as an importance drift in Example 2.4.2.*

In many cases it is difficult or impossible to implement the two-stage IS algorithm using drifts that are even moderately complicated. As such it is important to consider importance drifts that minimize $d(\theta)$ over a tractable subset of $\mathcal{L}^2[0,T]$. Proposition 2.3.7 below identifies the optimal importance drift, according to the entropy criteria, if one restricts oneself to piecewise constant drifts. It is interesting to note that the average path under the optimal importance measure will be piecewise linear, indeed it will simply be a linear interpolation of the ideal path $h_t = \mathbb{E}[W_t|W \in B]$.

Proposition 2.3.7. *Consider a fixed set of times $0 = t_0 < t_1 < \ldots < t_M = T$. If one restricts oneself to piecewise constant functions of the form*

$$\theta_t = \sum_{i=1}^{N} \theta^{(i)} \mathbf{1}_{\{t_{i-1} < t \leq t_i\}} \ , \qquad (\theta^{(1)}, \ldots, \theta^{(M)}) \in \mathbb{R}^M \ ,$$

then $d(\theta)$ is minimized by setting

$$\theta^{(i)} = [h_{t_i} - h_{t_{i-1}}]/[t_i - t_{i-1}] \ ,$$

where h_t is defined in Proposition 2.3.2.

Proof. In this case we must minimize

$$\int_0^T [\theta_t - h'_t]^2 \, dt = \sum_{i=1}^M \int_{t_{i-1}}^{t_i} [\theta^{(i)} - h'_t]^2 \, dt \ .$$

Now

$$\int_{t_{i-1}}^{t_i} [\theta^{(i)} - h'_t]^2 dt = [\theta^{(i)}]^2 (t_i - t_{i-1}) - 2\theta^{(i)} [h_{t_i} - h_{t_{i-1}}] + \int_{t_{i-1}}^{t_i} [h'_t]^2 \, dt \ ,$$

which is clearly minimized by setting $\theta^{(i)} = [h_{t_i} - h_{t_{i-1}}]/[t_i - t_{i-1}]$. □

2.3.3 Comparison Between Minimum Entropy and GR

In this section we benchmark the performance of our proposed criteria against the performance of the GR criteria using the problem described in Example 2.3.1. Recall that this problem appears to be one of the rare instances that one can explicitly identify the optimal GR drift, and that it involves a functional of the form $\mathbf{h}(x) = b[\exp(a \int_0^T x_t \, dt) - c]$ and a region of interest of the form $B = \{x \in \mathcal{C}[0,T] : \int_0^T x_t dt \geq \log(c)/a\}$. For a fixed value of a the parameter c dictates how rare the event of interest is.

Both criteria here lead to linear importance drifts of the form $\theta_t = \alpha(T - t)$ for some

constant $\alpha \in \mathbb{R}$. As discussed in Example 2.3.1 the GR criteria leads to the unique value of $\hat{\alpha}$ for which

$$a\hat{\alpha}T^3 + 3\log[(\hat{\alpha} - a)/c\hat{\alpha}] = 0 \ ,$$

while (2.16) shows that the entropy criteria leads to the constant

$$\widetilde{\alpha} = \sqrt{3/T^3}\phi(-\sqrt{3/T^3}\log(c)/a)/\Phi(-\sqrt{3/T^3}\log(c)/a) \ .$$

As illustrated in Figure 2.2 the discrepancy between $\hat{\alpha}$ and $\widetilde{\alpha}$ appears to vanish as the event of interest becomes increasingly rare, i.e. as c increases without bound.

Remark 2.3.8. *Both $\hat{\alpha}$ and $\widetilde{\alpha}$ are asymptotic, as $c \to \infty$, to $\alpha^* = 3\log(c)/aT^3$. Indeed using the well-known asymptotic relation $(1 - \Phi(x)) \sim \phi(x)/x$ as $x \to \infty$ it is trivial to verify that $\widetilde{\alpha} \sim \alpha^*$ as $c \to \infty$, to α^*. To see that $\hat{\alpha}$ is asymptotic to the same value re-write the defining equation as*

$$\hat{\alpha} + (3/aT^3)\log(1 - (a/\hat{\alpha})) = 3\log(c)/aT^3 \ .$$

Since $\hat{\alpha} > a > 0$ it follows that $\hat{\alpha} > 3\log(c)/aT^3$, which means that $\hat{\alpha}$ will grow without bound as c does. Since a is fixed it is now obvious that $\hat{\alpha} \sim \alpha^*$. Note that the algorithm still performs extremely well for events that are not that rare.

Having established that the two criteria will be asymptotically identical for truly rare events we now consider their performance with regard to variance reduction. To this end, for $\alpha \in \mathbb{R}$ define $v(\alpha)$ as the coefficient of variation of the IS estimator that uses drift $\theta_t = \alpha(T - t)$. As illustrated in Figure 2.3 both $\hat{\alpha}$ and $\widetilde{\alpha}$ are very close to the optimal value $\arg\min_{\alpha\in\mathbb{R}} v(\alpha)$, indicating that among all IS estimators using linear drift, both criteria lead to very nearly optimal IS estimators. In other words it appears that, for sufficiently large c, $\alpha^* \approx \arg\min_{\alpha\in\mathbb{R}} v(\alpha)$.

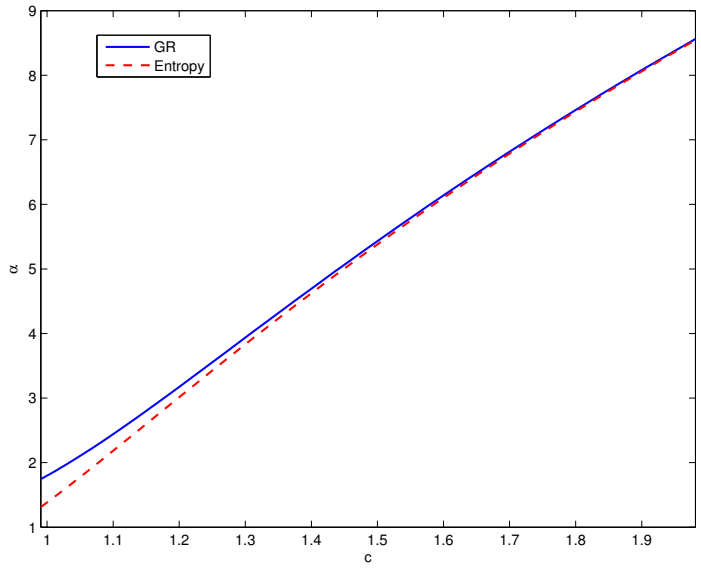

Figure 2.2: *This figure compares the constants $\hat{\alpha}$, the slope of the linear drift selected by the GR criteria, and $\widetilde{\alpha}$, the slope selected by the entropy criteria, in the context of Example 2.3.1 b, a* and *T are respectively* 1.4, 0.25 and 1 *in this example. As c varies along the horizontal axis here the probability $\mathbb{P}(W \in B)$ varies from* 0.5 *to* 4.356×10^{-6}.

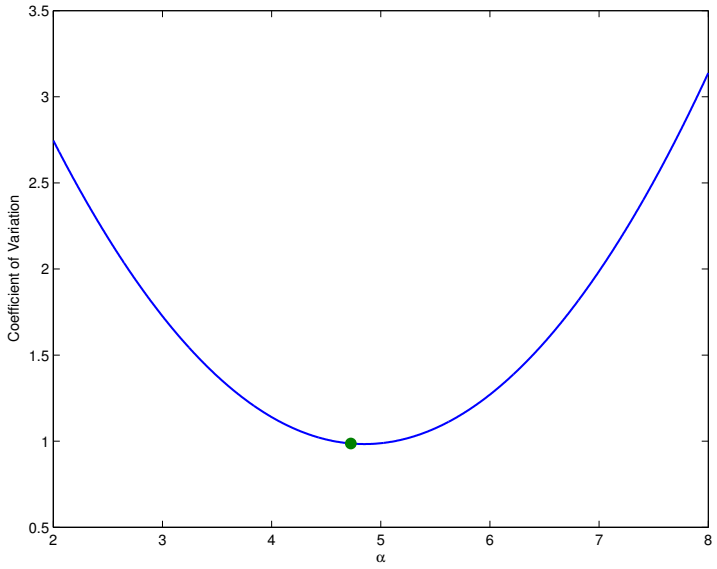

Figure 2.3: *This figure illustrates a quadratic fit to the function $v(\alpha)$, where $v(\alpha)$ is defined as the coefficient of the IS estimator for Example 2.3.1 that uses linear drift $\theta_t = \alpha(T-t)$. Parameter values are $c = 1.3869$, $b = 1.4$, $a = 0.25$ and $T = 1$ which leads to $\mathbb{P}(\int_0^T W_t dt \in B) \approx 0.0117$. We ran one simulation of size 10^6 for each value of α in a very fine equally-spaced grid, estimated the coefficient of variation and then fit a quadratic to the resulting data. There are actually two solid dots in the plot corresponding to the fitted (and estimated) values of $v(\widetilde{\alpha})$ and $v(\hat{\alpha})$; since the two constants are so close the dots cannot be distinguished by the nake eye.*

2.4 Functions of Extremes

In this section we return to the main problem and implement our two-stage IS algorithm for more general diffusions. In order to clearly demonstrate the utility of the second stage application we concentrate on functionals for which the first stage application is well-developed. In particular we restrict ourselves to functionals of the form

$$\mathbf{h}(S) = f(S_T, \overline{S}_T, \underline{S}_T) \cdot \mathbf{1}_{\{\overline{S}_T < r,\, \underline{S}_T > \ell\}} =: h(\overline{S}_T, \underline{S}_T, S_T) \ ,$$

where $f : \mathbb{R}^3 \mapsto \mathbb{R}$ is some function, and events of interest are of the form

$$B_{a,b} = \{x \in \mathcal{C}[0,T] : \ \overline{x}_T > a, \ \underline{x}_T < -b\}$$

where $a, b \in \mathbb{R}$. We define $B_a^+ = B_{a,-\infty}$ as the set of paths whose maximum exceeds the level a and $B_b^- = B_{-\infty,b}$ is the set whose minimum is below $-b$. See Chen and Huang (2012) for an elegant algorithm that can be used to implement the first stage of IS in this context.

Our estimator here is based on the fact that $\mathbb{E}[h(\overline{S}_T, \underline{S}_T, S_T) \cdot \mathbf{1}_{\{S \in B_{a,b}\}}]$ is equal to

$$\mathbb{E}[\exp(A(W_T^\theta)) \cdot \hat{C}_\phi(W^\theta, N) \cdot \Lambda_\theta^{-1} \cdot g(\overline{W}_T^\theta, \underline{W}_T^\theta, W_T^\theta) \cdot \mathbf{1}_{\{W \in B_{F(a),F(b)}\}}] \ ,$$

where $g(\overline{w}, \underline{w}, w) = h(F^{-1}(\overline{w}), F^{-1}(\underline{w}), F^{-1}(w))$. When $\theta = 0$ we refer to the resulting estimator as the "crude estimator"; in other words the crude estimator only involves the first stage application of IS and is not designed to address the rare event problem.

In order to implement this estimator one must be able to simulate

$$(\overline{W}_T^\theta, \underline{W}_T^\theta, W_{t_1}^\theta, \ldots, W_{t_n}^\theta, W_T^\theta, \Lambda_\theta)$$

for an arbitrary set of times $0 < t_1 < \ldots < t_n < T$. This does not appear possible for the optimal drift $\theta = h'$ determined in Proposition 2.3.2 and so we must content ourselves with selecting piecewise constant drifts using Proposition 2.3.7.

In the remainder of this section we describe the zero-drift algorithm presented by Chen and Huang (2012) and show how it can be generalized to the case of piecewise constant drift. In principle this allows the user to implement the two-stage estimator using drifts that are arbitrarily close to the optimal drift $\theta = h'$ determined in Proposition 2.3.2.

Zero Drift

When $\theta_t \equiv 0$ we have $\Lambda^{-1} = 1$ and, given simulated Poisson event times $0 < t_1 < \ldots < t_n$, we need only simulate

$$(\overline{W}_T, \underline{W}_T, W_{t_1}, \ldots, W_{t_n}, W_T) \, .$$

One way to accomplish this is to first simulate the triplet $(\overline{W}_T, W_T, \xi_T)$, where $\xi_T = \sup\{t \in [0,T] : \ W_t = \overline{W}_T\}$ is the temporal location of the maximum of W, and then make use of the fact that conditional on this triplet the trajectory of W can be decomposed into two independent Brownian meanders (one on either side of ξ_T); see Chen and Huang (2012) for a description. Skeletons of such processes can be simulated using results from Devroye (2009) and the minimum of a meander over $[t_{i-1}, t_i]$, conditional on its values at the endpoints, can be simulated using the inverse transform algorithm described at the end of Section 4.1 in Chen and Huang (2012). The global minimum $\underline{W}_T$ can then be determined from the local minima in the obvious way.

There are a variety of ways to carry out this general idea, and Algorithm 1 in Appendix

E provides a detailed outline of one such implementation. It follows the algorithm developed in Chen and Huang (2012) quite closely but uses a different method for simulating the triplet $(\overline{W}_T, W_T, \xi_T)$. In particular we generate the components of this triplet in a different order; in Chen and Huang (2012) ξ_T is generated first using inverse transform but it does not seem feasible to do so in the presence of drift, as such we generate the triplet in an order that is more amenable to the introduction of non-zero drift.

Remark 2.4.1. *In the interest of computational efficiency Steps 4 to 12 in Algorithm 1 should only be performed if $\overline{W}_T > F(a)$. This is because (i) generation of ξ_T uses acceptance-rejection and (ii) generation of the local minima uses numerical inverse transform. These steps can be expensive and therefore should only be carried out if absolutely necessary.*

Constant Drift

In this case we have, with an admitted abuse of notation, $\theta_t \equiv \theta$ for some constant $\theta \in \mathbb{R}$. Since $\Lambda_\theta = \exp(-\theta W_T^\theta + \theta^2 T/2)$ it suffices to simulate

$$(\overline{W}_T^\theta, \underline{W}_T^\theta, W_{t_1}^\theta, \ldots, W_{t_n}^\theta, W_T^\theta) \, .$$

In Appendix A we prove (see Theorem A.0.2) that the conditional law of W^θ, given the triplet $(\overline{W}_T^\theta, W_T^\theta, \xi_T^\theta)$, does not depend on the value of θ. Therefore if one is able to simulate the triplet one can then assume without loss of generality that $\theta = 0$ and proceed as in the case of zero drift by making use of independent meanders on either side of ξ_T^θ.

Algorithm 2 provides a detailed algorithm for implementing this idea. Note that steps 5 through 10 are the most expensive steps and should be performed only if necessary.

Piecewise Constant Drift

Here we fix a set of times $0 = s_0 < s_1 < \ldots < s_m = T$ and assume that θ_t takes on the constant value $\theta^{(i)} \in \mathbb{R}$ over the interval $(s_{i-1}, s_i]$; for completeness assume $\theta_0 = 0$. The basic idea is to begin by creating m independent Brownian bridges by first simulating $(W^{\theta}_{s_1}, \ldots, W^{\theta}_{s_m})$. If one is then able to generate the maximum value and its temporal location for each bridge (the temporal location of the i^{th} bridge will lie in $[s_{i-1}, s_i]$) then, since the law of the i^{th} bridge will not depend on $\theta^{(i)}$, one can use the decomposition into Brownian meanders over each subinterval $[s_{i-1}, s_i]$. A detailed algorithm for implementing this procedure is provided in Algorithm 3.

2.4.1 Example - CIR Process With Large Maximum

In this example we illustrate how effective the second stage IS application can be when the event of interest is truly rare. In particular we compare the performance of the crude estimator (i.e. the IS estimator using only the first stage) to the two-stage estimator using constant drift and find that the latter performs admirably in cases where the former fails completely. We consider estimating

$$\mathbb{E}[h(\overline{S}_T, \underline{S}_T, S_T) \cdot \mathbf{1}_{\{S \in B_a^+\}}] = \mathbb{E}[h(\overline{S}_T, \underline{S}_T, S_T) \cdot \mathbf{1}_{\{\overline{S}_T > a\}}] \tag{2.17}$$

where S_t denotes a so-called CIR process driven by the SDE

$$dS_t = \kappa\,(\alpha - S_t)\,dt + \sigma\sqrt{S_t}dW_t, \quad S_0 = s_0\ ,$$

and we recall that $h(\overline{s}, \underline{s}, s)$ is of the form $f(\overline{s}, \underline{s}, s) \cdot \mathbf{1}_{\{\underline{s}>0\}}$ for some function $f : \mathbb{R}^3 \mapsto \mathbb{R}$.

The crude estimator is based on the fact that (2.17) is equal to

$$\mathbb{E}[f(\overline{W}_T, \underline{W}_T, W_T) \cdot \exp(A(W_T)) \cdot \hat{C}_\phi(W, N) \cdot \mathbf{1}_{\{\overline{W}_T > F(a),\, \underline{W}_T > F(0)\}}] \,,$$

where $F(x) = (2/\sigma) \cdot (\sqrt{x} - \sqrt{s_0})$,

$$\phi(x) = \tfrac{1}{2}\left[\left(\tfrac{4\kappa\alpha-\sigma^2}{2\sigma^2(x+2\sqrt{s_0}/\sigma)} - \tfrac{\kappa}{2}\left(x + \tfrac{2\sqrt{s_0}}{\sigma}\right)\right)^2 - \tfrac{4\kappa\alpha-\sigma^2}{2\sigma^2(x+2\sqrt{s_0}/\sigma)^2} - \tfrac{\kappa}{2}\right] \,,$$

and

$$A(x) = \tfrac{4\kappa\alpha-\sigma^2}{2\sigma^2} \log\left(x + \tfrac{2\sqrt{s_0}}{\sigma}\right) - \tfrac{\kappa}{4}\left(x + \tfrac{2\sqrt{s_0}}{\sigma}\right)^2 .$$

If a is large then the event of interest $\{\overline{W}_T > F(a)\}$ is rare and one might expect this crude estimator to perform poorly.

The second stage application of IS is based on the fact that (2.17) is equal to

$$\mathbb{E}[f(\overline{W}^\theta_T, \underline{W}^\theta_T, W^\theta_T) \cdot \exp(A(W^\theta_T)) \cdot \hat{C}_\phi(W^\theta, N) \cdot \Lambda_\theta^{-1} \cdot \mathbf{1}_{\{\overline{W}^\theta_T > F(a),\, \underline{W}^\theta_T > F(0)\}}]$$

and for our numerical example we use the constant drift

$$\theta = T^{-1} \cdot \mathbb{E}[W_T | \overline{W}_T > F(a)] = F(a)/T \,,$$

which is the optimal constant drift according to our entropy-based optimality criteria (see Proposition 2.3.7).

In Figure 2.4 we compare the performance of the two estimators in the case that $f(\overline{s}, \underline{s}, s) \equiv 1$, which corresponds to simply estimating $\mathbb{P}(\overline{S}_T > a)$. Ideally we could plot the estimated coefficient of variation against $\mathbb{P}(\overline{S}_T > a)$ as a varies, as this would tell us how the estimator performs as a function of how rare the event of interest is. Unfortunately we do not have this in closed-form and so we plot the estimated coefficient of variation against $-\log_{10}(\mathbb{P}(\overline{W}_T > F(a)))$, which is available in closed-

form and is indicative of how rare the event of interest is. For each value of a in our grid the coefficient of variation is based on an independent sample of size $2 \cdot 10^6$ with $\lambda = 10$ (we found that using values of λ in excess of 10 did not appreciably impact the performance of the estimator but did increase CPU time substantially). We see that the two-stage IS procedure drastically outperforms the crude estimator and performs quite well in cases where the crude estimator fails completely, as it appears to occur for probabilities on the order of 10^{-6}.

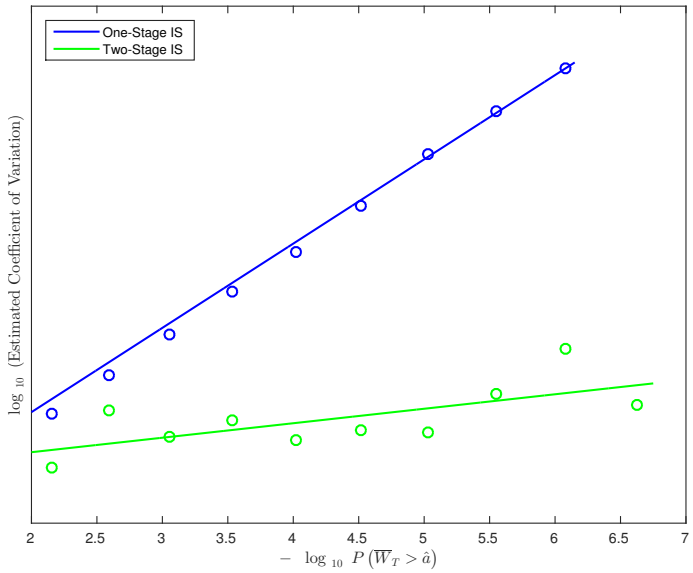

Figure 2.4: *Estimated coefficient of variation against* $\mathbb{P}(\overline{W}_T > \hat{a})$ *on the* log *scale. We are using* $2,000,000$ *trajectories when* $f(\overline{S}_T, \underline{S}_T, S_T) \equiv 1$ *with parameters* $\sigma = 0.15$, $\kappa = 0.5$, $\alpha = 0.06$, $T = 1$ *and* $s_0 = 0.06$. *We fit an exponential curve for the crude estimator and a linear curve for the two-stage estimator (the two-stage estimator uses constant drift).*

In Figure 2.5 we compare performance in the case that $f(\overline{s}, \underline{s}, s) = s^2$, which corresponds to estimating $\mathbb{E}[S_T^2 \cdot \mathbf{1}_{\{\overline{S}_T > a\}}]$. The plot is produced in the same manner as Figure 2.4 and tells an identical story, namely that the addition of the second stage of IS allows the estimator to perform at an acceptable level of precision in cases where the crude estimator fails completely. Applying a second stage of Importance Sampling is almost cost free since we restricted ourselves to deterministic IS drifts and since sampling from $(\overline{W_T}^\theta, \underline{W_T}^\theta, \xi_T^\theta)$ can be done very quickly.

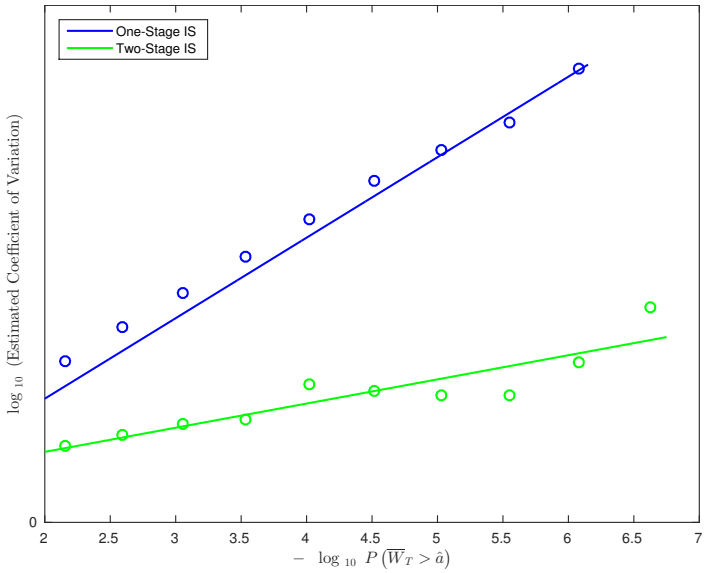

Figure 2.5: *Estimated coefficient of variation against* $\mathbb{P}(\overline{W}_T > \hat{a})$ *on the* log *scale. We are using* $2,000,000$ *trajectories when* $f(\overline{S}_T, \underline{S}_T, S_T) = S_T^2$ *with parameters* $\sigma = 0.15,\ \kappa = 0.5,\ \alpha = 0.06,\ T = 1$ *and* $s_0 = 0.06$*. We fit an exponential curve and a linear curve for he First-Stage and the Two-Stage Importance Sampling estimator, respectively.*

2.4.2 Example - OU Process With Large Maximum and Low Minimum

In the previous example the optimal importance drift, when optimized over the whole of $\mathcal{L}^2[0,T]$, is very nearly constant. As such we were able to obtain substantial variance reduction using a very simple importance drift. In this section we consider an example where the optimal drift is non-linear in order to demonstrate the importance of selecting an importance drift that captures the salient features of the optimal (entropy-based) IS drift.

