
THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

# Computational methods for sums of random variables

Patrick John Laub

BE(Software, Hons. I)/BSc(Mathematics), BSc(Mathematics, Hons. I)

*A thesis submitted for the degree of Doctor of Philosophy in 2018 at The University of Queensland, School of Mathematics and Physics in association with Aarhus University, Department of Mathematical Sciences*

**Abstract**

A typical problem in the field of rare-event estimation is to find the probability $\mathbb{P}(S > \gamma)$ where $S := X_1 + \cdots + X_d$ for a fixed $d \in \mathbb{N}_+$ and where the $\gamma \in \mathbb{R}$ is large or increasing. In applications we often wish to understand the behaviour of a combination of random factors. Hence the random variable $S$ is ubiquitous in real-world modeling problems. It can model, for example, aggregate risk or portfolio value for holding $d$ risky assets [123, 150], the aggregate losses for $d$ insurance policy claims [14, 106], and the combined signal interference from $d$ wireless transmission sources [72]. Probabilities of this form are used to understand how a system would behave under extreme scenarios such as a market crash, a power surge, or a natural disaster. One is typically interested in, not just the quantity $\mathbb{P}(S > \gamma)$, but the behaviour of the summands when the extreme event $\{S > \gamma\}$ occurs.

This probability is available in a closed form for only a few basic cases, when the density of $S$ (which is a $d$-fold convolution) has a known form; see [130]. For example, when the summands are independent and identically distributed (iid), then it is sometimes simple to calculate (for exponential, gamma, normal, binomial, geometric, or negative binomial summands) and sometimes intractable (for lognormal, Weibull, Laplace, or Beta). However, requiring the assumption of independence (let alone iid-ness) of the summands is a stifling restriction when modeling real-world events; a notorious example would be the partial blame of the 2008–09 global financial crisis on mathematicians' inappropriate use of a simplistic dependence model (the Gaussian copula) [151].

This thesis outlines methods for approximating quantities related to sums of random variables. Two chapters consider the use of *orthogonal polynomial expansions* in approximating probability density functions; one focuses on sums of correlated lognormal random variables, and the other on random sums which are used in insurance. We also introduce an *importance sampling estimator* for the survival function of a sum distribution which uses knowledge of the asymptotic form of the sum. We also give the results of an *asymptotic analysis* of the Laplace transform for the sum of lognormal random variables. A

related problem, of estimating the probability of the maximum of a random vector exceeding a large threshold, is also considered.

## Beregningsmetoder for summer af stokastiske variable

*Abstrakt.*

Et typisk problem i forbindelse med estimation af sjældne hændelser er at finde sandsynligheden $\mathbb{P}(S > \gamma)$ hvor $S := X_1 + \cdots + X_d$ for et fast $d \in \mathbb{N}_+$, og hvor $\gamma \in \mathbb{R}$ er stor eller voksende. I anvendelser er man ofte interesseret i at forstå opførslen af en kombination af stokastiske faktorer. Derfor er den stokastiske variabel $S$ naturligt forekommende i praktiske modelleringsproblemer. Den kan f.eks. bruges som til modellering af en samlet risiko, af en porteføljeværdi ved en beholdning af $d$ risikofyldte aktiver [123, 150], af det samlede tab for $d$ forsikringsforpligtelser [14, 106] og af den kombinerede signalinterferens fra $d$ trÃ̆dløse transmissionskilder [72]. Sandsynligheder af denne form bruges til at forstå, hvordan et system vil opføre sig under ekstreme forhold som f.eks. et børskrak, et strømsvigt eller en naturkatastrofe. Typisk er man ikke kun interesseret i størrelsen $\mathbb{P}(S > \gamma)$, men også i opførslen af summanderne når den ekstreme hændelse $\{S > \gamma\}$ indtræffer.

Kun i ganske få tilfælde er denne sandsynlighed tilgængelig i lukket form, når tætheden af $S$ (som svarer til $d$ foldinger) har en kendt form; se [130]. For eksempel når summanderne er uafhængige og identisk fordelte (iid), er tætheden nogle gange nem at beregne (hvis summandernes fordeling er eksponentiel, gamma, normal, binomial, geometrisk eller negativ binomial), og andre gange er det ikke muligt (hvis summandernes fordeling er lognormal, Weibull, Laplace eller Beta). Dog er antagelsen om uafhængighed (og i særdeleshed om iid) mellem summanderne en seriøs begrænsning ved modellering af virkelige hændelser; et notorisk eksempel er den delvise skyld i den globale finanskrise i 2008–2009, som matematikerne bærer, grundet upassende brug af en simpel afhængighedsmodel (den gaussiske copula) [151].

Denne afhandling beskriver metoder til approksimation af størrelser relateret til summer af stokastiske variable. To kapitler omhandler anvendelsen af *ortogonale polynomielle udviklinger* til approksimation af tæthedsfunktioner; det ene fokuserer på summer af korrelerede lognormale stokastiske variable og det andet på stokastiske summer, der anvendes i forsikring. Vi introducerer også en *importance sampling-estimator* for overlevelsesfunktionen af en sumfordeling, som bygger på viden om den asymptotiske form af summen.