To this end we consider estimation of

$$\mathbb{E}[h(\overline{S}_T, \underline{S}_T, S_T) \cdot \mathbf{1}_{\{S \in B_{a,b}\}}] = \mathbb{E}[h(\overline{S}_T, \underline{S}_T, S_T) \cdot \mathbf{1}_{\{\overline{S}_T > a,\ \underline{S}_T < -b\}}]\,, \tag{2.18}$$

where S_t is an Ornstein-Uhlenbeck process driven by the SDE

$$dS_t = \kappa\left(\alpha - S_t\right)dt + \sigma dW_t, \quad S_0 = s_0.$$

Since the boundaries $\pm\infty$ are unattainable here there is no implicit indicator in h. In this OU case we have $F(s) = (s - s_0)/\sqrt{\sigma}$,

$$\phi\left(s\right) = \tfrac{1}{2}\left[\left(\tfrac{\kappa(\alpha - s_0 - \sigma s)}{\sigma}\right)^2 - \kappa\right], \tag{2.19}$$

and

$$A(s) = \tfrac{\kappa}{\sigma}\left(\left(\alpha - s_0\right)s - \tfrac{\sigma s^2}{2}\right). \tag{2.20}$$

We will compare the performance of three estimators - first is the crude estimator based only on the first stage application of IS, second is the two-stage estimator using constant drift $\theta = T^{-1}\cdot\mathbb{E}[W_T|\overline{W}_T > a,\ \underline{W}_T < -b]$ and third is the two-stage estimator using piecewise constant drift of the form

$$\begin{aligned}\theta_t = {} & t_1^{-1}\cdot\mathbb{E}[W_{t_1}|\overline{W}_T > a,\ \underline{W}_T < -b]\cdot\mathbf{1}_{\{0<t\le t_1\}} \\ & + (T-t_1)^{-1}\cdot\mathbb{E}[W_T - W_{t_1}|\overline{W}_T > a,\ \underline{W}_T < -b]\cdot\mathbf{1}_{\{t_1<t\le T\}}\,.\end{aligned}$$

Motivated by the behaviour observed in Figure 2.1, in the case $a > b$ we select t_1 as that point where h_t attains its minimum value (in the case $a < b$) replace minimum with maximum) and determine the temporal location of this minimum numerically.

Figure 2.6 compares performance of the three estimators in the case that $h(\overline{s}, \underline{s}, s) \equiv 1$, which corresponds to estimating $\mathbb{P}(\overline{S}_T > a,\ \underline{S}_T < -b)$. As in the previous example we plot the coefficients of variation against $-\log_{10}\mathbb{P}(\overline{W}_T > F(a), \underline{W}_T < -F(b))$ for several (a, b) pairs, and each coefficient is estimated using an independent sample of size $2\cdot 10^6$.

In cases where $\mathbb{P}(\overline{W}_T > F(a), \underline{W}_T < -F(b)) < 10^{-6}$ we did not obtain a single informative scenario using the crude estimator, and in cases where $\mathbb{P}(\overline{W}_T > F(a), \underline{W}_T < -F(b)) < 10^{-8}$ we did not obtain a single informative scenario using the two-stage estimator with constant drift. By contrast the two-stage estimator with piecewise constant drift performed quite well over the entire range.

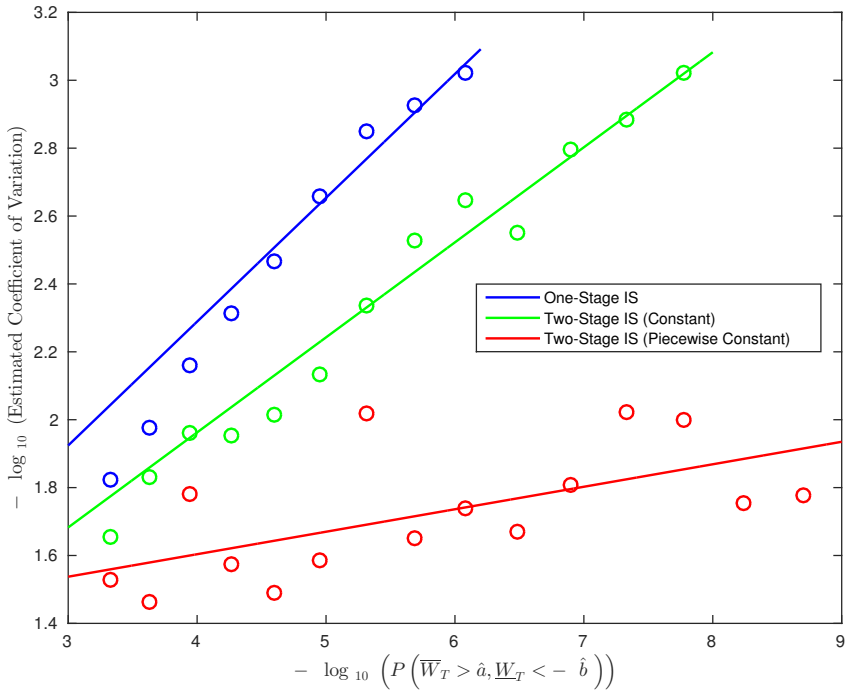

Figure 2.6: *Estimated coefficient of variation versus* $\mathbb{P}(\overline{W}_T > \hat{a},\ \underline{W}_T < -\hat{b})$ *on the* log *scale for the First-Stage estimator, Two-Stage estimator with constant drift and the Two-Stage estimator with piecewise constant drift. We fit an exponential curve for the First-Stage estimator and Two-Stage estimator with constant drift, and we fit a linear curve for the Two-Stage estimator with piecewise constant drift. The parameters used in this Figure are* $\kappa = 1,\ \alpha = 0,\ \sigma = \sqrt{2},\ s_0 = 0.06$ *and* $T = 1$.

The evidence here is consistent with that obtained in the previous example, namely that the two-stage estimator can perform well even when the crude estimator fails completely. However we also see that this performance boost is by no means guaranteed, and we see that one must select an importance drift that is sufficiently reflective of h_t.

One can potentially improve the performance of the Piecewise Constant Importance Sampling estimator by adding more subintervals. However, doing so would increase

computation time substantially since it increases the number of minima that have to be generated. If we have M subintervals and n trajectories, the Piecewise Constant Importance Sampling estimator will have to generate an extra $2n(M-1)$ minima on average. Therefore, it might be better to increase the number of simulations rather than adding subintervals. This tradeoff is problem-specific and depends highly on the shape of the minimum entropy drift. Consequently, one should analyze its behaviour before implementing the piecewise linear estimator.

2.5 Concluding Remarks

This chapter proposes an importance sampling procedure that can be used to estimate expectations of path functionals of diffusion processes that is free of discretization error and performs well when crude estimators fail completely. We pay particular attention to the so-called measure selection problem for standard Brownian motion and propose a simple criteria for selecting effective importance drifts for standard Brownian motion. The criteria is based on entropy minimization (between the importance measure and conditional law of the Brownian motion, given that its trajectory lies in some region of interest) and leads to an intuitive choice that is easy to implement. Moreover in the example considered in this paper the performance of importance sampling estimators based on this criteria is virtually indistinguishable from that of an alternative criteria, based on large deviations, that is known to be asymptotically optimal in a certain sense, but difficult to implement in practice.

In the case where the functional of interest depends only on the extreme and terminal values of the underlying diffusion we demonstrate that our proposed algorithm offers substantial variance reduction, and performs admirably in cases where crude estimators fail completely. Obvious extensions that would be of interest include the

development of algorithms that can be used for more general path functionals, a deeper understanding of the connection between our entropy criteria and the large deviations criteria proposed by Guasoni and Robertson (2008), as well as extending our approach to multivariate diffusions (where the Radon-Nykodym derivatives can be much more complicated to work with) and/or jump diffusions.

Bibliography

Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 1st edition, 2007.

Søren Asmussen, Dirk P. Kroese, and Reuven Y. Rubinstein. Heavy Tails, Importance Sampling And Cross–Entropy. *Stoch. Models*, 21:57–76, 2005.

Alexandros Beskos and Gareth O. Roberts. Exact Simulation of Diffusions. *Ann. Appl. Probab.*, 15(4):2422–2444, 2005.

Nan Chen and Zhengyu Huang. Brownian Meanders, Importance Sampling and Unbiased Simulation of Diffusion Extremes. *Oper. Res. Lett.*, 40(6):554–563, 2012.

Luc Devroye. *On Exact Simulation Algorithms for Some Distributions Related to Brownian Motion and Brownian Meanders*, 2009.

Joe DiCesare. *Imputation, Estimation and Missing Data in Finance*. PhD thesis, University of Waterloo, 2006.

Joe DiCesare and Don McLeish. Simulation of Jump Diffusions and the Pricing of Options. *Insurance: Mathematics and Economics*, 43(3):316–326, 2008.

K. Giesecke and D. Smelov. Exact Sampling of Jump-Diffusions. *Operations Research*, 61(4):894–907, August 2013.

Paolo Guasoni and Scott Robertson. Optimal Importance Sampling With Explicit Formulas in Continuous Time. *Finance Stoch.*, 12:1–19, 2008.

Adam Metzler and Alexandre Scott. Rare Event Simulation for Diffusion Processes via Two-Stage Importance Sampling. *Monte Carlo Methods and Appl.*, 20(2): 77–100, March 2014. doi: 10.1515/mcma-2013-0019.

Chapter 3

Rare Event Simulation for Portfolio Losses

3.1 Introduction

This chapter addresses the problem of estimating[1]

$$\mathbb{P}(L_N \geq x), \tag{3.1}$$

where L_N is the loss on a portfolio of N obligors and x is a large user-defined threshold. The two primary means of estimating (3.1) in the conditional independence framework are convolution (see Merino and Nyfeler (2002)) and Monte Carlo. If N is large, as is often the case in practice, the former approach becomes unwieldy and the latter approach is the only feasible method. If in addition x is large, as the case when computing risk measures such as VaR or CVaR, the variability of the crude Monte Carlo estimator tends to be unacceptably large. If, as is so often the case in practice, the computational budget is limited then there is clear value in implementing variance reduction schemes.

Several authors have developed effective importance sampling (IS) algorithms for esti-

[1] A version of this chapter has been submitted for publication (see Scott and Metzler (2015)).

mating (3.1) in particular models. At the present time, the most effective algorithms in the Gaussian and t models (assuming non-random loss given default, at least) are those developed by Glasserman and Li (2005) and Chan and Kroese (2010), respectively. This chapter develops a more general procedure that can be applied in a wide variety of models, including but not limited to the Gaussian and t. The proposed algorithm is more general, offers comparable performance when benchmarked against these "gold standards" (both of which are known to have good asymptotic properties), requires approximately the same computation time as Glasserman and Li (2005) and requires substantially less computational time than the Chan and Kroese (2010) algorithm.

3.1.1 Summary of Proposed Algorithm

There are two components to any effective IS algorithm. The first is the identification of a parametric family of candidate measures from which to simulate. This family must be tractable in the sense that all members are straightforward to simulate from. The second is the criteria used to select an effective (in terms of variance reduction) member of the candidate family. In general this component involves both theoretical characterization of the optimal parameter value(s) and computational methods for approximating them.

Throughout this chapter, we employ a generalization of the well-known exponential tilt for the first component of our algorithm. In particular we apply an exponential tilt to the distribution of the natural sufficient statistics of the systematic risk factors and then, conditional on the simulated values of the systematic factors, apply an exponential tilt to the conditional distribution of portfolio loss. Being unable to locate a precise definition or treatment of such a change of measure (which is in general infinite dimensional since the tilt parameter applied in the second stage is a

function of the systematic factors) in the literature we term it a sequential exponential tilt in what follows. Exponential tilts are well known to provide very effective variance reduction in a wide variety of contexts (see Asmussen and Glynn (2007) and references therein) and the idea of using (what we call) sequential tilts in the portfolio problem has appeared elsewhere in the literature, most notably in Glasserman and Li (2005), albeit in less generality.

The second component of our algorithm consists of choosing the tilt parameters so as to ensure that the resulting IS measure has minimal Kullback-Liebler divergence (KLD) from the ideal (but impractical) zero-variance measure. In this sense our algorithm can be seen as a variant of the well-known cross-entropy (CE) method for estimating rare event probabilities. When combined with the parametric family described in the previous section this criteria allows one to characterize optimal parameter values via intuitive moment-matching equations. Elementary properties of the parametric family ensure that these equations are well-behaved, and we exploit asymptotic features of the conditional independence structure to develop efficient approximations to their solutions. In the homogeneous case the approximations are available via simple quadrature, in more general cases efficient iterative algorithms (such as the adaptive CE algorithm described in Asmussen and Glynn (2007)) are available.

3.1.2 Brief Literature Review

Glasserman and Li (2005) develop an extremely effective IS algorithm in the Gaussian case. They apply a sequential tilt of the form described in the previous section and appeal to large deviations theory in order to select optimal parameters. They are able to prove that their algorithm is asymptotically optimal in the homogeneous case, and it is well known to perform extremely well under a moderate degree of heterogeneity.

For certain correlation structures the approach fails to be asymptotically optimal (but still seems to perform well in most cases of practical interest) and Glasserman et al. (2008) prove that asymptotic optimality can be achieved in these instances by using mixtures of sequential tilts. The only real drawback to the Glasserman and Li (2005) algorithm is that it is not clear how it could be extended to incorporate important phenomena such as heavy-tailed risk factors or correlation between default rates and loss given default.

At the present time the most effective available algorithm in the Student t copula case appears to be that of Chan and Kroese (2010), which combines conditional Monte Carlo with a clever observation regarding order statistics. The Chan and Kroese (2010) algorithms requires substantially less computational time than the algorithm developed by Bassamboo et al. (2006). The former algorithm is several times faster than the latter, and the algorithm proposed in this paper is several times faster than the former. The reason for this efficiency gain appears to be that Chan and Kroese (2010) requires one to sort high-dimensional vectors, which can be quite cumbersome when the number of obligors is large. In particular the complexity of the Chan and Kroese (2010) algorithm is of order $N \log(N)$ whereas the algorithm proposed in this paper is of order N.

3.1.3 Outline of the Chapter

Sections 3.2 and 3.3 describe the mathematical structures of the financial and simulation problems, respectively. Section 3.4 applies the general simulation methodology outlined in Section 3.3 to the financial problem outlined in Section 3.2. Sections 3.5 and 3.6 apply the methodology in the Gaussian and t cases, respectively, in each case benchmarking the performance of the proposed algorithm to the relevant gold standard.

3.2 Financial Setting

Consider a portfolio of obligors labelled $i = 1, 2, \ldots, N$, and an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let Y_i be the default indicator for obligor i, taking on the value one if obligor i defaults over a given time horizon and zero otherwise. In the conditional independence framework (where the default indicators are conditioned on $\mathbf{Z}$) default indicators take the form $Y_i = f_i(\mathbf{Z}, \varepsilon_i)$, where $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$ is a d-dimensional vector of systematic risk factors and $\varepsilon_1, \varepsilon_2, \ldots$ are idiosyncratic risk factors. All risk factors are uncorrelated and we may, without loss of generality, assume that they have zero mean and unit variance. That is, each member of the sequence $Z_1, \ldots, Z_d, \varepsilon_1, \ldots$ has zero mean and unit variance, and distinct members are uncorrelated. Finally, let $PD_i = \mathbb{P}(Y_i = 1)$ be the default probability for obligor i and let $p_i(\mathbf{z}) = \mathbb{P}(Y_i = 1 | \mathbf{Z} = \mathbf{z})$ denote the conditional default probability of obligor i, given $\mathbf{Z} = \mathbf{z}$.

Assume that loss given default (LGD_i), exposure at default (E_i) and default probability (PD_i) are known constants for each obligor. For $N \geq 1$ and $1 \leq i \leq N$ let $c_i^{(N)} = LGD_i E_i / \sum_{j=1}^{N} E_j$ be the relative effective exposure to obligor i, among the first N obligors. Then the percentage loss on a portfolio consisting of the first N obligors takes the form

$$L_N = \sum_{i=1}^{N} c_i^{(N)} \cdot Y_i \, .$$

It is well known that

$$\lim_{N \to \infty} (L_N - \mathbb{E}[L_N | \mathbf{Z}]) = 0 \ \text{ a.s.} \tag{3.2}$$

under mild conditions on the relative exposures. For instance a sufficient condition is that $c_N^{(N)} = O(N^{-1/2+\xi})$ for some $\xi > 0$ (see Gordy (2003)), which effectively ensures that no single exposure dominates the portfolio. Throughout the paper we

will assume that the convergence in (3.2) holds, and will refer to $\mathbb{E}[L_N|\mathbf{Z}]$ as the asymptotic approximation to L_N.

In a linear factor model, the object of this paper, one associates a credit quality variable X_i with each obligor. This variable is characterized by a d-dimensional vector a_i, assumed to satisfy $|a_i| < 1$ where $|\cdot|$ denotes vector norm, and is defined via $X_i = a_i^T\mathbf{Z} + \sqrt{1-|a_i|^2}\varepsilon_i$. Default indicators are then constructed via $Y_i = \mathbf{1}_{\{X_i \le x_{PD_i}\}}$, where x_{PD_i} is the $100PD_i\%$ quantile of X_i. We assume that all factor loadings are positive to ensure positive correlation between credit qualities.

For fixed values of x and N define the probability measure $\widetilde{\mathbb{P}}$ by $\widetilde{\mathbb{P}}(A) := \mathbb{P}(A|L_N \ge x)$. Then $\widetilde{\mathbb{P}}$ is absolutely continuous with respect to $\mathbb{P}$ (but not vice versa), with Radon-Nikodym derivative (RND)

$$\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}} = \frac{\mathbf{1}_{\{L_N \ge x\}}}{\mathbb{P}(L_N \ge x)} ,$$

It is well known that $\widetilde{\mathbb{P}}$ provides an ideal but impractical importance sampling (IS) measure in the sense that it (i) results in a zero-variance estimator but (ii) is not practical to simulate from (since it requires knowledge of the very quantity being estimated). We will henceforth refer to $\widetilde{\mathbb{P}}$ as the ideal IS measure, turning now to the problem of identifying that member of a tractable parametric family that is as close as possible to the ideal measure, in the sense of Kullback-Liebler divergence (KLD).

3.3 Statistical Setting

The portfolio problem is an instance of the following, more general, rare event problem. Suppose that one is interested in estimating $\mathbb{P}(h(\mathbf{V}) \ge x)$, where $\mathbf{V} : \Omega \mapsto \mathbb{R}^m$ is a random vector on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $h : \mathbb{R}^m \mapsto \mathbb{R}$ is a deterministic

function and x is large in the sense that the probability of interest is small. If the moment generating function (mgf) of $\mathbf{V}$ is available in closed form then the well-known cross-entropy (CE) method can provide very effective variance reduction; see Asmussen and Glynn (2007) and references therein for an excellent and comprehensive treatment of the method.

The basic idea behind the CE method is as follows. If the mgf of $\mathbf{V}$ is available in closed form then the parametric family of IS candidates defined by RNDs of the form

$$\frac{d\mathbb{P}^{\alpha}}{d\mathbb{P}} = \frac{\exp(\alpha^T \mathbf{V})}{\mathbb{E}[\exp(\alpha^T \mathbf{V})]} \tag{3.3}$$

constitutes an exponential family (see Bickel and Doksum (2001) for basic definitions and important properties) and therefore has many attractive properties. In particular $\mathbb{E}^{\alpha}[\mathbf{V}] = \nabla\Psi(\alpha)$ and $\mathbb{V}^{\alpha}[\mathbf{V}] = \mathrm{H}\Psi(\alpha)$, where (i) $\mathbb{E}^{\alpha}$ and $\mathbb{V}^{\alpha}$ denote mean vector and covariance matrix under $\mathbb{P}^{\alpha}$, (ii) ∇ and H denote gradient and Hessian and (iii) $\Psi(\alpha) = \log(\mathbb{E}[\exp(\alpha\mathbf{V})])$ is the cumulant generating function (cgf) of $\mathbf{V}$ under $\mathbb{P}$. The $\mathbb{P}^{\alpha}$-distribution of $\mathbf{V}$ is often called an exponential tilt of the $\mathbb{P}$-distribution of $\mathbf{V}$, with tilt parameter α, or simply the α-tilted distribution of $\mathbf{V}$.

Now let α^* denote that value of α that minimizes the Kullback-Liebler divergence (KLD) of the candidate family from the ideal measure $\widetilde{\mathbb{P}}$, where $\widetilde{\mathbb{P}}(A) := \mathbb{P}(A|h(\mathbf{V}) \geq x)$. Then $\mathbb{P}^{\alpha^*}$ can be considered the optimal member of the candidate family in the sense that it is as "close" as possible to the ideal measure. It is reasonably well known (see Lemma 2 in Hu et al. (2007), for instance) that α^* can be characterized by the fact that $\mathbb{E}^{\alpha^*}[\mathbf{V}] = \widetilde{\mathbb{E}}[\mathbf{V}]$ where $\mathbb{E}^{\alpha^*}$ is taken with respect to $\mathbb{P}^{\alpha^*}$ and $\widetilde{\mathbb{E}}$ with respect to $\widetilde{\mathbb{P}}$. Thus if one has a good estimate of $\widetilde{\mathbb{E}}[\mathbf{V}]$ then, since $\mathbb{E}^{\alpha}[\mathbf{V}]$ is in general a very well-behaved function of α, efficient numerical approximation of α^* is straightforward. Even if one is not able to develop an analytic approximation to $\widetilde{\mathbb{E}}[\mathbf{V}]$, the so-called adaptive CE algorithm (see Chapter VI.8 of Asmussen and Glynn (2007)) can be used

to construct a sequence α_n converging to α^* under mild conditions (in particular the initial guess must be carefully chosen).

The power of the CE method lies in its amenability, when combined with exponential candidate families, to the efficient approximation of optimal parameters. It is a shame, then, that it is not directly applicable in the portfolio context for the simple reason that the mgf of L_N is in general not available in closed form. In the remainder of this section we explain how the CE method can be generalized, and its computational efficiency can be retained, in cases where the mgf of $\mathbf{V}$ is not tractable but there exists another random vector $\mathbf{R}$ such that both the mgf of $\mathbf{R}$ and the conditional mgf of $\mathbf{V}$ given $\mathbf{R}$ are tractable.

3.3.1 Sequential Exponential Tilts

Suppose that $\mathbf{R} : \Omega \mapsto \mathbb{R}^n$ is a random vector on the same probability space such that (i) the mgf of $\mathbf{R}$ and (ii) the conditional mgf of $\mathbf{V}$, given $\mathbf{R}$, are available in closed form. To fix notation let $\mathcal{S}_{\mathbf{R}}$ denote the support of $\mathbf{R}$. For $\lambda \in \mathbb{R}^n$ let $\Psi(\lambda) = \log(\mathbb{E}[\exp(\lambda^T \mathbf{R})])$ and for $\mathbf{r} \in \mathcal{S}_{\mathbf{R}}$ and $\theta \in \mathbb{R}^m$ let $\psi_{\mathbf{r}}(\theta) = \log(\mathbb{E}[\exp(\theta^T \mathbf{V})|\mathbf{R} = \mathbf{r}])$. Then $\Psi(\cdot)$ is the cgf of $\mathbf{R}$ and $\psi_{\mathbf{r}}(\cdot)$ is the conditional cgf of $\mathbf{V}$, given $\mathbf{R} = \mathbf{r}$. Finally, let $\Lambda = \{\lambda \in \mathbb{R}^n : \Psi(\lambda) < \infty\}$ and for each $\mathbf{r} \in \mathcal{S}_{\mathbf{R}}$ let $\Theta_{\mathbf{r}} = \{\theta \in \mathbb{R}^m : \psi_{\mathbf{r}}(\theta) < \infty\}$. Note that Ψ and $\psi_{\mathbf{r}}$ are convex.

Now let $\lambda \in \Lambda$ and suppose that $\theta_\bullet : \mathbb{R}^n \mapsto \mathbb{R}^m$ is such that $\theta_{\mathbf{r}} \in \Theta_{\mathbf{r}}$ for all $\mathbf{r} \in \mathcal{S}_{\mathbf{R}}$. Define the probability measure $\mathbb{P}^{\lambda,\theta_\bullet}$ via the RND

$$\frac{d\mathbb{P}^{\lambda,\theta_\bullet}}{d\mathbb{P}} = \exp\left(\lambda^T \mathbf{R} - \Psi(\lambda)\right) \cdot \exp(\theta_{\mathbf{R}}^T \mathbf{V} - \psi_{\mathbf{R}}(\theta_{\mathbf{R}})) \,. \tag{3.4}$$

Then (i) the $\mathbb{P}^{\lambda,\theta_\bullet}$-distribution of $\mathbf{R}$ is the λ-tilt of the $\mathbb{P}$-distribution of $\mathbf{R}$ and (ii) the $\mathbb{P}^{\lambda,\theta_\bullet}$-conditional distribution of $\mathbf{V}$, given $\mathbf{R} = \mathbf{r}$, is the $\theta_{\mathbf{r}}$-tilt of the $\mathbb{P}$-conditional

distribution of $\mathbf{V}$, given $\mathbf{R} = \mathbf{r}$. Elementary results for exponential families (see Section 1.6 in Bickel and Doksum (2001), for example) ensure that (i) $\nabla\Psi(\lambda) = \mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{R}]$ and $\nabla\psi_{\mathbf{r}}(\theta_{\mathbf{r}}) = \mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{V}|\mathbf{R} = \mathbf{r}]$, where $\mathbb{E}^{\lambda,\theta_\bullet}$ denotes mean vector under $\mathbb{P}^{\lambda,\theta_\bullet}$, and (ii) $\mathrm{H}\Psi(\lambda) = \mathbb{V}^{\lambda,\theta_\bullet}(\mathbf{R})$ and $\mathrm{H}\psi_{\mathbf{r}}(\theta_{\mathbf{r}}) = \mathbb{V}^{\lambda,\theta_\bullet}(\mathbf{V}|\mathbf{R} = \mathbf{r})$, where $\mathbb{V}^{\lambda,\theta_\bullet}$ denotes covariance matrix under $\mathbb{P}^{\lambda,\theta_\bullet}$.

We call $\mathbb{P}^{\lambda,\theta_\bullet}$ a sequential exponential tilt of $\mathbb{P}$. Though particular instances of sequential tilts have appeared in the literature, see Glasserman and Li (2005) for an important example, to the best of our knowledge they have not been treated in any generality. In the next section we show that if one minimizes the KLD of a family of this form from the ideal IS measure $\widetilde{\mathbb{P}}$ then the resulting optimal parameters continue to solve intuitive moment-matching problems, and continue to be amenable to efficient numerical approximation.

3.3.2 Minimum Divergence with Sequential Exponential Tilts

The KLD of $\mathbb{P}^{\lambda,\theta_\bullet}$ from $\widetilde{\mathbb{P}}$ is

$$\widetilde{\mathbb{E}}\left[\log\left(\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^{\lambda,\theta_\bullet}}\right)\right] = \widetilde{\mathbb{E}}\left[\log\left(\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}}\right)\right] + \widetilde{\mathbb{E}}[-\lambda^T\mathbf{R} + \Psi(\lambda)] + \widetilde{\mathbb{E}}[-\theta_{\mathbf{R}}^T\mathbf{V} + \psi_{\mathbf{R}}(\theta_{\mathbf{R}})]\ , \quad (3.5)$$

where we have used the fact that $\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}^{\lambda,\theta_\bullet}} = \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}}\frac{d\mathbb{P}}{d\mathbb{P}^{\lambda,\theta_\bullet}}$. The following result shows that, so long as boundary solutions are ruled out, the optimal choices of λ and $\theta_\bullet$ will satisfy an intuitive moment-matching condition. This result can be seen as a generalization of Lemma 2 in Hu et al. (2007), which states the analogous result in the non-sequential case.

Theorem 3.3.1. *Suppose that Λ is open and that there exists a solution (in λ) to $\nabla\Psi(\lambda) = \widetilde{\mathbb{E}}[\mathbf{R}]$. Further suppose that for each $\mathbf{r} \in \mathcal{S}_{\mathbf{r}}$, $\Theta_{\mathbf{r}}$ is open and there exists*

a solution (in θ) to $\nabla\psi_{\mathbf{r}}(\theta) = \widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$. Then (3.5) *admits a global minimzer at* $(\lambda^*, \theta_\bullet^*)$*, where*

$$\begin{aligned} \mathbb{E}^{\lambda^*,\theta_\bullet^*}[\mathbf{R}] &= \widetilde{\mathbb{E}}[\mathbf{R}] \\ &\text{and} \\ \mathbb{E}^{\lambda^*,\theta_\bullet^*}[\mathbf{V}|\,\mathbf{R}=\mathbf{r}] &= \widetilde{\mathbb{E}}[\mathbf{V}|\,\mathbf{R}=\mathbf{r}] \end{aligned}$$

for every $\mathbf{r} \in \mathcal{S}_{\mathbf{R}}$.

Proof. In order to minimize (3.5) with respect to λ and $\theta_\bullet^*$, it clearly suffices to minimize the second and third terms individually. The map $\lambda \mapsto -\lambda^T \widetilde{\mathbb{E}}[\mathbf{R}] + \Psi(\lambda)$ is convex on Λ, therefore any solution to $\nabla\Psi(\lambda) = \widetilde{\mathbb{E}}[\mathbf{R}]$ is necessarily a global minimizer since Λ is assumed open. When it exists, let λ^* denote such a solution. Since $\nabla\Psi(\lambda) = \mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{R}]$ it follows that $\mathbb{E}^{\lambda^*,\theta_\bullet}[\mathbf{R}] = \widetilde{\mathbb{E}}[\mathbf{R}]$ for any $\theta_\bullet$.

For each fixed $\mathbf{r} \in \mathcal{S}_{\mathbf{R}}$ the map $\theta \mapsto -\theta^T\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}] + \psi_{\mathbf{r}}(\theta)$ is convex on $\Theta_{\mathbf{r}}$, therefore any solution to $\nabla\psi_{\mathbf{r}}(\theta) = \widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$ is necessarily a global minimizer since $\Theta_{\mathbf{r}}$ is assumed open. When it exists, let $\theta_{\mathbf{r}}^*$ denote such a solution. If $\theta_\bullet : \mathbb{R}^m \mapsto \mathbb{R}^n$ is any function then $-\theta_{\mathbf{R}}^*\mathbf{V} + \psi_{\mathbf{R}}(\theta_{\mathbf{R}}^*) \leq -\theta_{\mathbf{R}}\mathbf{V} + \psi_{\mathbf{R}}(\theta_{\mathbf{R}})$, whence $-(\theta_{\mathbf{R}}^*)^T\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}] + \psi_{\mathbf{R}}(\theta_{\mathbf{R}}^*) \leq -\theta_{\mathbf{R}}^T\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}] + \psi_{\mathbf{R}}(\theta_{\mathbf{R}})$ almost surely. Applying the tower property to the third term in (3.5) we conclude that it is minimized by setting $\theta_{\mathbf{r}} = \theta_{\mathbf{r}}^*$. Since $\nabla\psi_{\mathbf{r}}(\theta_{\mathbf{r}}) = \mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$ we have $\mathbb{E}^{\lambda,\theta_\bullet^*}[\mathbf{R}] = \widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$ for any λ. The result follows. □

3.3.3 General Algorithm

In this section we summarize the CE method using sequential tilts. Since $\mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{R}]$ does not depend on $\theta_\bullet$ we may, without loss of generality, simply write $\mathbb{E}^{\lambda}[\mathbf{R}]$ instead of $\mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{R}]$. Similarly we may write $\mathbb{E}^{\theta_\bullet}[\mathbf{V}|\mathbf{R}]$ instead of $\mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{V}|\mathbf{R}]$. Finally, with an

admitted abuse of notation we write $\mathbb{E}^{\theta_{\mathbf{r}}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$ instead of $\mathbb{E}^{\lambda,\theta_{\bullet}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$. The algorithm is as follows.

1. Estimate λ^* with, say, $\hat{\lambda}$. If $\widetilde{\mathbb{E}}[\mathbf{R}]$ is available in closed form then one can typically compute λ^* exactly (or to any desired degree of accuracy) using the characterization in Theorem 3.3.1. Otherwise this step would involve (i) developing an approximation to $\widetilde{\mathbb{E}}[\mathbf{R}]$, say $\widehat{\widetilde{\mathbb{E}}[\mathbf{R}]}$, and then (ii) solving $\mathbb{E}^{\hat{\lambda}}[\mathbf{R}] = \widehat{\widetilde{\mathbb{E}}[\mathbf{R}]}$ for $\hat{\lambda}$. This calculation only needs to be carried out once, at the outset of the algorithm.

2. Simulate M independent copies of $\mathbf{R}$ from its $\hat{\lambda}$-tilted distribution. Denote the simulated values $\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_M$.

3. For each $i = 1, 2, \ldots, M$ approximate $\theta^*_{\mathbf{r}_i}$. If $\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$ is available in closed form then one can typically compute $\theta^*_{\mathbf{r}_i}$ exactly (or to any desired degree of accuracy) using the characterization in Theorem 3.3.1. Otherwise one needs to develop an approximation to $\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$, say $\widehat{\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]}$, and then solve $\mathbb{E}^{\hat{\theta}_{\mathbf{r}_i}}[\mathbf{V}|\mathbf{R}=\mathbf{r}_i] = \widehat{\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}_i]}$ for $\hat{\theta}_{\mathbf{r}_i}$. This procedure needs to be carried out M times, and therefore rapid approximation of $\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}]$ and rapid solution of $\mathbb{E}^{\hat{\theta}_{\mathbf{r}_i}}[\mathbf{V}|\mathbf{R}=\mathbf{r}_i] = \widehat{\widetilde{\mathbb{E}}[\mathbf{V}|\mathbf{R}=\mathbf{r}_i]}$ are of paramount importance. In our experience it is advantageous to avoid the adaptive CE method here, if possible - in the portfolio problem its use at this stage of the algorithm increases total computational time by several hundred times.

4. For each $i = 1, 2, \ldots, M$ simulate $\mathbf{V}_i$ from its $\hat{\theta}_{\mathbf{r}_i}$-tilted conditional (on $\mathbf{R}=\mathbf{r}_i$) distribution. Denote the simulated values $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M$, and note that each value is drawn from a potentially different distribution.

5. Return the estimate $\frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{\{h(\mathbf{v}_i)\geq x\}} \exp(-\hat{\lambda}^T\mathbf{r}_i + \Psi(\hat{\lambda})) \exp(-\hat{\theta}^T_{\mathbf{r}_i}\mathbf{v}_i + \psi_{\mathbf{r}_i}(\hat{\theta}_{\mathbf{r}_i}))$.

3.3.4 Important Special Case

Suppose that conditional on $\mathbf{R} = \mathbf{r}$, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^T$ is a vector of N independent Bernoullis random variables with success probabilities $p_i(\mathbf{r}) = \mathbb{P}(Y_i = 1|\mathbf{R} = \mathbf{r})$. If $\mathbf{V} = \mathbf{c}^T\mathbf{Y}$ for some fixed $\mathbf{c} \in \mathbb{R}^N$, then under $\mathbb{P}^{\lambda,\theta_\bullet}$ and conditional on $\mathbf{R} = \mathbf{r}$, $\mathbf{Y}$ is also a vector of independent Bernoullis random variables with success probabilities

$$\mathbb{P}^{\lambda,\theta_\bullet}(Y_i = 1|\mathbf{R} = \mathbf{r}) = \frac{p_i(\mathbf{r})e^{\theta_\mathbf{r}\cdot c_i}}{1 + (e^{\theta_\mathbf{r} c_i} - 1)p_i(\mathbf{r})}, \quad i = 1, \ldots, N\ . \tag{3.6}$$

Thus

$$\mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{V}|\mathbf{R} = \mathbf{r}] = \sum_{i=1}^{N} \frac{c_i p_i(\mathbf{r})e^{\theta_\mathbf{r}\cdot c_i}}{1 + (e^{\theta_\mathbf{r} c_i} - 1)p_i(\mathbf{r})}\ . \tag{3.7}$$

Observe that (3.7) is smooth and monotone in $\theta_\mathbf{r} \in \mathbb{R}$, ranging from 0 as $\theta_\mathbf{r} \to -\infty$ to $\sum_{i=1}^N c_i$ as $\theta_\mathbf{r} \to \infty$. The implication is that solving $\mathbb{E}^{\lambda,\theta_\bullet}[\mathbf{V}|\mathbf{R} = \mathbf{r}] = K$ for $\theta_\mathbf{r}$ is a simple root-finding exercise for any constant $K \in [0, \sum_{i=1}^N c_i]$.

3.4 Application to the Portfolio Problem

In this section we apply the method of Section 3.3 to the problem of Section 3.2. For notational simplicty we henceforth suppress dependence of relative exposures on the size of the portfolio, writing c_i for $c_i^{(N)}$ and $\mathbf{c}$ for $\mathbf{c}_N$. In addition we write $\mathbf{Y}$ for $\mathbf{Y}_N$ when there is no danger of confusion.

We assume that the systematic risk factors come from an exponential family with natural sufficient statistic $\mathbf{R} = f(\mathbf{Z})$ and natural parameter λ, for some one-to-one function f. Setting $\mathbf{V} = L_N = \mathbf{c}^T\mathbf{Y}$ we are in the setting of Section 3.3 and the discussion in Section 3.3.4 applies. In what follows we condition on $\mathbf{Z}$ instead of $\mathbf{R}$,

since $\mathbf{Z}$ is the more natural object in the financial context (since f is one-to-one there is no difference between conditioning on $\mathbf{Z}$ or $\mathbf{R}$).

The ideal IS measure here is $\widetilde{\mathbb{P}}(A) := \mathbb{P}(A|L_N \geq x)$, hence the moments that need to be matched are $\widetilde{\mathbb{E}}[f(\mathbf{Z})] = \mathbb{E}[f(\mathbf{Z})|L_N \geq x]$ and $\widetilde{\mathbb{E}}[L_N|\mathbf{Z} = \mathbf{z}] = \frac{\mathbb{E}[L_N \mathbf{1}_{\{L_N \geq x\}}|\mathbf{Z}=\mathbf{z}]}{\mathbb{P}(L_N \geq x|\mathbf{Z}=\mathbf{z})}$. In general neither of these is available in closed and approximations must be developed, problems we address in Sections 3.4.1 and 3.4.2, respectively. Before proceeding we define

$$\ell_N(\mathbf{z}) := \mathbb{E}[L_N|\mathbf{Z} = \mathbf{z}] = \sum_{i=1}^{N} c_i p_i(\mathbf{z}) \ ,$$

so that $\ell_N(\mathbf{Z})$ is the asymptotic approximation to L_N.

3.4.1 Estimating $\widetilde{\mathbb{E}}[f(\mathbf{Z})]$ and Computing $\hat{\lambda}$

Recall that converges L_N almost surely to $\ell_N(\mathbf{Z})$ as $N \to \infty$. The good agreement between these quantities in the Gaussian and t models is illustrated in Figures 3.1 and 3.5, respectively. Thus a tractable approximation to $\widetilde{\mathbb{E}}[f(\mathbf{Z})]$ can be had by replacing L_N with its asymptotic approximation $\ell_N(\mathbf{Z})$, leading to the estimate

$$\widetilde{\mathbb{E}}[f(\mathbf{Z})] = \mathbb{E}[f(\mathbf{Z})|L_N \geq x] \approx \mathbb{E}[f(\mathbf{Z})|\ell(\mathbf{Z}) \geq x] \ .$$

Thus $\hat{\lambda}$ satisfies

$$\mathbb{E}^{\hat{\lambda}}[f(\mathbf{Z})] = \mathbb{E}[f(\mathbf{Z})|\ell(\mathbf{Z}) \geq x] \ . \tag{3.8}$$

For homogeneous portfolios the right hand side of (3.8) can often be put in semi-analytic form, in which case numerical determination of $\hat{\lambda}$ is straightforward and efficient. In other cases the adaptive CE method described in Chapter VI.8 of As-

mussen and Glynn (2007) provides a very efficient means of estimating $\hat{\lambda}$. Adapted to the present setting one initializes that algorithm by choosing the value of the user-specified constant $\eta \in (0,1)$ and setting $\lambda_0 = \mathbf{0}_d$ where $\mathbf{0}_d$ denotes the d-dimensional vector whose entries are 0. The algorithm proceeds iteratively as follows. For $n \geq 0$ one simulates $K << M$ independent copies $\mathbf{Z}_1^n, \ldots, \mathbf{Z}_K^n$ from $\mathbb{P}^{\lambda_n}$, where $\frac{d\mathbb{P}^{\lambda_n}}{d\mathbb{P}} = \exp(\lambda_n^T \mathbf{R}^n - \Psi(\lambda_n))$, and sets x_n to be the minimum of (i) the $100(1-\eta)\%$ quantile of the simulated values $\ell_N(\mathbf{Z}_1^n), \ldots, \ell_N(\mathbf{Z}_K^n)$ and (ii) x. One then computes λ_{n+1} by solving

$$\mathbb{E}^{\lambda_{n+1}}[\mathbf{R}] = \frac{\sum_{k=1}^K \exp(-\lambda_n^T \mathbf{R}_k^n + \Psi(\lambda_n)) \cdot \mathbf{R}_k^n \cdot \mathbf{1}_{\{\ell_N(\mathbf{Z}_k^n) \geq x_n\}}}{\sum_{k=1}^K \exp(-\lambda_n^T \mathbf{R}_k^n + \Psi(\lambda_n)) \cdot \mathbf{1}_{\{\ell_N(\mathbf{Z}_k^n \geq x_n\}}} \ . \tag{3.9}$$

If $x_n = x$ the algorithm is terminated, no further iterations are carried out and λ_{n+1} is returned as an approximation to $\hat{\lambda}$. In our experience we have found that this algorithm typically terminates within a very small number of iterations and produces a very accurate estimate of $\hat{\lambda}$.

3.4.2 Estimating $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}]$

In order to develop a tractable approximation to $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}]$ we begin with the observation (to be formally justified shortly) that

$$\widetilde{\mathbb{E}}[L_N|\mathbf{Z}] \geq \max(\mathbb{E}[L_N|\mathbf{Z}], x) \ , \tag{3.10}$$

with equality obtaining as $x \to 0$ or $x \to \sum_{i=1}^N c_i$. In many cases the agreement between the two sides of (3.10) is quite good for intermediate values of x as well, particularly if N is large (see Figure 3.2 for an illustrative example in the context of the Gaussian copula model). A tractable approximation to $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}]$ can therefore be had by replacing $\mathbb{E}[L_N|\mathbf{Z}]$ with its asymptotic approximation $\ell_N(\mathbf{Z})$ in the right-hand

side of (3.10), leading to the estimate

$$\widetilde{\mathbb{E}}[L_N|\mathbf{Z}] \approx \max(\ell_N(\mathbf{Z}), x) \ .$$

Thus $\hat{\theta}_{\mathbf{z}}$ satisfies

$$\mathbb{E}^{\hat{\theta}_{\mathbf{z}}}[L_N|\mathbf{Z} = \mathbf{z}] = \max(\ell_N(\mathbf{z}), x) \ , \tag{3.11}$$

and we recall that

$$\mathbb{E}^{\hat{\theta}_{\mathbf{z}}}[L_N|\mathbf{Z} = \mathbf{z}] = \sum_{i=1}^{N} \frac{c_i p_i(\mathbf{z}) e^{\hat{\theta}_{\mathbf{z}} \cdot c_i}}{1 + (e^{\hat{\theta}_{\mathbf{z}} c_i} - 1) p_i(\mathbf{z})} \ .$$

If $\hat{\theta}_{\mathbf{z}} = 0$ then the left-hand side of (3.11) reduces to $\ell_N(\mathbf{z})$, whence $\hat{\theta}_{\mathbf{z}} = 0$ for all $\mathbf{z}$ such that $\ell_N(\mathbf{z}) \geq x$. Otherwise, if $\mathbf{z}$ is such that $\ell_N(\mathbf{z}) < x$, one must solve (3.11) numerically (an easy exercise in root-finding). Observe that since the left-hand side of (3.11) is increasing in $\hat{\theta}_{\mathbf{z}}$ and is equal to $\ell_N(\mathbf{z})$ when $\hat{\theta}_{\mathbf{z}} = 0$ it follows that $\hat{\theta}_{\mathbf{z}} > 0$ whenever $\mathbf{z}$ is such that $\ell_N(\mathbf{z}) < x$. Intuitively, if the simulated value $\mathbf{z}$ is such that $\mathbb{E}[L_N|\mathbf{Z} = \mathbf{z}] > x$ then large losses are no longer rare and one need not adjust conditional default probabilities. Otherwise one increases default probabilities just enough to ensure that the expected loss is exactly x. It is prudent to note that this is identical to the procedure suggested by Glasserman and Li (2005) in the context of the Gaussian copula model.

Lemma 3.4.1. *We have* $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}] \geq \max(\mathbb{E}[L_N|\mathbf{Z}], x)$.

Proof. Since $L_N > x$ almost surely under $\widetilde{\mathbb{P}}$, it follows that $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}] \geq x$. Convergence as $x \to \sum_{i=1}^{N} c_i$ follows from the fact that the distribution of L_N under $\widetilde{\mathbb{P}}$ converges to a point mass at $\sum_{i=1}^{N} c_i$ as $x \to \sum_{i=1}^{N} c_i$. Next, observe that by the so-called abstract version of Bayes' rule one has

$$\widetilde{\mathbb{E}}[L_N|\mathbf{Z}] = \frac{\mathbb{E}[L_N \mathbf{1}_{\{L_N \geq x\}}|\mathbf{Z}]}{\mathbb{P}(L_N \geq x|\mathbf{Z})} \, .$$

Thus

$$\begin{aligned} \mathbb{E}[L_N|\mathbf{Z}] &= \mathbb{E}[L_N \mathbf{1}_{\{L_N \geq x\}}|Z] + \mathbb{E}[L_N \mathbf{1}_{\{L_N < x\}}|Z] \\ &\leq \widetilde{\mathbb{E}}[L_N|\mathbf{Z}]\mathbb{P}(L_N \geq x|\mathbf{Z}) + x\mathbb{P}(L_N < x|\mathbf{Z}) \\ &= \widetilde{\mathbb{E}}[L_N|\mathbf{Z}] - (\widetilde{\mathbb{E}}[L_N|\mathbf{Z}] - x)\mathbb{P}(L_N < x|\mathbf{Z}) \\ &\leq \widetilde{\mathbb{E}}[L_N|\mathbf{Z}] \, . \end{aligned}$$

Since $\widetilde{\mathbb{P}}$ converges weakly to $\mathbb{P}$ as $x \to 0$, equality obtains as $x \to 0$. □

3.5 Normal Copula Model

In this section we apply our methodology in the context of the normal copula model and benchmark its performance against the gold standard (and in many cases asymptotically optimal, in a certain sense) developed by Glasserman and Li (2005). In this model $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, where $\mathbf{0}_d$ and $\mathbf{I}_d$ denote the d-dimensional zero vector and identity matrix, respectively, $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. standard normal random variables independent of $\mathbf{Z}$, and $Y_i = \mathbf{1}_{\{X_i < \Phi^{-1}(PD_i)\}}$, where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution. Conditional default probabilities are given by

$$p_i(\mathbf{z}) := \Phi\left(\frac{\Phi^{-1}(PD_i) - a_i^T \mathbf{z}}{\sqrt{1 - |a_i|^2}}\right) = \mathbb{P}(Y_i = 1|\mathbf{Z} = \mathbf{z}). \tag{3.12}$$

The asymptotic approximation to L_N is $\ell_N(\mathbf{Z}) = \sum_{i=1}^N c_i p_i(\mathbf{Z})$, and Figure 3.1 illustrates the good agreement between these quantities. For simplicity and without loss of generality we assume that recovery rates are zero, i.e. that $\sum_{i=1}^n c_i = 1$.

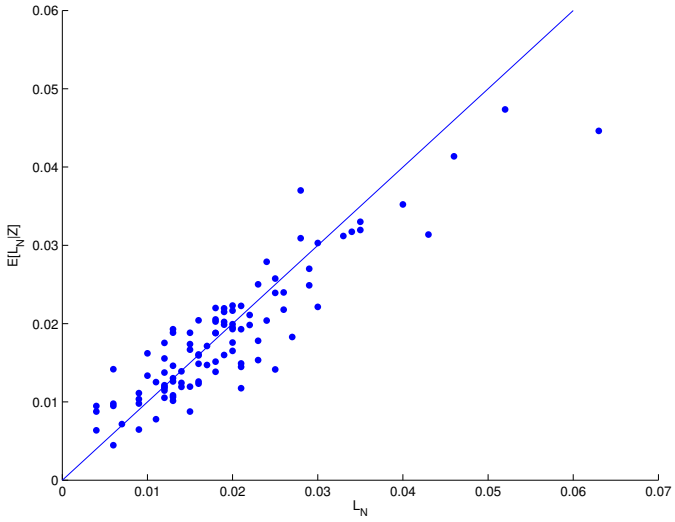

Figure 3.1: 100 *simulated pairs* $(L_N, \mathbb{E}[L_N|\mathbf{Z}])$ *in the single-factor (i.e.* $d = 1$*) homogeneous Gaussian copula model. Parameters are* $N = 1000$, $PD_i = PD = 0.02$, *and* $a_i = 0.2$ *for all* i. *The empirical correlation between the two series exceeds 90%.*

We restrict ourselves to candidates under which the components of $\mathbf{Z}$ are uncorrelated normal with unit variance, i.e. the IS distribution amounts to a shift of the mean of the systematic factors (numerical evidence indicates that allowing for non-unit variances does not produce notable efficiency gains). The natural sufficient statistic for the parameteric family of densities $\{\phi_\mu(\mathbf{z}) : \mu \in \mathbb{R}^d\}$, where $\phi_\mu(\mathbf{z})$ is the density of $N(\mu, \mathbf{I}_d)$, is $f(\mathbf{z}) = \mathbf{z}$ and the natural parameter is μ. Thus $\mathbf{R} = \mathbf{Z}$, $\lambda = \mu$ and $\Psi(\mu) = \frac{1}{2}|\mu|^2$.