Endvidere giver vi resultaterne af en *asymptotisk analyse* af Laplace-transformen af summen af lognormale stokastiske variable. Endeligt behandler vi et relateret problem, som vedrører estimation af sandsynligheden for, at maksimum af en stokastisk vektor overskrider et givet stort niveau.

## Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

**Publications included in this thesis**

Lars Nørvang Andersen, Patrick J. Laub, Leonardo Rojas-Nandayapa (2016), *Efficient simulation for dependent rare events with applications to extremes*, Methodology and Computing in Applied Probability

Søren Asmussen, Pierre-Olivier Goffard, Patrick J. Laub (2015), *Orthonormal polynomial expansions and lognormal sum densities*, Risk and Stochastics: Ragnar Norberg at 70 (Mathematical Finance Economics), World Scientific

Patrick J. Laub, Søren Asmussen, Jens Ledet Jensen, Leonardo Rojas-Nandayapa (2015), *Approximating the Laplace transform of the sum of dependent lognormals*, Advances in Applied Probability.

## Submitted manuscripts included in this thesis

Pierre-Olivier Goffard, Patrick J. Laub (2017), *Two numerical methods to evaluate stop-loss premiums*, Scandinavian Actuarial Journal (submitted)

Thomas Taimre, Patrick J. Laub (2018), *Rare tail approximation using asymptotics and L1 polar coordinates*, Statistics and Computing (submitted)

## Other publications during candidature

### Peer-reviewed papers

Lars Nørvang Andersen, Patrick J. Laub, Leonardo Rojas-Nandayapa (2016), *Efficient simulation for dependent rare events with applications to extremes*, Methodology and Computing in Applied Probability

Søren Asmussen, Enkelejd Hashorva, Patrick J. Laub, Thomas Taimre (2017), *Tail asymptotics of light-tailed Weibull-like sums*, Probability and Mathematical Statistics

Patrick J. Laub, Søren Asmussen, Jens Ledet Jensen, Leonardo Rojas-Nandayapa (2015), *Approximating the Laplace transform of the sum of dependent lognormals*, Advances in Applied Probability.

### Book chapters

Søren Asmussen, Pierre-Olivier Goffard, Patrick J. Laub (2015), *Orthonormal polynomial expansions and lognormal sum densities*, Risk and Stochastics: Ragnar Norberg at 70 (Mathematical Finance Economics), World Scientific

## Contributions by others to the thesis

The chapters in this thesis (except for the first chapter) represent published or submitted work done with co-authors. For these chapters, each co-author contributed to the conception, planning, mathematics, and writing. See the page preceding each chapter, labelled "Authorship Statement", for further details.

**Statement of parts of the thesis submitted to qualify for the award of another degree**

No works submitted towards another degree have been included in this thesis.

**Research Involving Human or Animal Subjects**

No animal or human subjects were involved in this research.

**Acknowledgements**

*"A man must love a thing very much if he practices it without any hope of fame or money, but even practices it without any hope of doing it well."*

G. K. Chesterton

I have been incredibly lucky to have been supported by many institutions, academics, mentors, friends, and family members, without whom this thesis would not exist.

Firstly, I must sincerely thank my supervisors in Aarhus and Brisbane, Søren Asmussen and Phil Pollett. Søren welcomed me into Denmark, gave me a path in applied probability, and calmly guided me along the way. It is due to his meetings, emails, lectures, and books that I have learned so much these past years, and met so many friendly mathematicians all across Europe. Having joined the hiking trip in Greenland which he organised, and seen his pre-departure presentation, inspires me to go and make the most of each day.

Phil Pollett is the reason why I became passionate about probability in the first place. It was only by chance that I enrolled in his probability course STAT3004 in 2012. His well prepared and entertaining lectures (replete with a *schadenfreude*-inducing tyranny aimed at the unpunctual students) inspired me to pursue mathematics research instead of a standard engineering career. As my honours supervisor Phil was unfailingly positive and helpful, always emphasising that the enjoyment of the journey was of paramount importance. He gave me the confidence to begin my PhD. I can say, with probability one, that two better supervisors than Søren and Phil cannot be found!

Thomas Taimre has taught me an incredible amount during his lectures and chats in his office or over Merlo coffee. He dissected draft after draft of my honours thesis, explaining to me a menagerie of minuscule typography, TEX, and grammar rules. I'm very grateful for the help of Leonardo Rojas-Nandayapa, who recommended me to Søren, undertook a vast amount of negotiation and paperwork to setup this joint PhD, and supervised the first half of my PhD.

I must sincerely thank my collaborators: Jens Ledet Jensen, Pierre-Olivier Goffard, Lars Nørvang Andersen, Enkelejd Hashorva, Robert Salomone, Zdravko I. Botev, Jevgenijs Ivanovs, and Hailiang Yang. I hope to work with you again soon!

Being part of the research group ACEMS has been a wonderful experience. They sup-

ported me in a variety of ways, including a stipend, conference travel funding, and attendance at the yearly retreat. In particular, Peter Taylor has been a great mentor for me, and I'm thankful that he hosted me in Melbourne in 2017.

I benefited from many great teachers over the years. Joel Fenwick's memorable and meticulous course on programming taught me so much that I felt like a totally different person after completing it. And before this, I am grateful for my Mackay teachers Janelle Agius, Brendan Gunning, Pauline Hendry, and Glen Smith.