The candidate family is parametrized by a vector $\mu \in \mathbb{R}^d$ and a function $\theta_\bullet : \mathbb{R}^d \mapsto \mathbb{R}$, and members are defined by RNDs of the form

$$\frac{d\mathbb{P}^{\mu,\theta_\bullet}}{d\mathbb{P}} = \mathbb{G}(\mu, \mathbf{Z}) \cdot \mathbb{B}(\theta_{\mathbf{Z}}, \mathbf{Z}, \mathbf{Y}), \tag{3.13}$$

where $\mathbb{G}(\mu, \mathbf{z}) = \exp(\mu^T \mathbf{z} - \frac{1}{2}\mu^T \mu)$, $\mathbb{B}(\theta_\bullet, \mathbf{z}, \mathbf{y}) = \exp\left(\theta_{\mathbf{z}} \cdot \mathbf{c}^T \mathbf{y} - \psi_{\mathbf{z}}(\theta_{\mathbf{z}})\right)$ and $\psi_{\mathbf{z}}(\theta) = \sum_{i=1}^N \log\left(1 + p_i(\mathbf{z})[e^{\theta c_i} - 1]\right)$. Under $\mathbb{P}^{\mu,\theta_\bullet}$ (i) the distribution of $\mathbf{Z}$ is $\mathcal{N}(\mu, \mathbf{I}_d)$ and (ii) conditional on $\mathbf{Z} = \mathbf{z}$, $\mathbf{Y}$ is a vector of independent Bernoullis with success probabilities given by (3.6).

Glasserman and Li (2005) use the same candidate family but different criteria for parameter selection. In what follows we let μ^{**} and $\theta_\bullet^{**}$ be the values of the parameters suggested by the Glasserman and Li (2005) criteria, and $\hat{\hat{\mu}}$ and $\hat{\hat{\theta}}_\bullet$ denote approximations suggested by those authors.

3.5.1 Computing $\hat{\mu}$

The suggested approximation to μ^* is $\hat{\mu} = \mathbb{E}[\mathbf{Z}|\ell_N(\mathbf{Z}) \geq x]$, and we recall that it is based on the fact that the discrepancy between L_N and $\mathbb{E}[L_N|\mathbf{Z}]$ should be small if N is large. As illustrated in Figure 3.1 the agreement between these two quantities is quite good.

$\hat{\mu}$ is typically not available in closed form, but is easily and rapidly approximated using Algorithm 4 in Appendix E. The algorithm begins by maximizing $\phi(\mathbf{z})$ subject to the constraint that $\ell_N(\mathbf{z}) = x$ (an easy numerical optimization problem that can be motivated by the Laplace principle) to develop an initial approximation which helps identifying the most likely point on the boundary, and then refining this initial estimate with Monte Carlo. Since $\hat{\mu}$ is precomputed these initial steps do not have a noticeable impact on the overall CPU time required to implement the entire algorithm.

Glasserman and Li (2005) identify $\mu^{**} = \arg\max_{\mathbf{z}\in\mathbb{R}^d} \mathbb{P}(L_N \geq x|\mathbf{Z} = \mathbf{z})\phi(\mathbf{z})$ as the optimal value of the parameter and suggest a variety of approximations to μ^{**}. The simplest approximation is the so-called constant approximation $\hat{\hat{\mu}}_{\text{CON}} := \arg\max\{\phi(\mathbf{z}) : \mathbf{z} \in \mathbb{R}^d,\ \ell_N(\mathbf{z}) \geq x\}$. The following result suggests that, in practice, the two methodologies can be expected to provide similar first-stage tilt parameters.

Proposition 3.5.1. *In the single-factor homogeneous model, that is when $d = 1$ and parameter values do not vary across obligors, we have $\hat{\hat{\mu}}_{\text{CON}} \sim \hat{\mu}$ as $x \to 1$.*

Proof. Under the stated conditions $\mathbf{z} \in \mathbb{R}$, the function $\ell_N(\mathbf{z})$ is monotone increasing,

whence $\hat{\hat{\mu}}_{\text{CON}} = \ell_N^{-1}(x) \cdot \mathbf{1}_{\{\mathbf{z} \geq 0\}}$ and $\hat{\mu} = \mathbb{E}[\mathbf{Z} | \mathbf{Z} \leq \ell_N^{-1}(x)] = -\frac{\phi(\ell_N^{-1}(x))}{\Phi(\ell_N^{-1}(x))}$. Since $-\frac{\phi(x)}{\Phi(x)} \sim x$ as $x \to -\infty$ and $\ell_N^{-1}(1) = -\infty$ it follows that $\hat{\hat{\mu}}_{\text{CON}} \sim \hat{\mu}$ as $x \to 1$. □

3.5.2 Computing $\hat{\theta}_\bullet$

Recall that our approximation $\hat{\theta}_\bullet$ is based on the estimate $\widetilde{\mathbb{E}}[L_N | \mathbf{Z}] \approx \max(\ell_N(\mathbf{Z}), x)$. Figure 3.2 illustrates that the agreement between these quantities is perfect as $x \to 0$ and $x \to 1$, and is also quite good for intermediate values of x. Note also that the most extreme error occurs when $x = \mathbb{E}[L_N]$, which in practice will be quite far from values of interest. As noted in Section 3.4.2 our suggested approximation to $\theta_\bullet^*$ coincides with that of Glasserman and Li (2005), that is $\hat{\theta}_\bullet = \hat{\hat{\theta}}_\bullet$.

In the single-factor homogeneous model, that is when $d = 1$ and parameter values do not vary across obligors, $\hat{\theta}_{\mathbf{z}}$ can be computed in closed form. Indeed in this case one has $\hat{\theta}_{\mathbf{z}} = N \log\left(\frac{x(1-p(\mathbf{z}))}{(1-x)p(\mathbf{z})}\right) \cdot \mathbf{1}_{\{\ell(\mathbf{z}) < x\}}$, whence the conditional default probability applied in the second stage is

$$q(\mathbf{z}) = x \cdot \mathbf{1}_{\{\ell(\mathbf{z}) < x\}} + p(\mathbf{z}) \cdot \mathbf{1}_{\{\ell(\mathbf{z}) \geq x\}}. \tag{3.14}$$

In other cases one must numerically determine $\hat{\theta}_{\mathbf{z}}$.

3.5.3 Numerical Results

In practice it is common to aggregate obligors into different groups and assume homogeneity within each group (this is often necessary given limited computational budgets). To this end we now suppose that the portfolio is composed of G groups defined by (i) common exposure, (ii) common factor loadings and (iii) common default probabilities. For $1 \leq g \leq G$ let E_g be the exposure to each member of group g, N_g the number of obligors in group g and a_g the vector of factor loadings for each

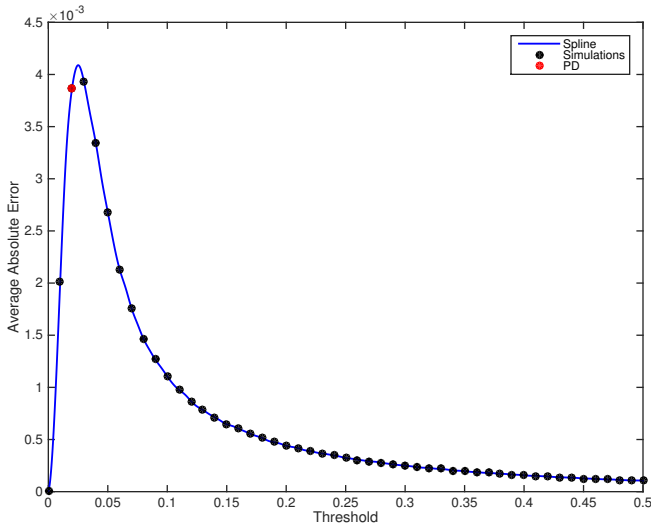

Figure 3.2: *Average absolute error between* $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}=\mathbf{z}]$ *and* $\max\{\mathbb{E}[L_N|\mathbf{Z}=\mathbf{z}], x\}$ *in the single-factor (i.e.* $d=1$*) Gaussian copula model. Parameters are* $N=200$, $PD_i=0.02$, $a_i=0.2$ *for all* i*. For each value of* x*, we simulated* $10,000$ *values of* $\mathbf{Z}$ *and for each value of* $\mathbf{Z}$ *we computed* $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}=\mathbf{z}]$ *using the fact that* $N\cdot L_N|\mathbf{Z}=\mathbf{z}\sim \mathrm{Bin}(N,p(\mathbf{z}))$*. The error* $\widetilde{E}[L_N|\mathbf{Z}]-\max(x,\ell_N(\mathbf{z}))$ *was then averaged across all simulated values and the solid blue line is a cubic spline fitted to that averaged data. The maximum error is located near* $x=PD$ *(point in red), and we recall that in the rare event setting one is typically interested in* $x\gg PD$*.*

member of group g. Conditional default probabilities for group g are therefore given by

$$p^{(g)}(\mathbf{z})=\Phi\left(\frac{\Phi^{-1}(PD_g)-a_g^T\mathbf{z}}{\sqrt{1-|a_g|^2}}\right)\ .$$

The portfolio loss is $L_N=\sum_{g=1}^G w_g\cdot L_N^{(g)}$, where $w_g=E_gN_g/\sum_{k=1}^G E_kN_k$ is the relative exposure to group g, $L_N^{(g)}=\frac{1}{N_g}\sum_{i=1}^{N_g}Y_{i,g}$ and $Y_{i,g}$ is the default indicator for the i^{th} obligor in group g. The asymptotic approximation to L_N is therefore $\ell_N(\mathbf{Z})=\sum_{g=1}^G w_gp^{(g)}(\mathbf{Z})$. In the notation of Section 3.2 we have $c_{i,g}=w_g/N_g$. Recall that in order to estimate $\hat{\mu}$ we use Algorithm 4 in Appendix E, and for a given value of $\mathbf{z}$ we estimate $\hat{\theta}_{\mathbf{z}}$ by solving the following equation for θ

$$\sum_{g=1}^G\frac{w_gp_g(\mathbf{z})e^{w_g\theta/N_g}}{1+p_g(\mathbf{z})\left(e^{w_g\theta/N_g}-1\right)}=\max(\ell_N(\mathbf{z}),x)\ .$$

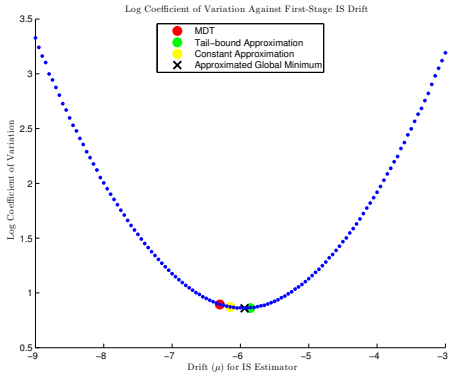

Figure 3.3: *This figure illustrates the near-optimal performance of the proposed algorithm in the single-factor homogeneous Gaussian model. The horizontal axis is the tilt parameter used in the first stage (in all cases* (3.14) *is used in the second stage) and the vertical axis is the log of the estimated coefficient of variation (estimated using* $1,000,000$ *simulations). Other parameters are* $N = 1000$*,* $x = 0.2$*,* $a_i = 0.2$*,* $PD_i = 0.02$ *for all* i*.*

Figures 3.3 and 3.4 illustrate that, in low dimensions at least, among all IS estimators that use an exponential tilt in the first stage and $\hat{\theta}_\bullet = \hat{\hat{\theta}}_\bullet$ in the second stage, the performance of the proposed algorithm is very nearly optimal. For the values used to produce Figure 3.3 it is worth noting that the expected difference between $\widetilde{\mathbb{E}}[L_N|\mathbf{Z}]$ and $\max(\ell_N(\mathbf{Z}, x)$ is approximately 5 basis points. Table 3.1 illustrates that in higher dimensions, i.e. when the number of groups and/or risk factors is large, the proposed algorithm offers comparable performance to that of Glasserman and Li (2005), which is known to be asymptotically optimal for most realistic portfolios. As N increases, the proposed algorithm performs even better since the difference between L_N and $\ell(\mathbf{Z})$ vanishes almost surely as $N \to \infty$. Since

3.6 t Copula Model

In this section we apply our methodology to the t copula model. We show that it offers comparable accuracy for substantially less computational time than the gold standard developed by Chan and Kroese (2010). The main reason for this efficiency gain is that, in contrast to Chan and Kroese (2010), the proposed algorithm does not

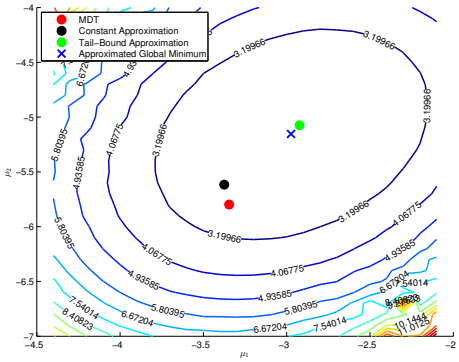

Figure 3.4: *This figure illustrates the near-optimal performance of the proposed algorithm in the two-factor inhomogeneous Gaussian model. Contour plot of $CV(\mu)$ where $CV(\mu)$ is the coefficient of variation assuming the first stage tilt parameter is $\mu \in \mathbb{R}^2$ and the second stage tilt parameter is $\hat{\theta}_\bullet$. $CV(\mu)$ is estimated by Monte Carlo using $1,000,000$ simulations. Parameters are $a_1 = [0.1 \;\; 0.1]^T$, $a_2 = [0.1 \;\; 0.2]^T$, $PD = [0.025 \;\; 0.0145]^T$, $x = 0.2$ with 500 assets distributed equally among each group (i.e. $N_1 = N_2 = 250$).*

Table 3.1: *This table compares the performance of the proposed algorithm to the performance of the Glasserman and Li (2005) algorithm. We assume there are ten groups and fifteen risk factors. Factor loadings, marginal default probabilities and relative exposures can be found in Appendix C. Note that factor loadings are chosen to ensure that $|a_g|^2 \leq 0.11$, meaning a low correlation environment. In all cases obligors are divided equally across groups. i.e. $N_g = N/G$.*

Estimator	Mean	Standard Error	CV
$x = 0.15,\ N = 200$			
Proposed	4.29×10^{-5}	1.45×10^{-4}	3.37
Glasserman and Li	4.33×10^{-5}	8.19×10^{-5}	1.89
$x = 0.30,\ N = 200$			
Proposed	1.70×10^{-9}	6.91×10^{-9}	4.06
Glasserman and Li	1.68×10^{-9}	3.84×10^{-9}	2.28
$x = 0.15,\ N = 2000$			
Proposed	6.85×10^{-6}	1.52×10^{-5}	2.22
Glasserman and Li	6.78×10^{-6}	2.06×10^{-5}	3.04
$x = 0.30,\ N = 2000$			
Proposed	7.94×10^{-11}	2.17×10^{-10}	2.74
Glasserman and Li	7.99×10^{-11}	2.30×10^{-10}	2.88

require sorting high-dimensional vectors, which can be quite time-consuming when the number of obligors is large. Indeed the algorithm of the proposed algorithm is of order N, as opposed to $N\log(N)$ in the Chan and Kroese (2010) algorithm.

3.6.1 Multivariate t Distribution

Suppose that $(\mathbf{Z}, \varepsilon_1, \ldots, \varepsilon_N)$ has a multivariate t distribution with zero mean and identity covariance matrix, with ν degrees of freedom. Then we have the representation $(\mathbf{Z}, \varepsilon_1, \ldots, \varepsilon_N) = S \cdot (\hat{\mathbf{Z}}, \hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N)$, where $S = \sqrt{\nu/\chi^2_\nu}$, χ^2_ν follows a chi-squared distribution with ν degrees of freedom, $\hat{\mathbf{Z}} \sim N(\mathbf{0}_d, I_d)$, $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N$ are i.i.d. standard normal random variables.

In this case $X_i = S \cdot (a_i^T \hat{\mathbf{Z}} + \sqrt{1 - |a_i|^2}\hat{\varepsilon}_i)$ and $Y_i = \mathbf{1}_{\{X_i < t_\nu^{-1}(PD_i)\}}$, where $t_\nu^{-1}(\cdot)$ denotes the inverse cdf of the t distribution with ν degrees of freedom. Conditional default probabilities are given by

$$p_i(\hat{\mathbf{z}}, s) = \mathbb{P}(Y_i = 1 | \hat{\mathbf{Z}} = \hat{\mathbf{z}},\ S = s) = \Phi\left(\frac{t_\nu^{-1}(PD_i)s^{-1} - a_i^T\hat{\mathbf{z}}}{\sqrt{1 - |a_i|^2}}\right), \tag{3.15}$$

and we note that one must condition on both S and $\hat{\mathbf{Z}}$ in order to induce independence among default indicators. The asymptotic approximation to L_N here is $\ell_N(\hat{\mathbf{Z}}, S) = \sum_{i=1}^N c_i p_i(\hat{\mathbf{z}}, s)$. The agreement between L_N and its asymptotic approximation is at least as good as in the Gaussian case, see Figure 3.5 for an illustration.

3.6.2 Candidate Family

We restrict ourselves to candidates under which $\hat{\mathbf{Z}} \sim \mathcal{N}(\mu, \mathbf{I}_d)$ independently of $S \sim \sqrt{(\nu + \alpha)/\chi^2_{\nu+\alpha}}$, where $\mu \in \mathbb{R}^d$ and $\alpha \in (-\nu, \infty)$. In other words we shift the mean of $\hat{\mathbf{Z}}$ and adjust the degrees of freedom of S, retaining independence between the two.

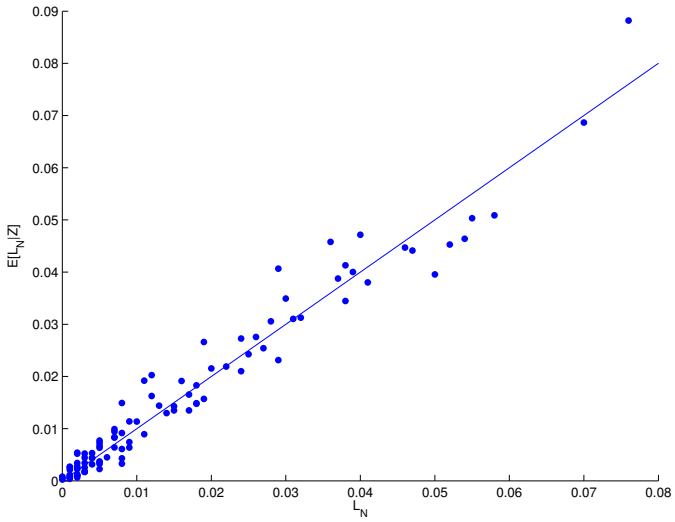

Figure 3.5: 100 *simulated pairs* $(L_N, \mathbb{E}[L_N|\mathbf{Z}])$ *in the t copula model. Parameters are* $N = 1000$, $\nu = 15$, $PD_i = PD = 0.02$ *and* $a_i = 0.3$ *for all* i. *The empirical correlation between the two series exceeds 98%.*

For fixed ν let $g_{\nu+\alpha}(s)$ be the density of $\sqrt{(\nu+\alpha)/\chi^2_{\nu+\alpha}}$, that is

$$g_{\nu+\alpha}(s) = \frac{2((\alpha+\nu)/2)^{(\alpha+\nu)/2}}{\Gamma((\alpha+\nu)/2)} \exp(-(\alpha+\nu)[\log(s) + (2s^2)^{-1}]) \frac{1}{s} , \qquad s > 0 .$$

where $\Gamma(\cdot)$ is the gamma function. This is clearly an exponential family with natural parameter α and natural sufficient statistic $T(s) = -\log(s) - (2s^2)^{-1}$. Using this representation and elementary properties of exponential families (Section 1.6 in Bickel and Doksum (2001)) it is also clear that

$$\mathbb{E}[\exp(\alpha T(S))] = -\frac{\alpha+\nu}{2} \log\left(\frac{\alpha+\nu}{2}\right) + \log\left(\Gamma\left(\frac{\alpha+\nu}{2}\right)\right) + \frac{\nu}{2} \log\left(\frac{\nu}{2}\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) ,$$

whence

$$\mathbb{E}^{\alpha}[-\log(S) - (2S^2)^{-1}] = -\frac{1}{2} - \frac{1}{2} \log\left(\frac{\nu+\alpha}{2}\right) + \frac{1}{2}\gamma\left(\frac{\nu+\alpha}{2}\right) ,$$

where $\gamma(\cdot)$ is the digamma function.

The candidate family is parametrized by $\lambda = (\mu, \alpha) \in \mathbb{R}^d \times (-\nu, \infty)$ and $\theta_{\bullet,\bullet} : \mathbb{R}^{d+1} \mapsto \mathbb{R}$, and is defined by RNDs of the form

$$\frac{d\mathbb{P}^{\alpha,\mu,\theta_{\bullet,\bullet}}}{d\mathbb{P}} = \mathbb{S}(\alpha, S) \cdot \mathbb{G}(\mu, \hat{\mathbf{Z}}) \cdot \mathbb{B}(\theta_{\hat{\mathbf{Z}},S}, S, \hat{\mathbf{Z}}, \mathbf{Y}) \ ,$$

where

$$\mathbb{S}(\alpha, S) = \frac{g_{\alpha+\nu}(S)}{g_\nu(S)} = \frac{\Gamma(\nu/2)}{\Gamma((\alpha+\nu)/2)} \cdot \frac{((\alpha+\nu)/2)^{(\alpha+\nu)/2}}{(\nu/2)^{\nu/2}} \cdot \exp(-\alpha[\log(S) + (2S^2)^{-1}]) \ ,$$

$\mathbb{G}(\mu, \mathbf{Z}) = \exp(\mu^T \hat{\mathbf{z}} - \frac{1}{2}\mu^T \mu)$, $\mathbb{B}(\theta_{\bullet,\bullet}, s, \hat{\mathbf{z}}, \mathbf{y}) = \exp\left(\theta_{\hat{\mathbf{z}},s} \mathbf{c}^T \mathbf{y} - \psi_{\hat{\mathbf{z}},s}(\theta_{\hat{\mathbf{z}},s})\right)$ and

$$\psi_{\hat{\mathbf{z}},s}(\theta) = \sum_{i=1}^{N} \log(1 + p_i(\hat{\mathbf{z}}, s)(\exp(\theta \cdot c_i) - 1)) \ .$$

3.6.3 Computing $\hat{\mu}$ and $\hat{\alpha}$

The suggested approximation to μ^* is $\hat{\mu} = \mathbb{E}[\mathbf{Z} | \ell_N(\hat{\mathbf{Z}}, S) \geq x]$, and the suggested approximation to α^*, $\hat{\alpha}$, will solve

$$\frac{1}{2} + \frac{1}{2}\log\left(\frac{\nu+\hat{\alpha}}{2}\right) - \frac{1}{2}\gamma\left(\frac{\nu+\hat{\alpha}}{2}\right) = \mathbb{E}[\log S + (2S^2)^{-1} | \ell_N(\hat{\mathbf{Z}}, S) \geq x] \ . \qquad (3.16)$$

In general $\hat{\mu}$ and $\hat{\alpha}$ must be computed numerically. One option is to apply the adaptive CE method by noting that (3.9) reduces here to

$$\mu_{n+1} = \frac{\sum_{k=1}^{K} \hat{\mathbf{Z}}_k \mathbf{1}_{\{\ell_N(\hat{\mathbf{Z}}_k, S_k) \geq x_n\}} [\mathbb{G}(\mu_n, \mathbf{Z}_k)\mathbb{S}(\alpha_n, S_k)]^{-1}}{\sum_{k=1}^{K} \mathbf{1}_{\{\ell_N(\hat{\mathbf{Z}}_k, S_k) \geq x_n\}} [\mathbb{G}(\mu_n, \mathbf{Z}_k)\mathbb{S}(\alpha_n, S_k)]^{-1}} \ ,$$

$$-\tfrac{1}{2} - \tfrac{1}{2}\log\left(\tfrac{\nu+\alpha_{n+1}}{2}\right) + \tfrac{1}{2}\gamma\left(\tfrac{\nu+\alpha_{n+1}}{2}\right) = \frac{\sum_{k=1}^{K} T(S_k)\mathbf{1}_{\{\ell_N(\hat{\mathbf{Z}}_k,S_k)\geq x_n\}}[\mathbb{G}(\mu_n,\mathbf{Z}_k)\mathbb{S}(\alpha_n,S_k)]^{-1}}{\sum_{k=1}^{K}\mathbf{1}_{\{\ell_N(\hat{\mathbf{Z}}_k,S_k)\geq x_n\}}[\mathbb{G}(\mu_n,\mathbf{Z}_k)\mathbb{S}(\alpha_n,S_k)]^{-1}} ,$$

where $(\hat{\mathbf{Z}}_k, S_k), \ldots, (\hat{\mathbf{Z}}_K, S_K)$ are i.i.d. copies from the (μ_n, α_n)-tilted distribution of $(\hat{\mathbf{Z}}, S)$. It is worth noting that elementary properties of exponential families ensure that the left-hand sides in above display is monotone in α_{n+1}. Thus, given simulated values of the systematic risk factors, solving for α_{n+1} is an easy root-finding exercise.