The Australian and Danish governments have been essential in their financial support of my research. I am very thankful to the graduate schools in both universities for allowing this joint degree.

My time in these maths departments has been enriched by the friends I've made there. In Australia, this included Alice, Azam, Leslie, Liam, Marielle, and Rob, and in Denmark, Claudio, Julie, Mikkel, Pierre-Olivier, Thorbjørn, Victor, and Itô.[1] I can't thank you enough for the coffee and *kage* (and the occasional *øl*) breaks which were needed to preserve one's sanity. Also, it was an unlikely pleasure to meet housemates in Nordre Ringgade and Hoogley Street who are such great people.

I'm blessed to have met my partner Vivian, who has been such a tremendous source of joy and energy and peace. We are certainly taking the road less traveled by — let it keep going, as winding as unpredictable as ever! My closest friends, Autumn, Blake, Bryce, Evan, Nina, Tiffany, Tzara, and Will, have helped me in infinitely many ways over the years. Thank-you. Also, I thank my capitalist friends Elliot and Rory for their company.

Finally, and most importantly, I have to thank my family. To Mum and Dad, I'm thankful for everything. From day one (and as Mum would hasten to add, from much earlier than that!) you have worked hard to support every one of my endeavours. You taught me patience, humility, and the value of education. Thanks Chris, Flick, Nanna, and Sam for your long-distance company, care, advice, and Skype calls.

**Patrick J. Laub**
Brisbane, 2018

[1] Note, I am not referring to the esteemed mathematician Kiyoshi Itô — whose work has brought me many a headache — but to Victor & Line's cute dog, who would visit the office and is named after him.

**Financial support**

This research was supported by an Australian Government Research Training Program Scholarship. A top-up scholarship and travel funds were also provided by the Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS), under grant number CE140100049. Further support from Aarhus University was provided by a grant to Søren Asmussen.

**Keywords**: monte carlo, sums, maxima, dependence, rare events, lognormal distribution, stop-loss premium, orthogonal expansions, asymptotic analysis

**Australian and New Zealand Standard Research Classifications (ANZSRC)**

ANZSRC code: 010404 Probability Theory, 60%
ANZSRC code: 010201 Approximation Theory and Asymptotic Methods, 30%
ANZSRC code: 010205 Financial Mathematics, 10%

**Fields of Research Classification**

FoR code: 0104, Statistics, 60%
FoR code: 0102, Applied Mathematics, 40%

# Contents

## Abbreviations and Notation

*Abbreviations*:

| | |
|---|---|
| a.s. | almost surely |
| cdf | Cumulative distribution function $F(x)$ |
| CMC | Crude Monte Carlo |
| iid | Independent and identically distributed |
| MCMC | Markov Chain Monte Carlo |
| pdf | Probability density function $f(x)$ |
| pgf | Probability generating function |
| pmf | Probability mass function $f(n)$ |
| resp. | respectively |
| VaR | Value-at-Risk |
| w.l.o.g. | without loss of generality |
| w.r.t. | with respect to |

*Collections of numbers*:

| | |
|---|---|
| $\mathbb{C}$ | Complex numbers, $\{a + \mathrm{i}b : a, b \in \mathbb{R}\}$ |
| $\mathbb{N}_+$ | Natural numbers, $\{1, 2, \dots\}$ |
| $\mathbb{N}_0$ | Natural numbers including zero, $\{0, 1, 2, \dots\}$ |
| $\mathbb{R}$ | Real numbers |
| $\mathbb{R}_+$ | Positive real numbers, $x > 0$ |
| $\overline{\mathbb{R}}$ | Extended reals, $\mathbb{R} \cup \{-\infty, \infty\}$ |
| $\mathbb{Z}$ | Integers, $\{\dots, -2, -1, 0, 1, 2, \dots\}$ |

*Fonts*:

| | |
|---|---|
| Lowercase letters | Constants, e.g. $a = 5$, $\lambda = 1$ |
| Uppercase Roman letters | Random variables, e.g. $X \sim \mathsf{Normal}(\mu, \sigma^2)$ |
| Boldface lowercase letters | Vectors, e.g. $\boldsymbol{x} = (x_1, x_2, x_3)$, $\boldsymbol{y} = (1, 0, 1)$ |
| Boldface uppercase letters* | Random vectors, e.g. $\boldsymbol{X} = (X_1, X_2, X_3)$ |
| | Matrices, e.g. $\boldsymbol{A}$, $\boldsymbol{H}$, $\boldsymbol{\Sigma}$ |
| Sans serif font | Distributions, e.g. $\mathsf{Gamma}(r, m)$, $\mathsf{Poisson}(\lambda)$ |
| Small caps | Software packages, e.g. MATHEMATICA, MATLAB |
| Teletype font | Functions in software packages, e.g. the `HermiteH` function |

*I have let random vectors take letters near the end of the Roman alphabet, leaving the Greek letters and the remaining Roman letters to denote matrices.

*Other notation*:

| | |
|---|---|
| $\Re(z)$, $\Im(z)$ | Real and imaginary parts of a number, i.e., $z = \Re(z) + \mathrm{i}\Im(z)$ |
| $f^{(n)}(x)$ | The $n$-th derivative of function $f(x)$ |
| $\mathbb{P}(A)$ | Probability of event $A$ |
| $\mathbb{E}[X]$ | Expectation of random variable $X$ |
| $\mathbb{V}\mathrm{ar}[X]$ | Variance of random variable $X$ |
| $\mathbb{I}\{A\}$ | Indicator function for event $A$ |
| $\xrightarrow{\text{a.s.}}$ | Convergence almost surely |
| $\xrightarrow{\mathcal{D}}$ | Convergence in distribution |
| $\overset{\mathcal{D}}{=}$ | Equal in distribution |
| $\sim$ | Distributed as, e.g. $X \sim \mathsf{Normal}(0,1)$, $Y \sim f_Y$ |
| $\overset{\text{iid}}{\sim}$ | Independently and identically distributed as |
| $\overset{\text{ind}}{\sim}$ | Independently distributed as |
| $\dot{\sim}$ | Approximately distributed as |
| $\Phi$ | Standard normal cdf |
| $:=$ | Left-hand side defined as right-hand side |
| $=:$ | Right-hand side defined as left-hand side |
| $\boldsymbol{\Sigma}^\top$ | Transpose of matrix $\boldsymbol{\Sigma}$ |

*Parametrisations of probability distributions:*

Uniform distribution: denoted $\mathsf{Uniform}(a,b)$ where $a, b \in \mathbb{R}$ and $a < b$, has pdf

$$f(x) = \frac{1}{b-a}\,, \quad a < x < b\,.$$

Exponential distribution: denoted $\mathsf{Exponential}(\lambda)$ where $\lambda > 0$, has pdf

$$f(x) = \lambda \mathrm{e}^{-\lambda x}\,, \quad x > 0\,.$$

Gamma distribution: denoted $\mathsf{Gamma}(r, m)$ where $r > 0$ and $m > 0$, has pdf

$$f(x) = \frac{x^{r-1}\mathrm{e}^{-\frac{x}{m}}}{\Gamma(r)m^r}\,, \quad x \in \mathbb{R}_+\,,$$

where $\Gamma$ is the gamma function.

Erlang distribution: denoted $\mathsf{Erlang}(n, m) = \mathsf{Gamma}(n, 1/m)$ where $n \in \mathbb{N}_+$ and $m > 0$.

Normal distribution: denoted $\mathsf{Normal}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, has pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,, \quad x \in \mathbb{R}\,.$$

Lognormal distribution: denoted $\mathsf{Lognormal}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, has pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x}\mathrm{e}^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}\,, \quad x \in \mathbb{R}_+\,.$$

Multivariate normal distribution: denoted $\mathsf{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d,d}$ is positive semi-definite, has pdf

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \mathrm{d}\boldsymbol{x}\,, \quad \boldsymbol{x} \in \mathbb{R}^d\,.$$

Multivariate lognormal distribution: denoted $\mathsf{Lognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d,d}$ is positive semi-definite, has pdf

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}) \prod_{i=1}^d x_i}} \exp\left\{-\frac{1}{2}(\mathbf{log}\,(\boldsymbol{x})-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{log}\,(\boldsymbol{x})-\boldsymbol{\mu})\right\} \mathrm{d}\boldsymbol{x}\,, \quad \boldsymbol{x} \in \mathbb{R}_+^d\,.$$

Dirichlet distribution: denoted $\mathrm{Dirichlet}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}_+^d$, has pdf

$$f(\boldsymbol{\theta}) = \frac{\Gamma\left(\sum_{i=1}^d \alpha_i\right)}{\prod_{i=1}^d \Gamma\left(\alpha_i\right)} \prod_{i=1}^d \theta_i^{\alpha_i - 1}\,, \qquad \boldsymbol{\theta} \in \mathbb{R}_+^d \text{ and } \boldsymbol{\theta}^\top \mathbf{1} = 1\,.$$

Sum of lognormals distribution: denoted $\mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d,d}$ is positive semi-definite. This is the distribution of $S = X_1 + \cdots + X_d$ where $\boldsymbol{X} \sim \mathsf{Lognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It does not have a closed-form density.

Weibull distribution: denoted $\mathsf{Weibull}(\beta, k)$ where $\beta > 0$ and $k > 0$, has pdf

$$f(x) = \beta(x/k)^{\beta-1}\mathrm{e}^{-(x/k)^\beta}, \quad x > 0\,.$$

This is light-tailed if $\beta \geq 1$ and heavy-tailed if $\beta \in (0, 1)$. We sometimes write $\mathsf{Weibull}(\beta)$ to denote $\mathsf{Weibull}(\beta, 1)$.