In the homogeneous single-factor case, that is when $d = 1$ and parameter values do not vary across obligors, both $\mathbb{E}[\hat{\mathbf{Z}}|\ell_N(\hat{\mathbf{Z}}, S) \geq x]$ and $\mathbb{E}[T(S)|\ell_N(\hat{\mathbf{Z}}, S) \geq x]$ can be put in semi-analytic form and the adaptive CE algorithm can be avoided. Indeed in this case we have

$$\ell_N(\hat{\mathbf{z}}, s) = \Phi\left(\frac{t_\nu^{-1}(PD)s^{-1} - \rho\hat{\mathbf{z}}}{\sqrt{1-\rho^2}}\right) ,$$

and if $PD < 0.5$ (so that $t_\nu^{-1}(PD) < 0$) then

$$\ell_N(\hat{\mathbf{z}}, s) \geq x \quad \text{if and only if} \quad \hat{\mathbf{z}} \leq z^*(s, x) := \frac{\Phi^{-1}(x)\sqrt{1-\rho^2} - t_\nu^{-1}(PD)s^{-1}}{\rho} .$$

By the tower property the requisite building blocks are given by

$$\mathbb{P}(\ell(\hat{\mathbf{Z}}, S) > x) = \int_0^\infty \Phi(z^*(s, x))\mathbb{P}(S \in ds) ,$$

$$\mathbb{E}\left[\hat{\mathbf{Z}} \cdot \mathbf{1}_{\{\ell(\hat{\mathbf{z}},S)>x]\}}\right] = -\int_0^\infty \phi(z^*(s, x))\mathbb{P}(S \in ds) ,$$

and

$$\mathbb{E}\left[T(S)\cdot\mathbf{1}_{\{\ell(\hat{\mathbf{z}},S)>x\}}\right]=-\int_0^\infty\left(\log s+(2s^2)^{-1}\right)\Phi(z^*(s,x))\mathbb{P}(S\in ds)\ ,$$

3.6.4 Computing $\hat{\theta}_{\bullet,\bullet}$

If $(\hat{\mathbf{z}},s)$ is such that $\ell_N(\hat{\mathbf{z}},s)\geq x$ then $\hat{\theta}_{\mathbf{z},s}=0$. Otherwise $\hat{\theta}_{\mathbf{z},s}=0$ must be determined numerically as the solution to

$$\sum_{i=1}^N\frac{c_ip_i(\hat{\mathbf{z}},s)e^{c_i\hat{\theta}_{\mathbf{z},s}}}{1+p_i(\hat{\mathbf{z}},s)[e^{c_i\hat{\theta}_{\mathbf{z},s}}-1]}=x\ .$$

3.6.5 Numerical Results

As in Chan and Kroese (2010), henceforth referred to as CK, we consider the homogeneous case where $d=1$ and parameter values do not vary across obligors. Table 3.2 presents the results. We see that the CK algorithm provides a coefficient of variation (CV) that is smaller than the proposed algorithm except when x is large[2]. Keeping in mind the size of the probabilities being estimated, however, one sees that the proposed algorithm provides substantial variance reduction relative to crude Monte Carlo and performance that is broadly comparable to CK.

Given the comparable levels of performance, the proposed algorithm has a clear advantage over CK when the number of obligors is large. Indeed when there are 5,000 obligors in the portfolio (a very realistic number) the proposed algorithm is nearly 7 times faster than the CK algorithm. The reason is that the CK algorithm requires one to sort high-dimensional vectors but the proposed algorithm does not. The implication is that the computational time required for the CK algorithm is of order

[2]This is because the CK algorithm assumes that the probability of default vanishes as the number of obligors goes to infinity, they are not considering the situation where the threshold x goes to 1.

Table 3.2: *This table compares the performance of the proposed algorithm to the performance of the algorithm proposed by Chan and Kroese (2010). Base parameters are* $N = 2000$, $\nu = 15$, $a_i = 0.3$ *and* $PD_i = 0.029$ *for all* i, $x = 0.4$. *CPU Time is given in seconds.*

	Proposed Algorithm				Chan and Kroese (2010)			
m	Mean	CV	Time	Time·CV	Mean	CV	Time	Time·CV
500	4.29×10^{-5}	3.85	10.3	39.78	4.27×10^{-5}	2.06	34.0	70.14
1625	3.93×10^{-5}	3.91	14.1	55.24	3.93×10^{-5}	1.08	60.0	64.64
2750	3.95×10^{-5}	3.96	18.2	71.92	3.90×10^{-5}	0.471	91.7	43.16
3875	3.92×10^{-5}	3.94	21.6	85.30	3.88×10^{-5}	0.413	125	51.84
5000	3.93×10^{-5}	3.93	24.2	95.08	3.88×10^{-5}	0.546	168	91.63
ρ	Mean	CV	Time	Time·CV	Mean	CV	Time	Time·CV
0.1	5.66×10^{-9}	5.77	14.7	84.99	5.53×10^{-9}	0.89	65.26	57.95
0.2	1.27×10^{-6}	4.77	14.8	70.58	1.27×10^{-6}	4.45	70.27	312.87
0.3	3.88×10^{-5}	3.94	14.7	57.89	3.94×10^{-5}	1.45	71.78	103.71
0.4	3.65×10^{-4}	3.29	14.7	48.46	3.66×10^{-4}	0.17	78.47	13.41
0.5	1.63×10^{-3}	2.85	14.7	42.07	1.64×10^{-3}	0.59	72.15	42.82
x	Mean	CV	Time	Time·CV	Mean	CV	Time	Time·CV
0.1	4.53×10^{-2}	2.21	14.7	32.53	4.52×10^{-2}	0.14	71.21	9.98
0.225	2.36×10^{-3}	3.1	14.6	45.37	2.37×10^{-3}	0.18	70.450	12.72
0.35	1.32×10^{-4}	3.7	14.7	54.53	1.32×10^{-4}	0.777	70.24	54.54
0.475	5.66×10^{-6}	4.23	14.7	62.13	5.73×10^{-6}	0.591	71.12	42.04
0.6	1.44×10^{-7}	4.76	15.0	71.59	1.37×10^{-7}	18.02	69.52	1252.50
ν	Mean	CV	Time	Time·CV	Mean	CV	Time	Time·CV
2	8.55×10^{-3}	2.75	14.6	40.22	8.66×10^{-3}	0.177	69.6	12.31
5	1.21×10^{-3}	3.29	14.9	49.07	1.21×10^{-3}	0.172	70.7	12.15
7	4.63×10^{-4}	3.52	14.6	51.59	4.65×10^{-4}	0.256	69.4	17.79
10	1.51×10^{-4}	3.72	14.8	55.13	1.52×10^{-4}	0.286	71.2	20.36
12	8.25×10^{-5}	3.85	14.8	56.88	8.33×10^{-5}	0.602	70.3	73.73

$N \log(N)$, whereas the time required to implement the proposed algorithm is of order N.

A further advantage of the proposed algorithm relative to Chan and Kroese (2010) is that it can be used to estimate more general risk measures, whereas Chan and Kroese (2010) is specifically designed to estimate loss probabilities. For instance a consistent IS estimator for the tail expectation $\mathbb{E}[L_N | L_N \geq x]$ is

$$\frac{\sum_{i=1}^{M} L_N^{(i)} \cdot \mathbf{1}_{\{L_N^{(i)} \geq x\}} \cdot \Lambda_i}{\sum_{i=1}^{M} \mathbf{1}_{\{L_N^{(i)} \geq x\}} \cdot \Lambda_i}, \tag{3.17}$$

Table 3.3: *This table compares the performance of the IS estimator* (3.17) *to that of the crude estimator, when estimating the conditional tail expectation* $\mathbb{E}[L_N|L_N \geq x]$ *in the t copula model. Portfolio consists of* $N = 500$ *obligors, evenly distributed among* $G = 10$ *groups (i.e.* $N_g = 50$ *for each g), and there are* $D = 15$ *Gaussian risk factors. Degrees of freedom are* $\nu = 5$*. Default probabilities, factor loadings and exposures are given in Appendix C, and loss given defaults are all set to 100%. For each value of x in the table, 20 realizations of each estimator were simulated using* $M = 10,000$ *in each case. Reported mean and standard deviation are the sample mean and sample standard deviation of the 20 realizations. A reported value of NA means that at least one of the 20 samples of 10,000 did not contain a single observation for which* $L_N \geq x$*, in which case both numerator and denominator of the crude estimator is zero. The reported value of* $\mathbb{P}(L_N \geq x)$ *is based on an IS estimate using a sample of size* $200,000$*.*

		Mean		Standard Deviation		
x	$\mathbb{P}(L_N \geq x)$	Crude	IS	Crude	IS	Ratio
0.20	8.2×10^{-3}	0.263	0.263	6.1×10^{-3}	1.4×10^{-3}	4.3
0.24	4.1×10^{-3}	0.308	0.306	9.2×10^{-3}	2.1×10^{-3}	4.5
0.29	2.1×10^{-3}	0.352	0.348	1.1×10^{-2}	1.4×10^{-3}	8.3
0.33	9.7×10^{-4}	0.392	0.389	1.6×10^{-2}	1.4×10^{-3}	11.2
0.38	5.1×10^{-4}	NA	0.430	NA	1.1×10^{-3}	NA
0.42	2.2×10^{-4}	NA	0.470	NA	1.8×10^{-3}	NA
0.47	8.5×10^{-5}	NA	0.511	NA	1.3×10^{-3}	NA
0.51	3.4×10^{-5}	NA	0.552	NA	2.2×10^{-3}	NA
0.56	1.2×10^{-5}	NA	0.592	NA	2.0×10^{-3}	NA
0.60	3.7×10^{-6}	NA	0.633	NA	1.4×10^{-3}	NA

where Λ_i is the RND given by the right-hand side of (3.13) and where $L_N^{(1)}, L_N^{(2)}, \ldots, L_N^{(M)}$ are M independent copies of the loss generated using our proposed IS estimator. As illustrated in Table 3.3, which assumes 500 obligors evenly distributed among ten groups ($G = 10$) and fifteen Gaussian risk factors ($D = 15$), (3.17) performs quite well over a wide range of thresholds x. Indeed it is substantially more accurate than the crude estimator and the relative performance of the IS estimator improves as one moves deeper into the tail of the loss distribution.

3.7 Concluding Remarks

This chapter develops a novel and general importance sampling algorithm for estimating portfolio loss probabilities in linear factor models, that can be applied in a wide variety of models. In the Gaussian and t cases the algorithm offers comparable

performance in terms of accuracy, with a notable advantage in computational time in the t case. Note that the cross-entropy minimization problem we have presented in this section has a unique solution. However, in a multi-dimensional setting, this can be problematic if there is many regions, see the discussion in Glasserman et al. (2008) for more details. Future research will involve applying the cross-entropy methodology in a more general setting, where we might have a non-unique solution.

The algorithm is also limited to the situation where $L_N|\mathbf{Z}$ is equivalent to tilting $p_i(\mathbf{z})$, which is not the case in presence of PD-LGD correlation. In the next chapter, we will discuss how one can combine rejection sampling with our two-stage IS algorithm in order to get a fast and efficient algorithm for models with PD-LGD correlation.

Bibliography

S. Asmussen and P.W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.

Achal Bassamboo, Sandeep Juneja, and Assaf Zeevi. Portfolio Credit Risk with Extremal Dependence: Asymptotic Analysis and Efficient Simulation . *Oper. Res.*, 56(3):593–606, July 2006.

P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basice Ideas and Selected Topics*, volume 1. Prentice Hall, second edition, 2001.

Joshua C.C. Chan and Dirk P. Kroese. Efficient Estimation of Large Portfolio Loss Probabilities in t-Copula Models. *European J. of Oper. Res.*, 205(2):361–367, September 2010.

Paul Glasserman and Jingyi Li. Importance Sampling for Portfolio Credit Risk. *Management Science*, 51(11):1643–1656, November 2005.

Paul Glasserman, Wanmo Kang, and Perwez Shahabuddin. Fast Simulation of Multifactor Portfolio Credit Risk. *Oper. Res.*, 56(5):1200–1217, 2008.

M.B. Gordy. A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation*, 12:199–232, 2003.

J.H. Hu, M.C. Fu, and S.I. Marcus. A model reference adaptive search method for global optimization. *Operations Research*, 55:549–568, 2007.

Sandro Merino and Mark Nyfeler. Calculating portfolio loss. *Risk*, 15(8):82–86, August 2002.

Alexandre Scott and Adam Metzler. A General Importance Sampling Algorithm

for Estimating Portfolio Loss Probabilities in Linear Factor Models. Submitted, February 2015.

Chapter 4

Variance Reduction for Models with PD-LGD correlation

4.1 Introduction

The accurate estimation of loss probabilities in a portfolio is, in general, a difficult problem. This is mainly due to the fact that (i) typical portfolios have a large number of obligors and (ii) there is a complicated dependence structure between the probability of default (PD) and the loss given default (LGD). When there is no PD-LGD correlation, it is possible to use a convolution based method (see for example Merino and Nyfeler (2002)) to compute loss probabilities, however when either the number of obligors is large or systematic risk factors is large, this method becomes intractable. Another possibility is to use Monte Carlo simulations, however typical Monte Carlo simulations have two major issues. First, the variability of the estimator is often large relative to the loss probability and secondly, when the default probability is small, it takes a large number of simulations before observing even one default. To address these issues, we can use Importance Sampling (IS) to improve the quality and the speed of the estimator, but one problem remains: how do we choose an efficient IS measure to sample from?

In the context of the normal copula model with known LGD, Glasserman and Li (2005) present a two-stage IS technique for estimating tails of the loss distribution, and prove that this methodology is asymptotically optimal in the homogeneous case. Glasserman et al. (2008) extend this technique to the heterogeneous model by solving a combinatorial problem and using mixtures of IS densities. Scott and Metzler (2015) present an alternative two-stage scheme, based on the CE method (see Rubinstein (2005)[1]) that performs well in both normal and t copula models. Other authors such as Bassamboo et al. (2006), Chan and Kroese (2010; 2011) introduced IS-based for the t copula model.

To the best of our knowledge little progress has been made for models which allow for PD-LGD correlation. As noted by Miu and Ozdemir (2006), failure to account for PD-LGD correlation can materially underestimate capital requirements. Miu and Ozdemir (2006) introduce a model with PD-LGD correlation, and the correlation between the PD and LGD appears to cause problems for the aforementioned algorithms. In this chapter, we introduce a two-stage IS algorithm that performs well in the Miu and Ozdemir (2006) model. The first stage tilts the distribution of the systematic risk factors and the second stage tilts the conditional distribution of the loss, given the systematic risk factors. However, we cannot simulate directly from the tilted conditional distribution of the loss, thus, we introduce a rejection sampling algorithm.

4.2 The Model

In this section, we briefly review the model presented in Miu and Ozdemir (2006). Consider a portfolio with equal exposure of each of N obligors. Let $Y_i \in \{0, 1\}$ be

[1]See Ridder and Rubinstein (2007) and Rubinstein and Kroese (2004) for more details on applications of entropy minimization in the rare event setting.

the default indicator for obligor i and let $LGD_i \in [0,1]$ be the loss-given-default for obligor i. Let L_N be the *percentage loss* defined as

$$L_N = \frac{1}{N}\sum_{i=1}^{N} LGD_i \cdot Y_i.$$

In the model defined in Miu and Ozdemir (2006), we consider both Y_i and LGD_i to be random and correlated and driven by four main factors: the systematic risk factors P, L, and the idiosyncratic risk factors p_i and l_i. The variables (P, L) is a bivariate normal random vector with mean $(0, 0)$, and where $\mathrm{Var}(P) = \mathrm{Var}(L) = 1$, and where

$$\mathrm{Corr}(P, L) \quad = \quad \beta.$$

Furthermore, we suppose that the conditional distribution $(p_i, l_i)|(P, L)$ is a bivariate normal with mean $\eta = (R_{PD}P,\ R_{LGD}L)^T$, where R_{PD} and R_{LGD} are constants in $[0, 1]$, $\mathrm{Var}(p_i|P, L) = \sigma^2_{PD} = 1 - R^2_{PD}$, $\mathrm{Var}(l_i|P, L) = \sigma^2_{LGD} = 1 - R^2_{LGD}$ and correlation

$$\mathrm{Corr}(p_i, l_i|P, L) = \rho_i.$$

We assume that given (P, L), the vector (p_i, l_i) is conditional independent to (p_j, l_j), and that $\beta,\ R_{PD},\ R_{LGD},\ \rho_1, \ldots, \rho_N$ are constants in $[0, 1]$ to insure positive correlation between obligors. The variable LGD_i[2] is then defined as

$$LGD_i = B^{-1}(\Phi(l_i), a, b),$$

where $B(\cdot, a, b)$ is the cumulative distribution function of a beta random variable with shape parameters $a,\ b > 0$ and where $\Phi(\cdot)$ denotes the cumulative distribution

[2]For more information about the choice of LGD_i and how to estimate a, b we refer the reader to Miu and Ozdemir (2006).

function of a standard normal random variable. The default indicator Y_i is expressed as

$$Y_i = \mathbf{1}_{\{p_i < \Phi^{-1}(PD_i)\}} = \begin{cases} 1 & \text{if } p_i < \Phi^{-1}(PD_i), \\ 0 & \text{otherwise,} \end{cases}$$

where PD_i is the marginal and unconditional probability of default of obligor i.

The main subject of this chapter is to accurately estimate

$$\mathbb{P}(L_N > x) = \mathbb{E}\left[\mathbf{1}_{\{L_N > x\}}\right], \tag{4.1}$$

where $x \in [0, 1]$ is a user-defined threshold. To estimate (4.1), we consider applying an IS scheme to the distribution of the loss. The IS candidates are chosen via entropy minimization, following the criteria developed in Scott and Metzler (2015) for the normal copula model. In the next section, we identify an ideal family of IS densities to simulate the loss, and we discuss the challenges related to simulating from candidates in that ideal family. Then, we suggest an alternative IS scheme inspired by the work of Glasserman and Li (2005). This is a two-stage procedure by first applying IS to the distribution of $(P,\, L)$ since

$$L_N \to \mathbb{E}[L_N | P, L] \tag{4.2}$$

as the number of obligors N goes to infinity. Then, we apply a second round of IS to twist the conditional distribution of loss given (P, L).

4.3 Appropriate IS Densities

Recall that the goal is to provide a satisfactory unbiased estimator for (4.1) in the sense that the variability of that estimator is as low as possible relative to the loss

probability. This translates into observing more defaults or equivalently, simulating values of the loss in the region where $\{L_N > x\}$ more frequently. To do this, it seems reasonable to apply IS to the distribution of the main drivers of the portfolio loss, which is the distribution of (P, L) (this is a consequence of (4.2)). To twist the distribution of the loss directly, let us consider an IS measure $\mathbb{Q}$ whose Radon-Nykodym derivative (RND) can be expressed as

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\theta L_N - \Psi(\theta)\right), \tag{4.3}$$

where $\Psi(\theta) = \log \mathbb{E}[\exp(\theta L_N)]$ and note that $\Psi(\theta)$ is the cumulant generating function of the loss L_N. Since L_N is a random variable defined in $(0, 1)$, the dominated convergence theorem ensures that $|\Psi(\theta)| \leq |\theta|$, thus (4.3) is well-defined for every $\theta \in \mathbb{R}$. Choosing an appropriate value for θ is not trivial at all; if one chooses θ such that the estimated value of L_N is much larger than x, than the attached likelihood ratio will be small and could potentially result in a worse estimator (in terms of variability) than the crude Monte Carlo estimator. On the other hand, if the estimated value of L_N is much smaller than x, then we might be unable to observe an occurrence of L_N in the region where $\{L_N > x\}$. Applying the minimum divergence criteria (see Proposition 1 in Scott and Metzler (2015) for more details), we suggest to choose θ to be the solution of

$$\Psi'(\theta) = \mathbb{E}[L_N | L_N > x]. \tag{4.4}$$

Since $\Psi(\theta)$ is convex, there exists a unique solution to (4.4), and let us denote this solution as θ^*. Two problems arise from such methodology. Firstly, obtaining a value for θ^* is extremely complicated numerically speaking, and secondly, the correlation structure between the random variables P, L, $p_1, \ldots, p_N$, $l_1, \ldots, l_N$ makes sampling from a density with RND (4.3) using the inverse transform impossible (unless $\theta = 0$).

To see the first problem, we need to first derive an expression for the left-hand side of (4.4). Let $\Psi(\theta)$ be expressed as

$$\begin{aligned}\Psi(\theta) &= \log \mathbb{E}[\exp(\theta L_N)] \\ &= \log \mathbb{E}[\mathbb{E}[\exp(\theta L_N)|P, L]] \\ &= \log \mathbb{E}\left[\prod_{i=1}^{N} u_i\left(\frac{\theta}{N}, P, L\right)\right]\end{aligned}$$

where $u_i(\theta, P, L) = \mathbb{E}[\theta \cdot \exp(B^{-1}(\Phi(l_i), a, b) \cdot Y_i)|P, L]$. Therefore, $\Psi'(\theta)$ can be expressed as

$$\Psi'(\theta) = \frac{1}{N}\frac{\sum_{k=1}^{N} \mathbb{E}\left[r_k\left(\frac{\theta}{N}, P, L\right)\prod_{j\neq k} u_j\left(\frac{\theta}{N}, P, L\right)\right]}{\mathbb{E}\left[\prod_{i=1}^{N} u_i\left(\frac{\theta}{N}, P, L\right)\right]}, \tag{4.5}$$

where

$$r_k(\theta, P, L) = \frac{\partial u_k}{\partial \theta} = \mathbb{E}\left[B^{-1}(\Phi(l_k), a, b) \cdot Y_k \cdot \exp(\theta \cdot B^{-1}(\Phi(l_k), a, b) \cdot Y_k)|P, L\right].$$

Since N is typically very large in practise, computing (4.5) can be problematic.

Let us attack the second problem independently by supposing that we are able to obtain $\theta^* \in \mathbb{R}$ which satisfies (4.4). Since we cannot sample from a density whose RND is of the form (4.3) using the inverse transform, we have to use rejection sampling in order to generate exactly from this "ideal tilt" distribution. Since the random variable L_N is non-negative, an upper bound for the RND of the "ideal tilt" distribution is given by

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp(\theta^* L_N - \Psi(\theta^*)) \leq \exp\left(\theta^* \cdot \mathbf{1}_{\{\theta^* \in \mathbb{R}^+\}} - \Psi(\theta^*) \cdot \mathbf{1}_{\{\theta^* \in \mathbb{R}^-\}}\right) \tag{4.6}$$

We have $\Psi'(0) = \mathbb{E}[L_N] \leq \mathbb{E}[L_N | L_N > x]$ and since $\Psi'(\theta)$ is an increasing function with respect to θ, we have $\theta^* \in \mathbb{R}^+$. Therefore, (4.6) can be simplified to

$$\frac{d\mathbb{Q}}{d\mathbb{P}} \leq \exp(\theta^*) = M \quad \text{(rejection constant)}. \tag{4.7}$$

Recall that, once the rejection constant M has been determined, the probability of accepting a simulated value generated by the rejection algorithm is $1/M$. Using our "ideal algorithm", the probability of accepting a simulated value is $\exp(-\theta^*)$, which means that using rejection sampling can be extremely slow if θ^* is large.