Pareto distribution: denoted $\mathsf{Pareto}(a, b, \theta)$ where $a, b > 0$ and $\theta \in \mathbb{R}$, has survival function

$$\overline{F}(x) = \Big(1 + \frac{x-\theta}{a}\Big)^{-b}, \quad x > \theta\,.$$

Inverse Gaussian distribution: denoted $\mathsf{InverseGaussian}(\mu, \lambda)$ where $\mu, \lambda > 0$, has pdf

$$f(x) \propto x^{-3/2}\mathrm{e}^{-\lambda(x-\mu)^2/(2\mu^2 x)}\,.$$

Laplace distribution: denoted $\mathsf{Laplace}()$, cf. [68, 109], has pdf

$$f(\boldsymbol{x}) = 2(2\pi)^{-d/2}K_{(d/2)-1}\Big(\sqrt{2\boldsymbol{x}^\top\boldsymbol{x}}\Big)\Big(\sqrt{\tfrac{1}{2}\boldsymbol{x}^\top\boldsymbol{x}}\Big)^{1-(d/2)}, \quad \boldsymbol{x} \in \mathbb{R}^d\,,$$

where $K_n$ denotes the modified Bessel function of the second kind of order $n$.

Poisson distribution: denoted $\mathsf{Poisson}(\lambda)$ where $\lambda \in \mathbb{R}_+$, has pmf

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \in \mathbb{N}_0\,.$$

Binomial distribution: denoted $\mathsf{Binomial}(n, p)$ where $n \in \mathbb{N}_+$, $p \in (0, 1)$, and $p + q = 1$, has pmf

$$f(k) = \binom{n}{k}p^k q^{n-k}, \quad k = 0, 1, \ldots, n\,.$$

Pascal distribution: denoted $\mathsf{Pascal}(\alpha, p)$ where $\alpha \in \mathbb{N}_+$ and $p \in (0,1)$, has pmf

$$f(k) = \binom{\alpha + k - 1}{k} p^\alpha q^k \,, \quad k \in \mathbb{N}_0 \,.$$

If we relax $\alpha \in \mathbb{N}_+$ to $\alpha > 0$ this is the *negative binomial distribution.*

# Chapter 1

# Introduction

*"In theory there is no difference between theory and practice. In practice there is."*

Yogi Berra

## 1.1 A seemingly simple problem

I remember the moment when I was first confronted with the problem of sums of random variables. I was tutoring introductory probability, and one of my students (whose full-time job was in insurance) asked about how to find the distribution of a sum of random variables. My response was that there were simple cases when the distribution of the sum was known (sums of exponentials or gammas were gamma distributed, sums of normals were normal); indeed, these kind of calculations featured prominently in the course's assignments. But, the student asked, what about all the other situations which didn't fit into these niches?

As it turns outs, there is an abundance of solutions to this problem, and none of the answers are as simple as the original question. To be more specific about the problem, let us consider the standard procedure for statistical inference:

1. Collect data on the system of interest

2. Fit a statistical model to the data

3. Use the model to infer interesting quantities

4. (Optional) Perform sensitivity analysis on the inference

Steps 1 and 2 are the application of statistical techniques, and fall outside the scope of this thesis. Step 3 involves calculating probabilities of events occurring (e.g. bankruptcy, catastrophic climate events, physical infrastructure failure, Internet infrastructure failure), and expectations of random variables (e.g. expected profit, expected throughput for a communications hub, expected intensity of earthquakes or cyclones).

To summarise, practitioners are interested in evaluating $\mathbb{P}(S < \gamma)$, or $\mathbb{E}[g(S)]$ for some function $g$, where $S = X_1 + \cdots + X_d$ is the sum of $d$ summands. When $S$ is a *compound sum*, that is, a sum of random variables where the number of summands is also random, then the same quantities need to be estimated but with different techniques. The calculation of $\mathbb{P}(S < \gamma)$ when $\gamma$ is either very large or very small is treated as a special case, called a *rare-event problem*, since it poses unique challenges; rare-event problems are relevant for estimating the probability of Black Swan events (that is, situations which are so rare that we have no historical record of them occurring).

The goal of this thesis is to outline how one would answer real-world questions relating to sums of continuous random variables in an accurate and efficient way. This thesis will use numerical analysis, Monte Carlo methods, asymptotic analysis, and limit theorems.

The research was not undertaken with the goal of being solely dedicated to computational methods, it is simply the case that there tends to be (to a greater or lesser extent) a computational element to each solution. One can learn a great deal about how sums behave by performing asymptotic analysis of the relevant probabilities, expectations, and integral transforms; an exemplary example of this is the *principle of the single big jump*, to be discussed in Section 1.3.6. And since evaluating other types of distributions (like the maximum of a random vector) present similar challenges to sums, the research was not undertaken considering only sums. So, to misquote Voltaire, this thesis is neither (solely) about computational methods nor (solely) about sums of random variables.

Next I will outline some examples, from insurance and finance, where sums or compound

sums are of central importance. Then there is a background section which details the mathematics which the later chapters will rely upon. Finally, I will end this introduction with an outline of existing methods in the literature, and how this thesis contributes to the field.

## 1.2 Applications of sums of random variables

Sums of random variables are fundamental to modeling stochastic phenomena. In finance, risk managers need to predict the distribution of a portfolio's future value which is the sum of multiple assets; similarly, the distribution of the sum of an individual asset's returns over time is needed for valuation of some exotic (e.g. Asian) options [123, 150]. In insurance, the probability of ruin (i.e. bankruptcy) is determined by the distribution of aggregate losses (sums of individual claims of random size) [106, 14]. Lastly, wireless system engineers model total interference in a wireless communications network as the sum of all interfering signals (often lognormally distributed) [72].