To work around those two computational issues, we suggest an alternative idea: we first apply a change of measure to the distribution of (P, L), and then, we apply a rejection sampling algorithm to twist the conditional distribution of L_N given (P, L). The criteria for selecting an IS measure for (P, L) is discussed in Section 4.3.2 while the implementation of the rejection sampling algorithm is discussed in Section 4.3.3.

Example 4.3.1. *Consider the homogeneous model defined in Section 4.4 where* $\rho_i \equiv \rho = 0.2$, $\beta = 0.25$, $a = 0.63$, $b = 0.975$, $PD_i \equiv PD = 0.029$, $R_{PD} = 0.25$ *and* $R_{LGD} = \sqrt{0.0588}$. *Then, we can rewrite the right-hand side of* (4.5) *as*

$$\frac{\mathbb{E}\left[r\left(\frac{\theta}{N}, P, L\right) u\left(\frac{\theta}{N}, P, L\right)^{N-1}\right]}{\mathbb{E}\left[u\left(\frac{\theta}{N}, P, L\right)^{N}\right]},$$

where $r_i(\theta, P, L) \equiv r(\theta, P, L)$ *and* $u_i(\theta, P, L) \equiv u(\theta, P, L)$ *for all* i. *Therefore,* θ^* *must be the unique solution of*

$$\frac{\mathbb{E}\left[r\left(\frac{\theta}{N}, P, L\right) u\left(\frac{\theta}{N}, P, L\right)^{N-1}\right]}{\mathbb{E}\left[u\left(\frac{\theta}{N}, P, L\right)^{N}\right]} = x.$$

Since $\Psi(\theta)$ *is convex, the above equation is increasing in* θ, *and since* $\Psi'(25) \approx 0.0917$, *we conclude that the rejection constant will be greater than* $\exp(25) \approx 7.2 \times 10^{10}$. *This*

implies that the acceptance probability is of order 10^{-10}*, effectively making this method unusable.*

4.3.1 New Form for the Likelihood Ratio

To avoid the problems mentioned in the previous section, we will now consider RND of the form

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \mathbb{G}(\mu, P, L) \cdot \mathbb{B}(\theta_{P,L}, L_N, P, L) \tag{4.8}$$

where

- $\mathbb{G}(\mu, P, L)$ is of the form

$$\mathbb{G}(\mu, P, L) = \exp\left(\mu^T \Sigma^{-1} \begin{pmatrix} P \\ L \end{pmatrix} - \frac{1}{2}\mu^T \Sigma^{-1} \mu\right) \tag{4.9}$$

  where Σ is the covariance matrix of $(P, L)^T$. Note that under $\mathbb{Q}$, we have $(P,\ L)^T \sim \mathcal{N}(\mu, \Sigma)$, where μ is the IS parameter.

- $\mathbb{B}(\theta_{P,L}, L_N, P, L)$ is of the form

$$\mathbb{B}(\theta_{P,L}, L_N, P, L) = \exp\left(\theta_{P,L} L_N - \Psi_{P,L}(\theta_{P,L})\right) \tag{4.10}$$

  where $\theta_{P,L}$ is a constant in $\mathbb{R}^+$ and $\Psi_{P,L}(\theta_{P,L}) = \log \mathbb{E}[\exp(\theta_{P,L} \cdot L_N) |\, P, L]$. Note that the value of $\theta_{P,L}$ depends on the generated values of $(P,\ L)^T$.

This amounts to tilting the distribution of $(P,\ L)$, and then tilting the conditional distribution of $L_N |\, P, L$.

4.3.2 First Stage

Recall that the pair $(P, L)^T$ has a bivariate normal distribution with mean $(0, 0)^T$ and covariance matrix Σ where $\text{Var}(P) = \text{Var}(L) = 1$ and $\text{Cov}(P, L) = \beta$. To tilt the distribution of (P, L), we shall restrict ourselves to distributions whose RND is of the form (4.9). Applying the Kullback-Leibler criteria as in Proposition 1 in Scott and Metzler (2015), the optimal value μ^* can be expressed as

$$\mu^* = \mathbb{E}\left[\begin{pmatrix} P \\ L \end{pmatrix} \middle| L_N > x\right] \tag{4.11}$$

Recall that $\{L_N > x\}$ is the event whose probability we wish to estimate, and thus a closed-form solution is typically not available for μ^*. Let $\ell(P, L) = \mathbb{E}[L_N|P, L]$. Since $L_N \to \ell(P, L)$ as $N \to \infty$, we suggest replacing L_N by $\ell(P, L)$ in (4.11). This leads to the following approximation:

$$\hat{\mu}^* = \mathbb{E}\left[\begin{pmatrix} P \\ L \end{pmatrix} \middle| \ell(P, L) > x\right] \tag{4.12}$$

The region of integration $\{(p, l) \in \mathbb{R}^2 : \ \ell(p, l) > x\}$ cannot be determined in closed-form and we suggest to approximate (4.12) by first solving the problem

$$\widetilde{\mu} = \underset{(p,l)\in\mathbb{R}^2}{\arg\min} \begin{pmatrix} p & l \end{pmatrix} \Sigma^{-1} \begin{pmatrix} p \\ l \end{pmatrix}, \quad \text{subject to the condition } \ell(p, l) = x.$$

The above problem is quadratic and can be solved easily (note that this problem is analogous to the constant drift approximation outlined in Glasserman and Li (2005)). $\widetilde{\mu}$ is the most likely point on the boundary $\ell(p, l) = x$, and is used to refine the approximation for (4.12) (denoted μ_{ent}) by simulating $(P_i,\ L_i) \sim \mathcal{N}(\widetilde{\mu}, \Sigma)$ and reporting

$$\mu_{ent} = \frac{\sum_{i=1}^{M} \begin{pmatrix} P_i \\ L_i \end{pmatrix} \cdot \mathbf{1}_{\{\ell(P_i,L_i)>x\}} \cdot \Lambda_i^{-1}}{\sum_{j=1}^{M} \mathbf{1}_{\{\ell(P_j,L_j)>x\}} \cdot \Lambda_j^{-1}},$$

where M is the number of simulations defined by the user (in our case, we find that $10,000$ simulations is enough to get a decent estimate for μ_{ent}). The random variable Λ_i^{-1} is given by the expression:

$$\Lambda_i^{-1} = \exp\left(- \begin{pmatrix} P_i & L_i \end{pmatrix} \Sigma^{-1} \widetilde{\mu} + \frac{1}{2} \widetilde{\mu}^T \Sigma^{-1} \widetilde{\mu} \right).$$

Algorithm 7 in Appendix E summarizes the methodology.

Applying a first stage should ensure that many of the generations of $(P,\ L)$ are in the region $\{(P, L) \in \mathbb{R}^2 : \ell(P, L) > x\}$, however, the performance of the estimator can be substantially improved by twisting the conditional distribution of L_N given $(P,\ L)$. Unlike models where the LGD is known, twisting the conditional distribution of the default indicators Y_i is not sufficient to fully twist the conditional distribution of L_N given $(P,\ L)$. However, twisting simultaneously the LGD and PD component of the loss cannot be done directly in general, and in the next section, we discuss how a rejection sampling algorithm can be implemented to improve the estimate obtained in the first stage.

4.3.3 Second Stage

The goal of the second stage is to twist the conditional distribution of L_N when the simulated value of $(P,\ L) \notin \{(P,\ L) \in \mathbb{R}^2 : \mathbb{E}[L_N|P, L] > x\}$. Then, we need to apply a second round of IS in order to improve the estimation of $\mathbb{P}(L_N > x)$. To twist the conditional distribution of the loss, we shall restrict ourselves to densities whose RND is of the form (4.10), that is

$$\mathbb{B}(\theta_{P,L}, L_N, P, L) = \exp\left(\theta_{P,L} L_N - \Psi_{P,L}(\theta_{P,L})\right).$$

Because of the complicated form of L_N, we cannot use the inverse transform to generate from a density whose RND is of the form (4.10), so we have to use rejection sampling instead. Since $L_N \in (0,1)$ we have

$$\mathbb{B}(\theta_{P,L}, L_N, P, L) \leq \exp\left(\theta_{P,L} \cdot \mathbf{1}_{\left\{\theta_{P,L} \in \mathbb{R}^+\right\}} - \Psi_{P,L}\left(\theta_{P,L}\right) \cdot \mathbf{1}_{\left\{\theta_{P,L} \in \mathbb{R}^-\right\}}\right), \tag{4.13}$$

and provided a value for $\theta_{P,L}$, we can easily apply a rejection sampling algorithm using the right-hand side of (4.13) as the rejection constant. Using Kullback-Leibler minimization and Proposition 3 in Scott and Metzler (2015), the optimal value of $\theta_{P,L}$ (denoted $\theta^*_{P,L}$) is the unique solution of the non-linear equation

$$\Psi'_{P,L}\left(\theta^*_{P,L}\right) = \mathbb{E}[L_N |\ L_N > x, P, L]. \tag{4.14}$$

To simplify computations, we suggest to replace the right-hand side of (4.14) by $\max\{\mathbb{E}[L_N|P,L], x\}$ (see Scott and Metzler (2015) for further motivation and justification). This leads to finding the unique value of $\theta^*_{P,L}$ which satisfies

$$\Psi'_{P,L}\left(\theta^*_{P,L}\right) = x, \tag{4.15}$$

if $\ell(P,L) < x$, and set $\theta^*_{P,L} = 0$ otherwise. Note that, because the function $\Psi(\theta)$ is convex, (4.15) can easily be solved numerically and since $\Psi'_{P,L}(0) = \mathbb{E}[L_N|P,L] \leq \mathbb{E}[L_N|L_N > x, P, L] = \Psi'_{P,L}(\theta^*_{P,L})$, we have $\theta^*_{P,L} \in \mathbb{R}^+$. Therefore, the upper bound (4.13) can be refined as

$$\mathbb{B}\left(\theta^*_{P,L}, L_N, P, L\right) \leq \exp\left(\theta^*_{P,L}\right).$$

The algorithm is as follows: we first generate L_N and accept it if

$$\theta^*_{P,L}(1 - L_N) + \Psi_{P,L}\left(\theta^*_{P,L}\right) < E, \tag{4.16}$$

where $E \sim \text{Exp}(1)$. This algorithm allows us to generate exactly from the density whose RND is of the form (4.13), however, the acceptance probability is $\exp(-\theta^*_{P,L})$. In theory, this is not a problem, however by Remark 4.4.1, $\theta^*_{P,L}$ increases linearly with respect to N for the homogeneous model defined in Section 4.4. Thus, as N increases, the acceptance probability decreases exponentially fast.

4.3.4 Alternative for the Second Stage

Let $L_N^{(i)} = LGD_i \cdot Y_i$. This allows us to rewrite L_N as

$$L_N = \frac{1}{N}\sum_{i=1}^{N} L_N^{(i)}.$$

The algorithm defined in Section 4.3.3 is slow because we are applying an AR scheme on the joint distribution of $\mathbf{L} = (L_N^{(1)}, \ldots, L_N^{(N)})$. Thus, if only a few components of the vector $\mathbf{L}$ make the likelihood ratio big, then we will have to reject the entire vector $\mathbf{L}$. This is a waste of computational effort, so instead we shall consider an estimator based on the idea behind *Gibbs' Sampling*.

We shall apply a rejection sampling algorithm on the marginal distribution of $L_N^{(i)}$ given (P, L). Note that given (P, L), $L_N^{(i)}$ is conditionally independent of $L_N^{(j)}$ for $i \neq j$. Therefore, (4.10) can be rewritten as

$$\begin{aligned}
\mathbb{B}(\theta, L_N, P, L) &= \prod_{i=1}^{N} \mathbb{B}(\theta, L_N^{(i)}, P, L) \\
&= \prod_{i=1}^{N} \exp\left(\frac{\theta}{N} \cdot L_N^{(i)} - \Psi^{(i)}_{P,L}\left(\frac{\theta}{N}\right)\right),
\end{aligned}$$

where $\Psi_{P,L}^{(i)}(\theta) = \log \mathbb{E}\left[\exp\left(\theta \cdot L_N^{(i)}\right) \middle| P, L\right]$ and where $\Psi_{P,L}(\theta) = \sum_{i=1}^{N} \Psi_{P,L}^{(i)}(\theta)$. This implies that under the twisted measure, the RND for $L_N^{(i)}$ is

$$\exp\left(\frac{\theta}{N} \cdot L_N^{(i)} - \Psi_{P,L}^{(i)}\left(\frac{\theta}{N}\right)\right).$$

Therefore, we can apply an AR scheme on each $L_N^{(i)}$ independently, which should improve the speed of the algorithm overall. Since we have $L_N^{(i)} \in [0,1]$, the following upper bound holds

$$\mathbb{B}(\theta, L_N^{(i)}, P, L) \leq \exp\left(\frac{\theta}{N} \cdot \mathbf{1}_{\{\theta \in \mathbb{R}^+\}} - \Psi_{P,L}^{(i)}\left(\frac{\theta}{N}\right) \cdot \mathbf{1}_{\{\theta \in \mathbb{R}^-\}}\right), \tag{4.17}$$

with equality obtaining $\theta = 0$. Note that the criteria for selecting the optimal value of θ remains unchanged, thus $\theta_{P,L}^*$ is the unique solution of

$$\Psi'_{P,L}\left(\theta_{P,L}^*\right) = x,$$

if $\ell(P,L) < x$, and set $\theta_{P,L}^* = 0$ otherwise. As discussed earlier, we have $\theta_{P,L}^* \in \mathbb{R}^+$, thus the upper bound (4.17) can be refined as

$$\mathbb{B}(\theta_{P,L}^*, L_N^{(i)}, P, L) \leq \exp\left(\frac{\theta_{P,L}^*}{N}\right).$$

The algorithm is as follows: we first simulate $L_N^{(i)}$ and accept it if

$$\frac{\theta_{P,L}^*}{N}\left(1 - L_N^{(i)}\right) + \Psi_{P,L}^{(i)}\left(\frac{\theta_{P,L}^*}{N}\right) < E_i, \tag{4.18}$$

where $E_i \sim \text{Exp}(1)$. Note that, with this algorithm, accepting $L_N^{(i)}$ is independent of accepting $L_N^{(j)}$. Thus, the new algorithm requires $N \exp(\theta_{P,L}^*/N)$ steps on average. As discussed in Remark 4.4.1, $\theta_{P,L}^*$ is linearly increasing in N, which implies that our improvement on the AR scheme leads to an algorithm that has a linear complexity in

N. This improvement makes the problem *feasible* since the suggested AR scheme in Section 4.3.3 has a exponential complexity (it takes $\exp(\theta^*_{P,L})$ steps on average before accepting a generated value of L_N).

The proposed algorithm in this Section is fast and efficient due to change of measure on the systematic risk factors (P, L) which makes the solution of (4.15) close to 0. The relative speed of the algorithm can be observed in Figure 4.3.

4.4 Numerics

Recall that we would like to estimate

$$\mathbb{P}(L_N > x) = \mathbb{E}[\mathbf{1}_{\{L_N > x\}}],$$

where L_N is defined as

$$L_N = \frac{1}{N} \sum_{i=1}^{N} L_N^{(i)}.$$

Suppose that $PD_i \equiv PD,\ \rho_i \equiv \rho$. We shall refer to this model as the homogeneous model. Using the parameters in Table 4.1, the estimated value of $\mathbb{P}(L_N > 0.1)$ has an order of magnitude of approximately 1×10^{-7}, and therefore, it takes on average 10^7 simulations before observing a default using a standard Monte Carlo algorithm. Given unlimited time budget, this is not an issue, however if we want fast and reliable estimates, we need to incorporate variance reduction techniques to improve the quality of the simulation.

Recall that the approximation of the first stage drift is given by the solution of

Estimator	Mean	Standard Deviation	CV
first-stage	1.71×10^{-7}	3.78×10^{-6}	21.97
two-stage	1.85×10^{-7}	6.25×10^{-7}	3.37

Table 4.1: *Performance of the first-stage estimator and the two-stage estimator when* $PD = 0.029,\ a = 0.63,\ b = 0.975,\ \rho = 0.2,\ x = 0.1,\ \beta = 0.25,\ R_{LGD} = \sqrt{0.0588},\ R_{PD} = 0.25$ *and* 200 *obligors. We used* $1,000,000$ *simulations for each estimator.*

$$\widetilde{\mu} = \underset{(P,L)\in R^2}{\arg\min} \begin{pmatrix} P & L \end{pmatrix} \Sigma^{-1} \begin{pmatrix} P \\ L \end{pmatrix} \quad \text{subject to the condition } \ell(P,L) = x \tag{4.19}$$

where $\ell(P,L)$ is given by (D.1) in Appendix D and where

$$\Sigma = \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}.$$

To refine the approximation obtained by solving (4.19), we solve the problem

$$\mu_{ent} = \frac{\sum_{j=1}^{M} \begin{pmatrix} P_j \\ L_j \end{pmatrix} \cdot \mathbf{1}_{\{\ell(P_j,L_j)>x\}} \cdot \Lambda_j^{-1}}{\sum_{k=1}^{M} \mathbf{1}_{\{\ell(P_k,L_k)>x\}} \cdot \Lambda_k^{-1}},$$

where $(P_j, L_j)^T \sim \mathcal{N}(\widetilde{\mu}, \Sigma)$, M is the number of simulations and where

$$\Lambda_j^{-1} = \exp\left(- \begin{pmatrix} P_j & L_j \end{pmatrix} \Sigma^{-1} \widetilde{\mu} + \frac{1}{2} \widetilde{\mu}^T \Sigma^{-1} \widetilde{\mu} \right).$$

Once the drift μ_{ent} has been determined, we sample $(P,L)^T \sim \mathcal{N}(\mu_{ent}, \Sigma)$, and set $\theta^*_{P,L}$ such that

$$\Psi'\left(\theta^*_{P,L}\right) = x, \tag{4.20}$$

when $\ell(P,L) < x$, and set $\theta^*_{P,L} = 0$ otherwise. Note that this problem can be rewritten in terms of integrals, and the exact expression is given by (D.2) in Appendix D. This problem can be solved rapidly using numerical integration combined with a Newton or Steffenson root-finding algorithm. Once $\theta^*_{P,L}$ has been determined, we simulate

$L_N^{(i)} = B^{-1}(\Phi(l_i), a, b) \cdot \mathbf{1}_{\{p_i < \Phi^{-1}(PD_i)\}}$, where $(p_i,\, l_i)^T$ given (P, L) is a bivariate normal with mean $\eta = (R_{PD}P,\, R_{LGD}L)^T$ and where

$$\text{Corr}(p_i, l_i | P, L) = \rho$$

The simulated value of $L_N^{(i)}$ is accepted if (4.18) is satisfied. Then, we repeat the above procedure N times to generate $L_N^{(1)}, \ldots, L_N^{(N)}$, and we return an unbiased estimate of $\mathbb{P}(L_N > x)$:

$$\mathbf{1}_{\{L_N > x\}} \cdot \exp\left(-\mu_{ent}^T \Sigma^{-1} \begin{pmatrix} P \\ L \end{pmatrix} + \frac{1}{2} \mu_{ent}^T \Sigma^{-1} \mu_{ent} - \theta_{P,L}^* \cdot L_N + \Psi\left(\theta_{P,L}^*\right) \right),$$

where L_N is expressed as

$$L_N = \frac{1}{N} \sum_{i=1}^{N} L_N^{(i)}.$$

Repeating the algorithm n times and averaging the output leads to an accurate estimate for $\mathbb{P}(L_N > x)$.

Remark 4.4.1. *In Appendix D, we show that N and θ only appear in $\Psi'(\theta)$ via θ/N, which implies that $\theta_{P,L}^*$ is a linear function of N.*

4.4.1 Performance

In this section, we compare the performance of the first-stage estimator against our two-stage estimator. We also compare the performance of the crude Monte Carlo estimator, however we note that the crude Monte Carlo estimator fails to provide even one simulation in the region where $\{L_N > x\}$ when $x \geq 0.1$. As shown in Figure 4.1, the first-stage algorithm and the two-stage algorithm are both capable of producing accurate estimates of loss probabilities even when $\mathbb{P}(L_N > x)$ is of order

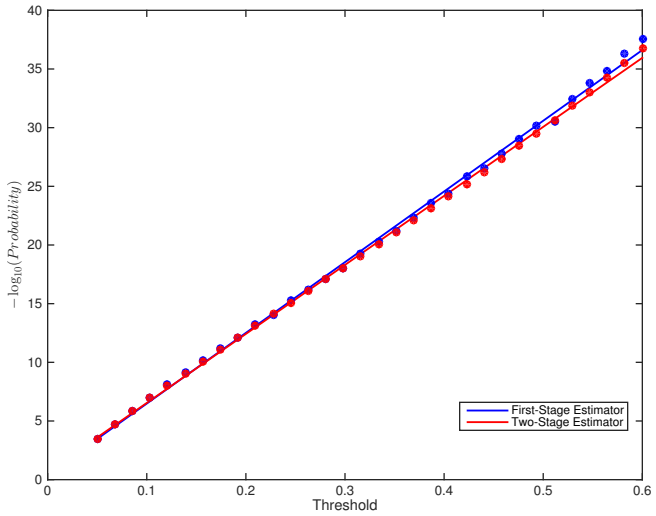

Figure 4.1: *Order of magnitude of the estimated probability* $\mathbb{P}(L_N > x)$ *against the threshold. The parameters are* $a = 0.63$, $b = 0.975$, $PD = 0.029$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$, $\rho = 0.2$, $\beta = 0.25$ *with* 200 *obligors. We used* $1,000,000$ *simulations per point.*

10^{-37}. However, the coefficient of variation of the two-stage estimator is far lower than the first-stage estimator's coefficient of variation, which indicates that applying a second stage significantly improves the performance of the estimator. This can be observed in Figure 4.2. The two-stage estimator performs incredibly well, with a coefficient of variation under 10 even for an estimated probability $\mathbb{P}(L_N > x)$ of approximately 10^{-37}.

We have yet to address the CPU time. Given that the second stage uses a rejection sampling algorithm, we may suspect that it is significantly slower than the first stage estimator. To this end, we compute the number of random variables that need to be generated for each algorithm. For the First-Stage algorithm, we need to generate:

- 2 normal random variables for P and L,
- $2N$ normal random variables for generating L_N

For the second stage, we need (on average):

- 2 normal random variables for P and L,
- $2N \exp(\theta^*_{P,L}/N)$ normal random variables for L_N, and

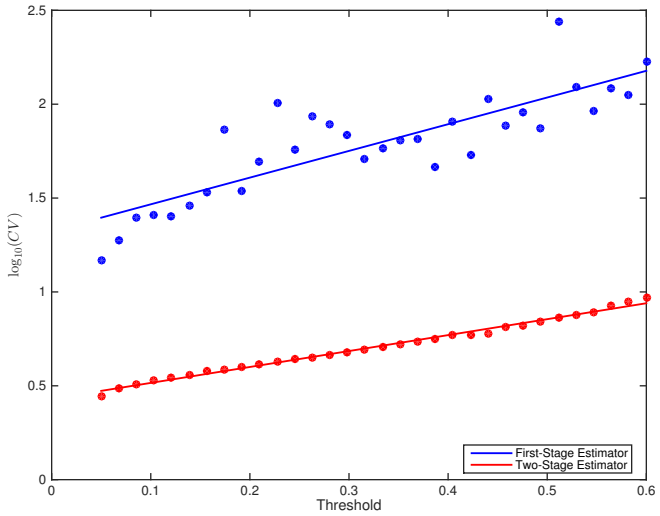

Figure 4.2: *Coefficient of variation of the estimator for* $\mathbb{P}(L_N > x)$ *against the threshold. The parameters are* $a = 0.63$, $b = 0.975$, $PD = 0.029$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$, $\rho = 0.2$, $\beta = 0.25$ *with* 200 *obligors. We used* $1,000,000$ *simulations per point.*

- $N \exp(\theta^*_{P,L}/N)$ exponential random variables.

Therefore, we require generating

$$\frac{2 + 3N \exp\left(\frac{\theta^*_{P,L}}{N}\right)}{2 + 2N} \sim \frac{3}{2} \exp\left(\frac{\theta^*_{P,L}}{N}\right)$$

more random variables. Note that this result is based on the fact that $\theta^*_{P,L}$ is a linear function of N. The ratio can be seen in Figure 4.3. The rejection sampling algorithm is approximately twice as slow as the First-Stage estimator, but its performance is far superior. The relatively low number of iterations required is due to choice of $\theta^*_{P,L}$ bringing many generations in the region $\{(P, L) \in \mathbb{R}^2 : \mathbb{E}[L_N | P, L] > x\}$.