**Example 1.1** (Asian options)**.** A standard put option has a payoff of $(X_T - a)_+$, where $X_T$ is the value of a stock $X_t$ at the time of maturity $T$ and $a$ is a predetermined threshold. Asian options differ in that their payoff is $(\overline{X} - a)_+$ where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_{t_i}\,.$$

Calculating the fair price of these options requires the evaluation of $\mathbb{E}[(\overline{X} - a)_+]$ which is of the form $\mathbb{E}[g(\sum_{i=1}^n X_i)]$. Glasserman [79, p. 99] writes of these, "there are no exact formulas for the prices of these options, largely because the distribution of $\overline{X}$ is intractable." ◇

**Example 1.2** (Barrier options)**.** Another so-called exotic option is a barrier option, where the payoff depends on a stockprice's trajectory over a period $t \in [0, T]$. For example, the *down-and-out* option [52] payoff is $\mathbb{I}\{\tau > a\}(X_T - b\}$ where $\tau = \min_{t\in[0,T]} X_t$. This is not an example of sums of random variables, but instead relates to the maxima of random variables, so the research in Chapter 6 is applicable. ◇

**Example 1.3** (Modern portfolio theory)**.** Modern portfolio theory, as first described by Markowitz [122], considers which assets to purchase so that a portfolio has optimally high returns with as little risk as possible. The value of a portfolio is a weighted sum

$$P = \sum_{i=1}^{n} x_i X_i$$

where $x_i$ describes the number of units of asset $i$ purchased, and $X_i$ is the random value of the asset at a fixed future time. The original model considers $\boldsymbol{X} \sim \mathsf{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and tries to find

$$\arg\max_{\boldsymbol{x}} \boldsymbol{x}^\top \boldsymbol{\mu} - \frac{\gamma}{2} \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \,.$$

The criticisms of this approach fall into two main categories: i) that the normal distribution is an unrealistic model of asset prices, or ii) that the portfolio variance is not an ideal measure of an investor's risk (e.g. when the distribution of $P$ is highly skewed). Neither assumption can be substantially relaxed without dealing with the fact that the distribution of the sum $P$ is no longer tractable. ◇

**Example 1.4** (Loss distribution approach)**.** A tool used by actuaries to control many types of risk is the *loss distribution approach (LDA)* [76]. The idea is to create a probabilistic model of the risk, use the model to calculate a *risk measure*, then use this measure to help manage the risk. An example is *operational risk* (which includes losses due to, e.g., theft and fraud), where the total loss related to this risk during a year is modeled as

$$L = \sum_{i=1}^{N} X_i$$

where $N$ is the random number of events (cases of fraud, et cetera), and the $X_i$ are the associated monetary losses for each event. The LDA specifies that practitioners fit a model for $N$ and for the $X_i$, such as $N \sim \mathsf{Poisson}(\lambda)$ and $X_i \overset{\text{iid}}{\sim} \mathsf{Lognormal}(\mu, \sigma)$, then calculate a risk measure such as *Value-at-Risk (VaR)*. The VaR at level $\alpha \in (0, 1)$ is defined such that the probability of losses exceeding the level VaR is at most $1 - \alpha$. We denote this $\alpha$-quantile as

$$\mathrm{VaR}_\alpha = \inf\{x \geq 0, F_L(x) \geq \alpha\}.$$

Depending on which regulations are applicable, the value of $\alpha$ is taken to be in the range

of 0.995 to 0.9997 see [101, 75].

The LDA is not limited to modelling operational risk, and indeed elements of the approach pervade nearly all domains of risk management [123]. It has received intense scrutiny amongst mathematicians, and one result is that the VaR has been found to behave incorrectly when various risks need to be aggregated (to be precise, the measure is not a *coherent* risk measure). There are many alternative risk measures to choose from, including *expected shortfall*

$$\mathrm{ES}_\alpha = \mathbb{E}[L \mid L > \mathrm{VaR}_\alpha]\,,$$

cf. [123] or [107]. ◇

**Example 1.5** (Premiums in non-life insurance)**.** In non-life insurance, the insurer must set premiums high enough to ensure that the insurer's reserves are not depleted by too many claims. The classical model for the level of an insurer's reserves is

$$R(t) = u + ct - \sum_{i=1}^{N(t)} U_i.$$

where $u$ is the initial reserves, premiums are collected continuously at rate $c$, $N(t)$ is the number of claims submitted before time $t$ and the $U_i$ are the claim severities. The model is called the Cramér–Lundberg model, after the original Swedish pioneers Lundberg [120] and Cramér [55].