Finally, we can easily adapt the two-stage estimator to estimate the expected shortfall which is defined as follows

$$\mathbb{E}[L_N | L_N > x] = \frac{\mathbb{E}\left[L_N \cdot \mathbf{1}_{\{L_N > x\}}\right]}{\mathbb{E}\left[\mathbf{1}_{\{L_N > x\}}\right]}. \tag{4.21}$$

By applying the two-stage estimator to the right-hand side of (4.21), we get a very

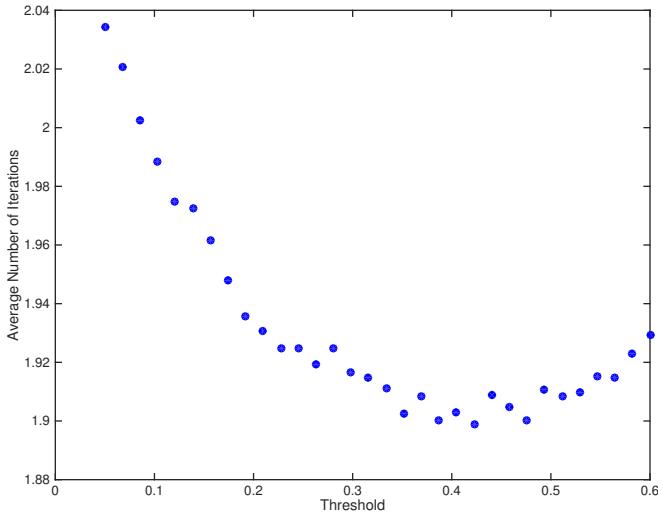

Figure 4.3: *Average iterations required for two-stage entropy algorithm against the threshold. The parameters are* $a = 0.63,\ b = 0.975,\ PD = 0.029,\ R_{LGD} = \sqrt{0.0588},\ R_{PD} = 0.25,\ \rho = 0.2,\ \beta = 0.25$ *with* 200 *obligors. We used* $1,000,000$ *simulations per point.*

accurate estimator for $\mathbb{E}[L_N|L_N > x]$. This can be observed in Figure 4.4. The coefficient of variation for the two-stage estimator is always a few times smaller than the first-stage estimator's coefficient of variation which makes the algorithm far more accurate.

4.5 Concluding Remarks

We have shown how to apply an efficient IS scheme to the model presented in Miu and Ozdemir (2006) using entropy minimization. As opposed to previous work, this model incorporates PD-LGD correlations, and thus, we need to sample from a rejection sampling algorithm in order to achieve efficiency. We compare the added value of incorporating rejection sampling to our estimator and find that it significantly increases the performance of the estimator. Future research will involve extending this approach to a model with multiple groups and applying these IS techniques to other models with PD-LGD correlation.

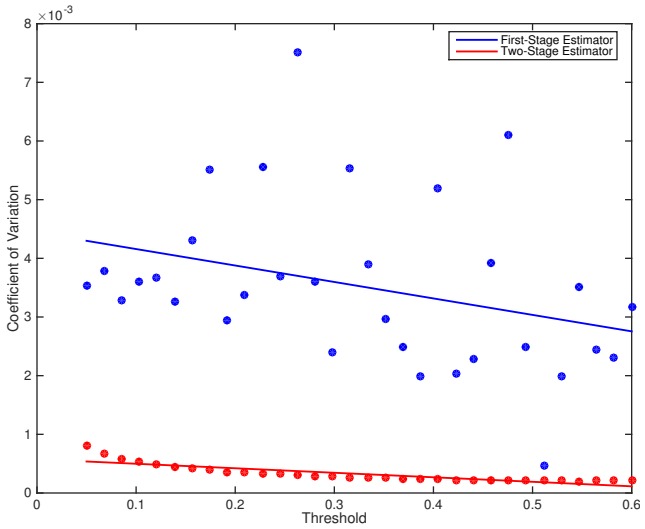

Figure 4.4: *Order of magnitude of the coefficient of variation of the estimation of* $\mathbb{P}(L_N|L_N > x)$ *for the One-Stage entropy criteria and the two-stage entropy criteria. The parameters are* $a = 0.63$, $b = 0.975$, $PD = 0.029$, $R_{LGD} = \sqrt{0.0588}$, $R_{PD} = 0.25$, $\rho = 0.2$, $\beta = 0.25$ *with* 200 *obligors. We used* $100,000$ *simulations per point.*

Bibliography

Achal Bassamboo, Sandeep Juneja, and Assaf Zeevi. Portfolio Credit Risk with Extremal Dependence: Asymptotic Analysis and Efficient Simulation . *Oper. Res.*, 56(3):593–606, July 2006.

Joshua C.C. Chan and Dirk P. Kroese. Efficient Estimation of Large Portfolio Loss Probabilities in t-Copula Models. *European J. of Oper. Res.*, 205(2):361–367, September 2010.

Joshua C.C. Chan and Dirk P. Kroese. Rare-Event Probability Estimation with Conditional Monte-Carlo. *Ann. of Oper. Res.*, 189(1):43–61, September 2011.

Paul Glasserman and Jingyi Li. Importance Sampling for Portfolio Credit Risk. *Management Science*, 51(11):1643–1656, November 2005.

Paul Glasserman, Wanmo Kang, and Perwez Shahabuddin. Fast Simulation of Multifactor Portfolio Credit Risk. *Oper. Res.*, 56(5):1200–1217, 2008.

Sandro Merino and Mark Nyfeler. Calculating portfolio loss. *Risk*, 15(8):82–86, August 2002.

Peter Miu and Bogie Ozdemir. Basel Requirement of Downturn LGD: Modeling and Estimating PD & LGD Correlations. *Journal of Credit Risk*, 2(2):43–68, 2006.

Ad Ridder and Reuven Rubinstein. Minimum cross-entropy methods for rare-event simulation. *Simulation: Transactions of the Society for Modeling and Simulation International*, 83:769–784, 2007.

Reuven Rubinstein. A stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation. *Methodol. Comput. Appl. Probab.*, 7:5–50, 2005.

Reuven Y. Rubinstein and Dirk P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Information Science and Statistics. Springer, 2004.

Alexandre Scott and Adam Metzler. A General Importance Sampling Algorithm for Estimating Portfolio Loss Probabilities in Linear Factor Models. Submitted, February 2015.

Chapter 5

Conclusion

In this thesis, we have shown how the cross-entropy methodology provides both an intuitive and an effective methodology to choose IS measures for sampling. In all cases, the cross-entropy methodology is equivalent to a simple moment-matching criteria which can be applied successively to provide efficient simulations. As seen in Chapter 2, we derive a two-stage IS estimator to first, transform the diffusion process to a process with unit volatility. From this point, we can simulate exactly from the transformed diffusion process. Because simulating from this transformed process is not significantly better than applying the crude Monte Carlo estimator, we apply a second round of IS using cross-entropy. We show that the results are much better for two examples: (i) estimating the probability that a CIR model reaches a user-defined large maximum and (ii) estimating the probability that an OU process reaches a very low minimum and a very large maximum. In both cases, entropy minimization performed incredibly well.

In Chapter 3, we extend the cross-entropy methodology in the context of estimating large losses in a portfolio. We were able to show that the entropy criteria is once again equivalent to a simple moment-matching criteria which could be applied successively in order to apply IS to the systematic risk factors and the idiosyncratic risk factors. Then, we consider an example on the normal copula model and compare our results

to the two-stage IS estimator found in Glasserman and Li (2005). The two models perform similarly, however, cross-entropy minimization can easily be extended to the t copula. We compare the performance of our estimator to Chan and Kroese (2010) and find that our algorithm perform much better when we consider a large number of obligors. The main drawback of their algorithm is that it requires sorting a large number of random vectors, while cross-entropy does not. This allows our algorithm to generate estimates much more rapidly.

In Chapter 4, we extend the cross-entropy methodology to the model defined in Miu and Ozdemir (2006). As opposed to the previous model we have studied, this particular model exhibits correlation between the probability of default and the loss given default. To the best of our knowledge, IS has never been applied to this model although its correlation structure makes it a very interesting model to consider in practice. We show that applying a two stage IS estimator based on cross-entropy is very effective to reduce the variance. For the first stage, we can proceed as in the normal copula scenario, however, the second stage becomes much more complicated. To sample from the IS distribution, we need to apply rejection sampling. By conditioning appropriately, we can make this rejection sampling algorithm efficient (we found that in practice, the algorithm is on average twice as slow as crude Monte Carlo). The algorithm is fast and can be easily implemented, which makes it valuable for practitioners. The variance reduction obtained is also very impressive: the algorithm is capable of producing accurate estimations of probabilities up to 10^{-37}.

Future work will include showing that cross-entropy minimization is asymptotically optimal for the Student-t copula and for the model in Miu and Ozdemir (2006). It would be also interesting to extend the cross-entropy methodology for elliptical copulas and see if we can achieve asymptotically optimality. It would also be interesting to study the connection between cross-entropy and the large deviation criteria.

Bibliography

Joshua C.C. Chan and Dirk P. Kroese. Efficient Estimation of Large Portfolio Loss Probabilities in t-Copula Models. *European J. of Oper. Res.*, 205(2):361–367, September 2010.

Paul Glasserman and Jingyi Li. Importance Sampling for Portfolio Credit Risk. *Management Science*, 51(11):1643–1656, November 2005.

Peter Miu and Bogie Ozdemir. Basel Requirement of Downturn LGD: Modeling and Estimating PD & LGD Correlations. *Journal of Credit Risk*, 2(2):43–68, 2006.

Appendix A

Supplementary Results and Proofs

The following result shows that continuity of the drift function of the Lamperti transformation is sufficient for identity (2.6) to hold. Though this is not necessarily the weakest possible conditions, it is enough for most cases of practical interest.

Theorem A.0.1. *Suppose that X is a solution to the SDE*

$$dX_t = a(X_t)\,dt + dW_t\,, \qquad X_0 = 0\,. \tag{A.1}$$

If a is continuous on (x_-, x_+), where $-\infty \le x_- < 0 < x_+ \le \infty$, then for any path functional $\mathbf{f}$ such that $\mathbb{E}[|\mathbf{f}(X)|] < \infty$ we have

$$\mathbb{E}[\mathbf{f}(X)\mathbf{1}_{\{\underline{X}_T > x_-,\, \overline{X}_T < x_+\}}] = \mathbb{E}[\Lambda^{-1}\mathbf{f}(W)\mathbf{1}_{\{\underline{W}_T > x_-,\, \overline{W}_T < x_+\}}]\,, \tag{A.2}$$

where

$$\Lambda = \exp(-\textstyle\int_0^T a(W_s)\,dW_s + \frac{1}{2}\int_0^T [a(W_s)]^2)\,ds)\,.$$

Proof. Define an increasing sequence of stopping times $\tau_1 < \tau_2 < \ldots$ as follows. If $x_- > -\infty$ and $x_+ < \infty$ then let τ_n be the first exit time of X from $(x_- + n^{-1}, x_+ - n^{-1})$, if $x_- = -\infty$ and $x_+ < \infty$ then let τ_n be the first exit time of X from $(-n, x_+ - n^{-1})$,

etc., and let $X_t^{(n)} := X_{t\wedge\tau_n}$. The process $Z_t^{(n)} = a(X_t^{(n)})$ is bounded, which means that

$$\begin{aligned}\Lambda_t^{(n)} &:= \exp(-M_t^{(n)} - \tfrac{1}{2} < M^{(n)} >_t) \\ &= \exp(-\textstyle\int_0^t a(X_s^{(n)})\, dX_s^{(n)} + \tfrac{1}{2}\int_0^t [a^2(X_s^{(n)})]^2\, ds)\end{aligned}$$

is a martingale, where $M_t^{(n)} := \int_0^t Z_s^{(n)}\, dW_s$ is a martingale and where $< M^{(n)} >_t := \int_0^t [Z_s^{(n)}]^2\, ds$ is its quadratic variation. By Girsanov's Theorem $X^{(n)}$ is a standard Brownian motion under the measure $\mathbb{P}^{(n)}$ defined via the Radon-Nikodym derivative $\frac{d\mathbb{P}^{(n)}}{d\mathbb{P}} := \Lambda_t^{(n)}$. Letting $\mathbb{E}^{(n)}$ denote expectation under this measure we have, for fixed $T > 0$,

$$\begin{aligned}\mathbb{E}[\mathbf{f}(X)\mathbf{1}_{\{\tau_n > T\}}] &= \mathbb{E}[\mathbf{f}(X^{(n)})\mathbf{1}_{\{\tau_n > T\}}] \\ &= \mathbb{E}^{(n)}[(\Lambda^{(n)})^{-1}\mathbf{f}(X^{(n)})\mathbf{1}_{\{\tau_n > T\}}] \\ &= \mathbb{E}[\Lambda^{-1}\mathbf{f}(W)\mathbf{1}_{\{\tau_n^W > T\}}]\,,\end{aligned}$$

where τ_n^W is defined in the obvious manner as the exit time of W from the intervals used to define τ_n. The desired result follows from the dominated convergence theorem.

□

The next theorem ensures that, given an algorithm for generating the terminal value, maximum value and temporal location of the maximum in the case of constant drift, the remaining portions of the zero-drift algorithm presented in Chen and Huang (2012) continue to be valid.

Theorem A.0.2. *Let $X_t = x_0 + \mu t + \sigma W_t$ be a Brownian motion with constant drift. Then for any $T > 0$ the conditional law of $\{X_t : t \in [0, T]\}$, given $(X_T, \overline{X}_T, \xi_T)$, does not depend on μ.*

In order to prove Theorem A.0.2 we may clearly assume without loss of generality that $\sigma = 1$. The strategy is then to show that the finite-dimensional distributions of the indicated conditional law do not depend on the value of μ. To this end the following result, which follows from part (ii) of Theorem 2 stated in Pitman and Yor (1996), is critical.

Lemma A.0.3. *Let $X_t^{x,\mu} = x + \mu t + W_t$ and let $T_z^{x,\mu}$ denote the first hitting time of $X^{x,\mu}$ to the level z. Given $(X_T^{x,\mu} = y, \xi_T^{x,\mu} = u, \overline{X}_T^{x,\mu} = z)$ the path fragments*

$$\{X_s^{x,\mu} : \ s \in [0,u]\} \quad \text{and} \quad \left\{X_{T-s}^{x,\mu} : \ s \in [0,T-u]\right\}$$

are independent, distributed respectively as

$$\{X_s^{x,\mu} : \ s \in [0,T_z^{x,\mu}]\}\,, \text{ given } T_z^{x,\mu} = u$$

and

$$\{X_s^{y,\mu} : \ s \in [0,T_z^{y,\mu}]\}\,, \text{ given } T_z^{y,\mu} = T-u\ .$$

Proof of Theorem A.0.2. The strategy is to show that for any $u \in [0,T]$ and $0 < t_1 < t_2 < \ldots t_n < T$ the quantity

$$\mathbb{P}\left(X_{t_1} \in dx_1, \ldots, X_{t_n} \in dx_n | X_T = y, \xi_T = u, \overline{X} = z\right) \tag{A.3}$$

does not depend on μ. We consider the cases $t_n < u$ and $t_1 > u$ separately, using the strong Markov property to combine them and establish the general case.

In the case that $t_n < u$ we may use Lemma A.0.3 to re-write (A.3) as

$$\mathbb{P}\left(X_{t_1} \in dx_1, \ldots, X_{t_n} \in dx_n | T_z = u\right)\ ,$$

which is equal to

$$\frac{\mathbb{P}\left(T_z \in du | X_{t_1} = dx_1, \ldots, X_{t_n} = dx_n\right) \mathbb{P}\left(X_{t_1} \in dx_1, \ldots, X_{t_n} \in dx_n\right)}{\mathbb{P}(T_z \in du)} \ . \tag{A.4}$$

By the Markov property the term

$$\mathbb{P}\left(T_z \in du | X_{t_1} = dx_1, \ldots, X_{t_n} = dx_n\right) \tag{A.5}$$

factors as the product of

$$\prod_{i=1}^{n} \left[1 - \exp\left(-\frac{2(z - x_{i-1})(z - x_i)}{t_i - t_{i-1}}\right)\right] \ ,$$

which is the probability the path remains below the level z over the interval $[0, t_n]$, given $X_{t_1} = x_1, \ldots, X_{t_n} = x_n$, and

$$\frac{z - x_n}{\sqrt{2\pi(u - t_n)^3}} \exp\left(-\frac{[(z - x_n) - \mu(u - t_n)]^2}{2(u - t_n)}\right) du \ ,$$

which is the (infinitesimal) probability that the process $\{X_{t_n+s} : s \geq 0\}$ strikes the level z for the first time at time $u - t_n$, given $X_{t_n} = x_n$. This quantity clearly factors into the product of

$$\exp\left(\mu(z - x_n) - \frac{\mu^2}{2}(u - t_n)\right) \tag{A.6}$$

and a term that does not involve μ. It is trivial to verify that the quantity

$$\mathbb{P}\left(X_{t_1} \in dx_1, \ldots, X_{t_n} \in dx_n\right)$$

factors as the product of

$$\exp\left(\mu(x_n - x_0) - \frac{\mu^2}{2}t_n\right) \tag{A.7}$$

and a term that does not involve μ, while the term $\mathbb{P}(T_z \in du)$ factors as the product of

$$\exp\left(\mu(z - x_0) - \frac{\mu^2}{2}u\right) \tag{A.8}$$

and a term that does not involve μ. Since (A.8) is equal to the product of (A.6) and (A.7), it follows that (A.4) does not depend on μ.

In the case $t_1 > u$ we may again use Lemma A.0.3 to re-write (A.3) as

$$\mathbb{P}\left(X_{T-t_n} \in dx_n, \ldots, X_{T-t_1} \in dx_1 | T_z = T - u,\, X_0 = y\right) \ ,$$

and it is clear from the preceding discussion that this quantity will not depend on μ.

For the more general case where $t_1 < u < t_n$ define $j = \max\{i :\ 1 \leq i \leq n,\ t_i \leq u\}$ and appeal to Lemma A.0.3 once again to see that (A.3) factors as the product of

$$\mathbb{P}\left(X_{t_1} \in dx_1, \ldots, X_{t_j} \in dx_j | X_T = y, \xi_T = u, \overline{X} = z\right)$$

and

$$\mathbb{P}\left(X_{t_{j+1}} \in dx_{j+1}, \ldots, X_{t_n} \in dx_n | X_T = y, \xi_T = u, \overline{X} = z\right) \ ,$$

neither of which depends on the value of μ. □

Appendix B

Useful Conditional Expectations for Standard Brownian Motion

In this section we discuss the computation of $\mathbb{E}[W_t|\overline{W}_T \geq a]$ and $\mathbb{E}[W_t|\overline{W}_T \geq a,\ \underline{W}_T \leq -b]$ for positive constants a and b.

B.1 Large Maximum

To begin we note that $\mathbb{E}[W_t|\overline{W}_T \geq a] = \mathbb{E}[W_t|T_a \leq T]$, where T_a is the hitting time of W to a. We begin with the special case of $t = T$, in which case the quantity of interest can be computed elegantly via Doob's Optional Sampling Theorem.

Proposition B.1.1. *Let W be a standard Brownian motion beginning at zero and for $a > 0$ let $T_a = \min\{t \geq 0 : W_t = a\}$. Then for any $T > 0$ we have $\mathbb{E}[W_T|T_a \leq T] = a$.*

Proof. Define $\overline{T}_a = \min(T, T_a)$. Then $\overline{T}_a$ is a bounded stopping time and the Optional Stopping Theorem ensures that $\mathbb{E}[W_T|\mathcal{F}_{\overline{T}_a}] = W_{\overline{T}_a} = a \cdot \mathbf{1}_{\{T_a \leq T\}} + W_T \cdot \mathbf{1}_{\{T_a > T\}}$. Therefore $\mathbb{E}[W_T \cdot \mathbf{1}_{\{T_a \leq T\}}|\mathcal{F}_{\overline{T}_a}] = a \cdot \mathbf{1}_{\{T_a \leq T\}}$. Taking expectations we get $\mathbb{E}[W_T \cdot \mathbf{1}_{\{T_a \leq T\}}] = a \cdot \mathbb{P}(T_a \leq T)$, whence $\mathbb{E}[W_T|T_a \leq T] = a$. □

Turning now to the intermediate case $t \in (0, T)$ observe that

$$\begin{aligned}\mathbb{E}[W_t \cdot \mathbf{1}_{\{T_a \le T\}}] &= \mathbb{E}[W_t \cdot \mathbf{1}_{\{T_a \le t\}}] + \mathbb{E}[W_t \cdot \mathbf{1}_{\{t < T_a \le T\}}] \\ &= a \cdot \mathbb{P}(T_a \le T) \\ &\quad + \mathbb{E}[W_t \cdot \mathbb{P}(\overline{W}_t < a | W_t) \cdot \mathbb{P}(\overline{W}_{t,T} \ge a | W_t)] \ ,\end{aligned}$$

where $\overline{W}_{t,T} = \max\{W_s : \ s \in [t, T]\}$ denotes the maximum value of W over the time interval $[t, T]$. Expressions for $\mathbb{P}(\overline{W}_t < a | W_t)$ and $\mathbb{P}(\overline{W}_{t,T} \ge a | W_t)$ are well known, leading to an expression for $\mathbb{E}[W_t | T_a \le T]$ that is easily computed numerically.

B.2 Large Maximum and Small Minimum

We now turn to the problem of computing $\mathbb{E}[W_t | \overline{W}_T > a, \ \underline{W}_T < -b]$. We begin with the observation that

$$\mathbb{E}[W_t | \overline{W}_T > a, \ \underline{W}_T < -b] = \int_{-\infty}^{\infty} w \cdot \mathbb{P}\left(W_t \in dw | \overline{W}_T > a, \ \underline{W}_T < -b\right) \ ,$$

The integrand can be computed using well-known expressions for the quantities

$$H(a, b, y, T) := \mathbb{P}\left(\overline{W}_T < a, \ \underline{W}_T > -b | W_T = y\right) \ , \tag{B.1}$$

$$G(a, b, T) := \mathbb{P}\left(\overline{W}_T < a, \ \underline{W}_T > -b\right) \ . \tag{B.2}$$

Note that

$$H(a, b, y, T) = H(a/\sqrt{T}, b/\sqrt{T}, y/\sqrt{T}, 1)$$

and that

$$G(a,b,T) = \int_{-b}^{a} H(a,b,y,T)\phi(y,T)\,dy = \tfrac{1}{\sqrt{T}} \cdot G(a/\sqrt{T}, b/\sqrt{T}, 1)\ , \tag{B.3}$$

where $\phi(y,T) = (2\pi T)^{-1/2}\exp(-y^2/2T)$.

According to Proposition 8.10 in Karatzas and Shreve (1991) we have, for $y \in (-b,a)$,

$$H(a,b,y,1) = \sum_{n=-\infty}^{\infty} \left[\phi(y+2n(a+b)) - \phi(y-2a-2n(a+b))\right]\ , \tag{B.4}$$

which is easily integrated against $\phi(y,1)$ in order to obtain a series expansion for $G(a,b,1)$.

Now in order to use these expressions in order to compute the quantity of interest begin with observation that

$$\mathbb{P}\left(W_t \in dw | \overline{W}_T > a,\ \underline{W}_T < -b\right) = \frac{\mathbb{P}(W_t \in dw, \overline{W}_T > a, \underline{W}_T < -b)}{\mathbb{P}(\overline{W}_T < a, \underline{W}_T > -b)}.$$

To compute the denominator simply note that

$$\begin{aligned}\mathbb{P}(\overline{W}_T > a,\ \underline{W}_T < -b) = 1 &- \mathbb{P}(\overline{W}_T < a) - \mathbb{P}(\underline{W}_T > -b) \\ &+ \mathbb{P}(\overline{W}_T < a,\ \overline{W}_T > -b)\ ,\end{aligned}$$

and to compute the numerator note that

$$\begin{aligned}\mathbb{P}(W_t \in dw, \overline{W}_T > a, \underline{W}_T < -b) = \mathbb{P}(W_t \in dw) &- \mathbb{P}(W_t \in dw, \overline{W}_T < a) \\ &- \mathbb{P}(W_t \in dw, \underline{W}_T > -b) \\ &+ \mathbb{P}(W_t \in dw, \overline{W}_T < a, \underline{W}_T > -b)\ .\end{aligned}$$

The first three terms of this expression are trivial, to compute the fourth observe that it is equal to

$$\mathbb{P}(\overline{W}_t < a, \underline{W}_t > -b | W_t = w)\mathbb{P}(\overline{W}_{T-t} < a - w, \underline{W}_{T-t} > -b - w)\mathbb{P}(W_t \in dw) \ ,$$

which can be computed using the functions H and G.