Typically $R(t)$ is used to calculate an insurer's ruin probability, that is, the probability that the financial reserves eventually fall below zero. Of interest are both the finite-time ruin probability $\psi(u, T)$ and the infinite-time ruin probability, also called the *probability of ultimate ruin*, $\psi(u)$, which are defined as

$$\psi(u, T) = \mathbb{P}\Big(\inf_{0 \le t \le T} R(t) \le 0\Big)\,,$$

and

$$\psi(u) = \mathbb{P}\Big(\inf_{t \ge 0} R(t) \le 0\Big)\,.$$

The book by Asmussen and Albrecher [14] considers the estimation of these probabilities under various models for the claim arrival process $N(t)$ and for the claim severity distribution. ◇

**Example 1.6** (Wireless systems analysis)**.** Wireless engineers measure the quality of one user's wireless connection by

$$\mathbb{P}(\text{SINR} < \alpha)$$

where $\alpha \in \mathbb{R}$ is a fixed threshold, and SINR is the *signal to interference plus noise ratio*. If the signal's power is denoted $X_0$ and there are $N$ interfering signals, each with power $X_1, \ldots, X_n$, and background noise $N_0$, then

$$\text{SINR} = \frac{X_0}{X_1 + \cdots + X_n + N_0}\,.$$

Fischione [72] and related papers take these random variables to be lognormally distributed, so that the denominator in SINR's definition is a sum of (correlated) lognormal random variables. ◇

## 1.3 Foundational background

### 1.3.1 Quadrature techniques

Every quantity that a probabilist finds interesting — probabilities, expectations, variances, et cetera — is simply an integral. *Quadrature methods* allow us to solve numerically complex integral problems. This section draws on Gautschi's textbook [78] and Hegland's lecture notes [97].

For most integrals we break up the range of integration, such as

$$\int_a^b f(x)\,\mathrm{d}x = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x + \int_{x_2}^{x_3} f(x)\,\mathrm{d}x + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,\mathrm{d}x$$

for $a = x_1 < x_2 < \cdots < x_n = b$, and evaluate the smaller integrals separately. For now, we consider a grid of constant step-size $h > 0$, so $x_i = x_1 + (i-1) \times h$.

The simplest quadrature technique approximates the integrand $f(x)$ as a constant value over each subinterval, and integrates the result. For example, the *midpoint Riemann sum*

*approximation*, illustrated in Figure 1.1, is

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n-1} f\left(\frac{x_i + x_{i+1}}{2}\right) h\,.$$

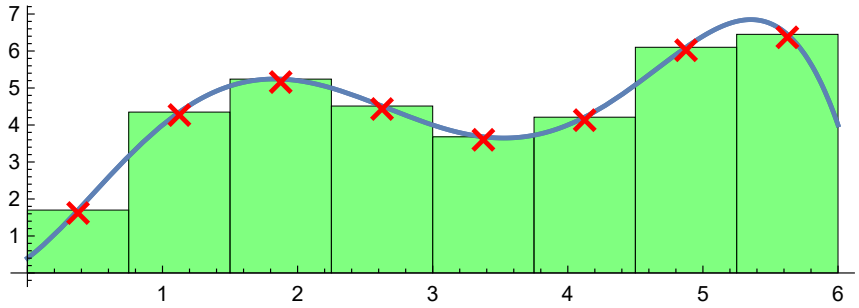



Figure 1.1: A midpoint Riemann sum approximation.

If we replace the piece-wise constant approximation by a piece-wise linear approximation we get a *trapezoidal rule approximation*. This approximation, illustrated in Figure 1.2, is

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n-1} \frac{f(x_i) + f(x_{i+1})}{2} h\,.$$

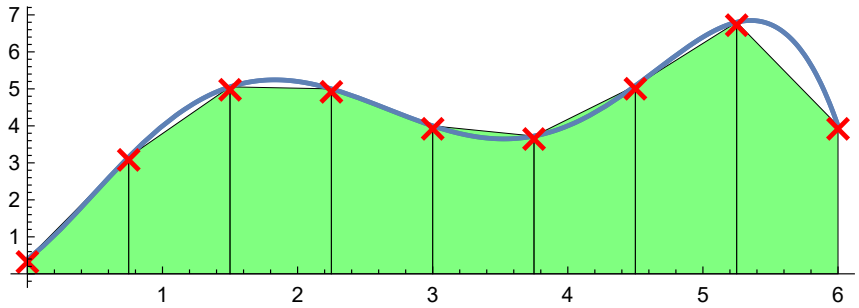



Figure 1.2: A trapezoidal rule approximation.

This integration scheme may seem to be too simple to ever be used, but it will appear in Section 1.3.2 on Laplace transform inversion. Trapezoidal integration has the property that it performs well on highly oscillatory integrands, such as integrals involving the

complex exponential function.

The process can continue, with the approximations becoming higher order polynomials which interpolate the integrand. For example, if we split the range of integration as

$$\int_a^b f(x)\,\mathrm{d}x = \int_{x_1}^{x_3} f(x)\,\mathrm{d}x + \int_{x_3}^{x_5} f(x)\,\mathrm{d}x + \cdots + \int_{x_{n-2}}^{x_n} f(x)\,\mathrm{d}x\,,$$

then we can approximately integrate $\int_{x_1}^{x_3} f(x)\,\mathrm{d}x$ by fitting a quadratic to the triplet

$$\Big\{\big(x_1, f(x_1)\big), \big(x_2, f(x_2)\big), \big(x_3, f(x_3)\big)\Big\}$$

and integrating the resulting fit. This is called a *Simpson's rule approximation*, and is illustrated in Figure 1.3.