Appendix C

Parameters

Parameters used in Table 3.1. Note that the vector E_g is normalized such that $\sum_{g=1}^{G} E_g \cdot N_g = N$.

$$a = \begin{pmatrix} 0.015 & 0.105 & 0.042 & 0.032 & 0.044 & 0.049 & 0.071 & 0.073 & 0.051 & 0.051 \\ 0.067 & 0.085 & 0.123 & 0.093 & 0.052 & 0.107 & 0.017 & 0.008 & 0.011 & 0.021 \\ 0.042 & 0.039 & 0.002 & 0.070 & 0.012 & 0.019 & 0.081 & 0.111 & 0.126 & 0.074 \\ 0.129 & 0.071 & 0.067 & 0.043 & 0.056 & 0.063 & 0.009 & 0.115 & 0.008 & 0.056 \\ 0.107 & 0.051 & 0.079 & 0.106 & 0.114 & 0.120 & 0.025 & 0.033 & 0.116 & 0.077 \\ 0.065 & 0.079 & 0.106 & 0.069 & 0.026 & 0.059 & 0.055 & 0.125 & 0.080 & 0.090 \\ 0.093 & 0.045 & 0.067 & 0.072 & 0.020 & 0.073 & 0.090 & 0.055 & 0.108 & 0.094 \\ 0.046 & 0.059 & 0.050 & 0.100 & 0.095 & 0.056 & 0.090 & 0.122 & 0.101 & 0.091 \\ 0.014 & 0.050 & 0.076 & 0.059 & 0.006 & 0.030 & 0.108 & 0.002 & 0.112 & 0.010 \\ 0.086 & 0.065 & 0.028 & 0.074 & 0.016 & 0.087 & 0.077 & 0.007 & 0.007 & 0.020 \\ 0.003 & 0.056 & 0.107 & 0.080 & 0.067 & 0.112 & 0.013 & 0.117 & 0.014 & 0.067 \\ 0.018 & 0.072 & 0.001 & 0.099 & 0.110 & 0.118 & 0.127 & 0.065 & 0.035 & 0.013 \\ 0.066 & 0.076 & 0.098 & 0.011 & 0.085 & 0.067 & 0.022 & 0.121 & 0.076 & 0.057 \\ 0.122 & 0.085 & 0.058 & 0.108 & 0.069 & 0.072 & 0.088 & 0.047 & 0.031 & 0.075 \\ 0.112 & 0.053 & 0.015 & 0.057 & 0.039 & 0.052 & 0.108 & 0.052 & 0.050 & 0.047 \end{pmatrix}^T$$

$$PD = \begin{pmatrix} 0.004 & 0.031 & 0.012 & 0.024 & 0.039 & 0.017 & 0.028 & 0.030 & 0.017 & 0.026 \end{pmatrix}^T$$

$$E = \begin{pmatrix} 1.522 & 0.478 & 2.457 & 0.316 & 0.924 & 0.746 & 1.668 & 0.287 & 1.273 & 0.329 \end{pmatrix}^T$$

$$w = \begin{pmatrix} 0.152 & 0.048 & 0.246 & 0.032 & 0.092 & 0.075 & 0.167 & 0.029 & 0.127 & 0.033 \end{pmatrix}^T$$

where T is the transpose of a vector.

Appendix D

Equations

The function $\ell(P, L) = \mathbb{E}[L_N|P, L]$ is given by

$$\begin{aligned}\ell(P,L) &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[B^{-1}(\Phi(l_i),a,b)\cdot Y_i|P,L] \\ &= \mathbb{E}[B^{-1}(\Phi(l_i),a,b)\cdot\mathbb{E}[Y_i|P,L,l_i]|P,L] \\ &= \int_{-\infty}^{\infty} B^{-1}(\Phi(z),a,b)\nu(z,P,L)\phi\left(z,R_{LGD}L,\sqrt{1-R_{LGD}^2}\right)dz, \qquad \text{(D.1)}\end{aligned}$$

where $\nu(z, P, L) = \mathbb{E}[Y_i|P, L, l_i]$, which is given via the expression

$$\nu(z,P,L) = \Phi\left(\frac{1}{\sqrt{1-\rho^2}}\left(\frac{\Phi^{-1}(PD_i)-R_{PD}P}{\sqrt{1-R_{PD}^2}} - \frac{\rho(z-R_{LGD}L)}{\sqrt{1-R_{LGD}^2}}\right)\right).$$

The cumulant generating function $\Psi_{P,L}(\theta) = \log\mathbb{E}[\exp(\theta\cdot L_N)|P, L]$ can be expressed as

$$\Psi_{P,L}(\theta) = N\log\left(\int_{-\infty}^{\infty}\left(1+\nu(z,P,L)\left(e^{\frac{\theta}{N}B^{-1}(\Phi(z),a,b)}-1\right)\right)\phi\left(z,R_{LGD}L,\sqrt{1-R_{LGD}^2}\right)dz\right),$$

therefore, the function $\Psi'_{P,L}(\theta)$ can be expressed as

$$\frac{\int_{-\infty}^{\infty} B^{-1}(\Phi(z),a,b)\nu(z,P,L)e^{\frac{\theta}{N}B^{-1}(\Phi(z),a,b)}\phi\left(z,R_{LGD}L,\sqrt{1-R_{LGD}^2}\right)dz}{\int_{-\infty}^{\infty}\left(1+\nu(z,P,L)\left(e^{\frac{\theta}{N}B^{-1}(\Phi(z),a,b)}-1\right)\right)\phi\left(z,R_{LGD}L,\sqrt{1-R_{LGD}^2}\right)dz}. \tag{D.2}$$

The variables θ and N appear on the left-hand side of (D.2) only via θ/N, therefore the optimal solution $\theta^*_{P,L}$ is linear in N.

Appendix E

Algorithms

Algorithm 1 Zero Drift

Simulate the variate W_T.

Simulate $\overline{W}_T$ conditioned on W_T using Corollary 1 in McLeish (2002).

if $\overline{W}_T \geq F(a)$ **then**

    Generate ξ_T conditioned on the pair $(W_T, \overline{W}_T)$ using the algorithm MAXLOCATION in Devroye (2009).

    Simulate $N_T \sim \text{Poisson}(\lambda T)$ and the event times $\tau_1 < \ldots < \tau_{N_T}$.

    Simulate $W_{\tau_1}, W_{\tau_2}, \ldots, W_{\tau_{N_T}}$. This is accomplished as follows. Given $(W_T, \overline{W}_T, \xi_T)$ the fragments $\{\overline{W}_T - W_{\xi_T - t} : 0 \leq t \leq \xi_T\}$ and $\{\overline{W}_T - W_{\xi_T + t} : 0 \leq t \leq T - \xi_T\}$ are two independent Brownian meanders[1] ending at $\overline{W}_T$ and $\overline{W}_T - W_T$ respectively. For more details about Brownian meanders and how to generate skeletons, see Devroye (2009).

    Generate the minimum value of W on each of the subintervals using the algorithm described in Section 4.1 in Chen and Huang (2012).

    if $\underline{W}_T \leq F(b)$ **then**

        Evaluate $\hat{C}_\phi(W, N)$.

        return $\exp(A(W_T)) \cdot \hat{C}_\phi(W, N) \cdot g(\overline{W}_T, \underline{W}_T, W_T)$

Algorithm 1 Zero Drift (continued)

else
  return 0
end if
else
  return 0
end if

Algorithm 2 Constant Drift

Simulate the variate W_T^θ.
Simulate $\overline{W}_T^\theta$ conditioned on W_T^θ using Corollary 1 in McLeish (2002).
Evaluate $\Lambda_\theta^{-1} = \exp\left(-\theta W_T^\theta + \theta^2 T/2\right)$.
if $\overline{W}_T^\theta \geq F(a)$ **then**
  Generate ξ_T^θ conditioned on the pair $(W_T^\theta, \overline{W}_T^\theta)$ using the algorithm MAXLOCATION in Devroye (2009).
  Simulate $N_T \sim \text{Poisson}(\lambda T)$ and the event times $\tau_1 < \ldots < \tau_{N_T}$.
  Simulate $W_{\tau_1}^\theta, W_{\tau_2}^\theta, \ldots, W_{\tau_{N_T}}^\theta$ using Step 6 in Algorithm 1.
  Generate the minimum value of W on each of the subintervals using the algorithm described in Section 4.1 in Chen and Huang (2012).
  if $\underline{W}_T^\theta \leq F(b)$ **then**
    Evaluate $\hat{C}_\phi^\theta(W, N)$.
    return $\exp(A(W_T^\theta)) \cdot \hat{C}_\phi^\theta(W, N) \cdot \Lambda_\theta^{-1} \cdot g(\overline{W}_T^\theta, \underline{W}_T^\theta, W_T)$
  else
    return 0
  end if
else
  return 0
end if

Algorithm 3 Piecewise Constant Drift

Set $W_0 = 0$ and $\Lambda^{-1}_{\theta_{(0)}} = 1$.
for $i = 1$ **to** N **do**
    Generate $W^{\theta}_{s_i}$ conditioned on $W^{\theta}_{s_{i-1}}$.
    Generate $\overline{W}^{\theta}_{(i)}$ conditioned on the pair $(W^{\theta}_{s_{i-1}}, W^{\theta}_{s_i})$ using Corollary 1 in McLeish (2002).
    Evaluate $\Lambda^{-1}_{\theta_{(i)}} = \Lambda^{-1}_{\theta_{(i-1)}} \times \exp\Big(-\theta_i(W^{\theta}_{s_i} - W^{\theta}_{s_{i-1}}) + \theta^2(s_i - s_{i-1})/2\Big)$.
end for
Set $\Lambda^{-1}_{\theta} = \Lambda^{-1}_{\theta_{(N)}}$.
Evaluate $\overline{W}^{\theta}_{T} = \max\{\overline{W}^{\theta}_{(i)},\ i = 1, \ldots, N\}$.
if $\overline{W}^{\theta}_{T} \geq F(a)$ **then**
    for $i = 1$ **to** N **do**
        Generate $\xi^{\theta}_{(i)}$ conditioned on the pair $(W^{\theta}_{s_i}, \overline{W}^{\theta}_{(i)})$ using the algorithm MAXLOCATION in Devroye (2009).
        Simulate $N_{(i)} \sim \text{Poisson}(\lambda(s_i - s_{i-1}))$ and the event times $\tau_1 < \ldots < \tau_{N_{(i)}}$.
        Simulate $W^{\theta}_{s_i+\tau_1}, W^{\theta}_{s_i+\tau_2}, \ldots, W^{\theta}_{s_i+\tau_{N_{T_i}}}$ using Step 6 in Algorithm 1.
        Generate $\underline{W}^{\theta}_{(i)}$ using the algorithm described in Section 4.1 in Chen and Huang (2012).
        Evaluate $\hat{C}^{\theta}_{i,\phi}(W^{\theta}_{(i)}, N_{(i)}) = \prod_{j=1}^{N_{(i)}} \frac{\lambda - \phi(W_{s_i+\tau_j})}{\lambda}$.
    end for
    Evaluate $\underline{W}^{\theta}_{T} = \min\{\underline{W}^{\theta}_{(i)},\ i = 1, \ldots, N\}$.
    if $\underline{W}^{\theta}_{T} \leq F(b)$ **then**
        Set $\hat{C}_{\phi}(W^{\theta}, N) = \prod_{i=1}^{N} \hat{C}^{\theta}_{i,\phi}(W^{\theta}_{(i)}, N_{(i)})$.
        return $\exp(A(W^{\theta}_{T})) \cdot \hat{C}_{\phi}(W^{\theta}, N) \cdot \Lambda^{-1}_{\theta} \cdot g(\overline{W}^{\theta}_{T}, \underline{W}^{\theta}_{T}, W^{\theta}_{T})$
    else
        return 0
    end if
else
    return 0
end if

Algorithm 4 Approximate Entropy Drift ($\mathbb{E}[Z|\ell(Z) > x]$)

Require: G : Number of group, x : Threshold, n : Number of iterations, w_g : Weights for group g.

Initial Estimate:

Set μ to be the solution of

$$\arg\max_{z \in \mathbb{R}^G} z^T z \quad \text{subject to } \ell(z) = x.$$

Refinement:

for $i = 1$ **to** n **do**
  Set $Z_i \sim \mathcal{N}(\mu, I_G)$
  Set $\Lambda_i := \exp\left(-\mu^T Z_i + \frac{|\mu|^2}{2}\right)$
  for $g = 1$ **to** G **do**
    Compute $p_g = \Phi\left(\frac{\Phi^{-1}(PD_g) - a_g^T Z}{1 - |a_g|^2}\right)$
  end for
  Set $\ell_i := \sum_{g=1}^{G} w_g p_g$
  if $\ell_i > x$ **then**
    Set $r_i := Z_i \cdot \Lambda_i$ (where Z_i is a vector)
    Set $p_i := \Lambda_i$
  else
    Set $r_i := 0$
    Set $p_i := 0$
  end if
end for
return $\mu = \frac{\sum_{i=1}^{N} r_i}{\sum_{i=1}^{N} p_i}$

Algorithm 5 Multiple Groups Algorithm to Estimate $\mathbb{P}(L_N > x)$

Require: μ: drift, x : threshold, G : number of groups, a_g : vector correlations for group g, PD_g : probability of default for group g, N_g : number of obligors in group g, w_g : Weights of group g, E_g : Exposure of obligors in group g, n : number of simulations.

Set $E = \sum_{g=1}^{G} E_g \cdot N_g$.

for $i = 1$ **to** n **do**

Set $Z \sim \mathcal{N}(\mu, 1)$

Set $\Lambda := \exp\left(-\mu^T Z + \frac{|\mu|^2}{2}\right)$

for $g = 1$ **to** G **do**

Set $p_g = \Phi\left(\frac{\Phi^{-1}(PD_g) - a_g Z}{\sqrt{1-|a_g^2|}}\right)$

end for

Set $\ell := \sum_{g=1}^{G} w_g \cdot p_g$

if $\ell < x$ **then**

Set θ_z to be the unique solution of the nonlinear problem

$$\sum_{g=1}^{N} w_g \frac{p_g e^{\theta E_g/E}}{1 + p_g\left(e^{\theta E_g/E} - 1\right)} = x.$$

for $g = 1$ **to** G **do**

Set $q_g = \frac{p_g e^{\theta_z E_g/E}}{1+p_g\left(e^{\theta_z E_g/E}-1\right)}$

end for

Set $\Psi(\theta_z) := \sum_{g=1}^{G} \log\left(1 + p_g\left(e^{\theta_z E_g/E} - 1\right)\right)$

else

Set $(q_1, \ldots, q_G) := (p_1, \ldots, p_G)$

Set $\Psi(\theta_z) := 0$

end if

Algorithm 5 Multiple Groups Algorithm to Estimate $\mathbb{P}(L_N > x)$ (continued)

for $g = 1$ **to** G **do**
Set $Y_g \sim \text{Bin}(N_g, q_g)$.
end for
Set $L_N := \sum_{g=1}^{G} \frac{w_g \cdot Y_g}{N_g}$
if $L_N > x$ **then**
Set $\text{simulation}_i := 1$
else
Set $\text{simulation}_i := 0$
end if
Set $\Lambda := \Lambda \cdot \exp(-\theta_z L_N + \Psi(\theta_z))$
Set $\text{simulation}_i := \text{simulation}_i \cdot \Lambda$
end for
return $\sum_{i=1}^{n} \frac{\text{simulation}_i}{n}$.

Algorithm 6 MDT-Multivariate t Copula

Require: ρ : correlation, N : number of assets, PD : probability of default, ν : degrees of freedom, n : number of simulations.

Set α as the unique solution of

$$\frac{1}{2}+\frac{1}{2}\log\left(\frac{\nu+\alpha}{2}\right)-\frac{1}{2}\phi\left(\frac{\nu+\alpha}{2}\right)=\mathbb{E}\left[\log S+\frac{1}{2S^2}\middle|\,\ell(\hat{\mathbf{Z}},S)>x\right]$$

where $S\sim\sqrt{\nu/\chi^2_\nu}$ and where $\phi(\cdot)$ is the digamma function.

Set $K=\log\left(\frac{(\nu/2)^{\nu/2}}{((\nu+\alpha)/2)^{(\nu+\alpha)/2}}\cdot\frac{\Gamma((\nu+\alpha)/2)}{\Gamma(\nu/2)}\right)$.

Set μ to be the solution of

$$\mu=\mathbb{E}[\hat{\mathbf{Z}}\mid\ell(\hat{\mathbf{Z}},S)>x].$$

for $i=1$ **to** n **do**

Simulate $S_i\sim\sqrt{\alpha/\chi^2_\alpha}$.

Simulate $Z_i\sim\mathcal{N}(\mu,1)$.

Compute $p_i:=\Phi\left(\frac{t_\nu^{-1}(PD)\cdot S_i^{-1}-\rho Z_i}{\sqrt{1-\rho^2}}\right)$.

if $p_i<x$ **then**

Set $q_i:=x$.

Set $\theta_i:=N\log\left(\frac{x(1-p_i)}{(1-x)p_i}\right)$

Set $\Psi_i:=N\log\left(\frac{1-p_i}{1-x}\right)$

else

Set $q_i:=p_i$.

Set $\theta_i:=0$.

Set $\Psi_i:=0$.

end if

Compute $L_i:=\mathrm{Bin}(q_i,N)/N$.

$$\Lambda_i:=\exp\left(\alpha\left(\log S_i+\frac{1}{2S_i^2}\right)+K-\mu Z_i+\frac{1}{2}\mu^2-\theta_i L_i+\Psi_i\right)$$

end for

return $\sum_{i=1}^{n}\frac{L_i\cdot\Lambda_i}{n}$

Algorithm 7 Approximate Entropy Drift

Require: x : Threshold, n : Number of iterations, β_1 : correlation between P and L, R_{PD} : driver for $\mathbf{p}$, R_{LGD} : driver for $\mathbf{l}$, ρ_i : correlation between p_i and l_i given (P, L), $a,\ b$: shape parameters.

Initial Estimate:

Compute $\Sigma := \begin{pmatrix} 1 & \beta_1 \\ \beta_1 & 1 \end{pmatrix}$

Set μ^* as the solution of the problem

$$\mu^* = \underset{(P\ \ L)\in\mathbb{R}^2}{\arg\min} \begin{pmatrix} P & L \end{pmatrix} \Sigma^{-1} \begin{pmatrix} P \\ L \end{pmatrix},$$

subject to $\mathbb{E}[L_N|P, L] = x$.

Refinement:

for $j = 1$ **to** n **do**

  Simulate $\begin{pmatrix} P_j & L_j \end{pmatrix}^T \sim \mathcal{N}(\mu^*, \Sigma)$.

  for $i = 1$ **to** N **do**

    Simulate $l_i \sim \mathcal{N}(R_{LGD} \cdot L_j, \sqrt{1 - R^2_{LGD}})$

    Compute $\nu_i := R_{PD} \cdot P_j + \rho_i \sqrt{\frac{1-R^2_{PD}}{1-R^2_{LGD}}}(l_i - R_{LGD} \cdot L_j)$.

    Compute $r_i := \Phi\left(\frac{\Phi^{-1}(PD_i)-\nu_i}{\sqrt{(1-R^2_{PD})(1-\rho_i^2)}}\right)$.

    Set $LGD_i := B^{-1}(\Phi(l_i), a, b)$.

    Simulate $Y_i \sim \text{Ber}(r_i)$.

  end for

  Set $loss_j := \frac{1}{N}\sum_{i=1}^{N} LGD_i \cdot Y_i$.

  Set $\Lambda_j^{-1} := \exp\left(-\begin{pmatrix} P_j & L_j \end{pmatrix} \Sigma^{-1}\mu^* + \frac{1}{2}\mu^{*T}\Sigma^{-1}\mu^*\right)$.

end for

return

$$\left(\sum_{k=1}^{n} \mathbf{1}_{\{loss_k > x\}} \cdot \Lambda_k^{-1}\right)^{-1} \cdot \sum_{j=1}^{n} \begin{pmatrix} P_j \\ L_j \end{pmatrix} \cdot \mathbf{1}_{\{loss_j > x\}} \cdot \Lambda_j^{-1}$$

Algorithm 8 Entropy Minimization Algorithm

Use Algorithm 7 to determine μ_{ent}.
for $j = 1$ **to** n **do**
    Simulate $\begin{pmatrix} P & L \end{pmatrix}^T \sim \mathcal{N}(\mu_{ent}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & \beta_1 \\ \beta_1 & 1 \end{pmatrix}$$

    if $\mathbb{E}[L_N|P, L] < x$ **then**
        Set θ to be the solution of

$$\Psi'(\theta) = x,$$

    where $\Psi(\theta) = \log \mathbb{E}[\exp(\theta L_N)|P, L]$.
    else
        Set $\theta = 0$.
    end if
    Set $\text{loss}_j := 0$.
    for $i = 1$ **to** N **do**
        Set acceptance = **false**.
        while acceptance is **false do**
            Simulate $\begin{pmatrix} p_i & l_i \end{pmatrix}^T \sim \mathcal{N}\left(\begin{pmatrix} R_{PD}P & R_{LGD}L \end{pmatrix}^T, \Gamma\right)$, where

$$\Gamma = \begin{pmatrix} \sigma_{PD}^2 & \rho\sigma_{PD}\sigma_{LGD} \\ \rho\sigma_{PD}\sigma_{LGD} & \sigma_{LGD}^2 \end{pmatrix},$$

    where $\sigma_{PD} = \sqrt{1 - R_{PD}^2}$ and $\sigma_{LGD} = \sqrt{1 - R_{LGD}^2}$.
            Set $LGD_i := B^{-1}(\Phi(l_i), a, b)$, where $B^{-1}(\cdot, a, b)$ is the inverse cumulative distribution function of a beta distribution with shape parameters $a,\ b > 0$.
            Set $Y_i := 1$ if $p_i < \Phi^{-1}(PD)$, otherwise set $Y_i = 0$.
            Set tempLoss $= LGD_i \cdot Y_i$.
            if $\frac{\theta}{N}(1 - \text{tempLoss}) - \frac{\Psi(\theta)}{N} < \text{Exp}(1)$ **then**
                Set acceptance := **true**.
            end if
        end while

Algorithm 8 Entropy Minimization Algorithm (continued)

Set $\text{loss}_j := \text{loss}_j + \text{tempLoss}$.
end for
if $\text{loss}_j > x$ **then**
Set $\Lambda_j^{-1} := \exp\left(-\begin{pmatrix} P & L \end{pmatrix}\Sigma^{-1}\mu_{ent} + \frac{1}{2}\mu_{ent}^T\Sigma^{-1}\mu_{ent} - \theta\cdot\text{loss}_j + \Psi(\theta)\right)$
Set $\text{simulatedLoss}_j := \Lambda_j^{-1}$
else
Set $\text{simulatedLoss}_j := 0$
end if
end for
return $\frac{1}{N}\sum_{j=1}^{N}\text{simulatedLoss}_j$.

Bibliography

Nan Chen and Zhengyu Huang. Brownian Meanders, Importance Sampling and Unbiased Simulation of Diffusion Extremes. *Oper. Res. Lett.*, 40(6):554–563, 2012.

Luc Devroye. *On Exact Simulation Algorithms for Some Distributions Related to Brownian Motion and Brownian Meanders*, 2009.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1991.

Donald L. McLeish. Highs and Lows: Some Properties of the Extremes of a Diffusion and Applications in Finance. *Canad. J. Statist.*, 30(2):243–267, June 2002.

Jim Pitman and Marc Yor. Decomposition at the Maximum for Excursions and Bridges of One-Dimensional Diffusions. *Itô Stochastic Calculus and Probability Theory*, pages 293–310, 1996.

Alexandre Scott's Curriculum Vitae

Post secondary Education and Degrees:

- Université du Québec à Montréal
  Montréal, Québec, Canada
  2006-2009 B.Sc.
- Université du Québec à Montréal
  Montréal, Québec, Canada
  2009-2011 M.Sc.
- The University of Western Ontario
  London, Ontario, Canada
  2011-2015 Ph.D.

Honours and Awards:

- Natural Sciences and Engineering Research Council
  2009-2010, 2011-2015
- Fonds de recherche du Québec: Nature et technologies
  2010-2011, 2011-2012

Related Work Experience:

- Teaching Assistant
  Université du Québec à Montréal
  2008-2011
- Teaching Assistant
  The University of Western Ontario
  2011-2014

Publications:

- Scott, Alexandre and Watier, François. (2011) "Goal achieving probabilities of constrained mean-variance strategies." *Statistics & Probability Letters*. 81: 1021-1026.

- Scott, Alexandre and Watier, François. (2012) "Bounds for Goal Achieving Probabilities of Mean-Variance Strategies with a No Bankruptcy Constraint." *Applied Mathematics*. 3-12.
- Adam Metzler and Alexandre Scott. "Rare Event Simulation for Diffusion Processes via Two-Stage Importance Sampling." *Monte Carlo Methods and Appl.*, 20(2):77-100, March 2014.