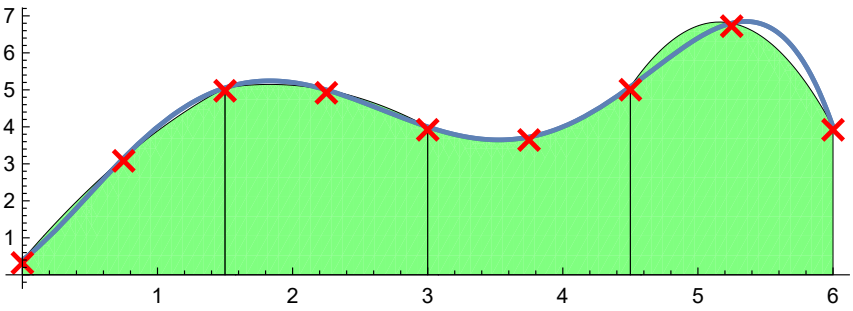



Figure 1.3: A Simpson rule approximation.

The *Newton–Cotes method* is the general case of this polynomial approximation. Before elaborating, let us rewrite the problem to be the evaluation of $\int_a^b f(x)w(x)\,\mathrm{d}x$ where $w(x)$ is called a *weight function*. The Newton–Cotes method approximates this integral as

$$\int_a^b f(x)w(x)\,\mathrm{d}x \approx \int_a^b p_{n-1}(f; x_1, \ldots, x_n; x)w(x)\,\mathrm{d}x \tag{1.1}$$

where $p_{n-1}(f; x_1, \ldots, x_n; x)$ is the unique polynomial of degree $n-1$ which interpolates $f$ on the points $\{x_1, \ldots, x_n\}$ evaluated at the point $x$. This interpolating polynomial, written in *Lagrangian form*, is

$$p_{n-1}(f; x_1, \ldots, x_n; x) = \sum_{k=1}^{n} f(x_k)\ell_k(x)\,, \quad \text{where} \quad \ell_k(x) = \prod_{i=1, i\neq k}^{n} \frac{x - x_i}{x_k - x_i}\,.$$

The approximating integral can be written as

$$\int_a^b p_{n-1}(f; x_1, \ldots, x_n; x) w(x)\,\mathrm{d}x = \sum_{k=1}^{n} w_k f(x_k)\,,$$

where

$$w_k = \int_a^b \ell_k(x) w(x)\,\mathrm{d}x \qquad \text{for } k = 1, \ldots, n\,. \tag{1.2}$$

For many common choices of weight function $w(x)$ (e.g. $w(x) = \mathrm{e}^{-x^2}$) and evaluation points $\{x_1, \ldots, x_n\}$ (e.g. the constant step-size grid over $[-1, 1]$, $x_i = -1 + 2\frac{i-1}{n-1}$) these constants $w_k$ are available in the literature or in software packages.

For an arbitrary choice of $\{x_1, \ldots, x_n\}$, the Newton–Cotes method has a surprising guarantee, which is that it will be exact when $f$ is a polynomial of degree $n-1$ or less. This is because the polynomial approximation of $f$ is just the original function $f$, i.e. $p_{n-1}(f; x_1, \ldots, x_n; x) = f(x)$, so the $\approx$ becomes $=$ in (1.1). The name for this property is that the Newton–Cotes method with $n$ points has *degree of exactness* $n-1$, and the natural next question is, can we achieve a greater degree of exactness than $n-1$ with just $n$ points?

We cannot achieve this. Specifically, we cannot do so by letting $\{x_1, \ldots, x_n\}$ be an arbitrary choice. If we consider the form of the approximation

$$\int_a^b f(x) w(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i)\,, \tag{1.3}$$

we can see that there are $2n$ free variables (the $x_i$ and the $w_i$) and only by choosing all $2n$ correctly we can achieve a degree of exactness of $2n-1$. The resulting approximation is called a *Gaussian quadrature*, and its construction is based on the following theorem. First, we define the *node polynomial* to be

$$\omega_n(x) = \prod_{k=1}^{n} (x - x_k)\,. \tag{1.4}$$

**Theorem 1.7.** *The approximation* (1.3) *will have degree of exactness of* $2n-1$ *iff:*

1. *the* $w_k$ *are given by* (1.2),

*2. and the node polynomial $\omega_n(x)$ satisfies, for every polynomial $q$ of order $\leq n-1$,*

$$\int_a^b q(x)\omega_n(x)w(x)\,\mathrm{d}x = 0\,. \tag{1.5}$$

*Proof.* We only prove the $\Leftarrow$ direction. For the full proof (which the following is based on), see Theorem 3.2.1. of [78].

We wish to show that, for $f$ which is a polynomial of degree $2n-1$ or less,

$$\int_a^b f(x)w(x)\,\mathrm{d}x = \sum_{i=1}^{n} w_i f(x_i)\,.$$

Divide $f$ by $\omega_n$, so

$$f(x) = q(x)\omega_n(x) + r(x) \tag{1.6}$$

where the quotient $q$ and the remainder $r$ are polynomials of order $\leq n-1$, so

$$\int_a^b f(x)w(x)\,\mathrm{d}x = \int_a^b q(x)\omega_n(x)w(x)\,\mathrm{d}x + \int_a^b r(x)w(x)\,\mathrm{d}x\,.$$

The first integral on the right-hand side is zero by (1.5), and since $r$ is of order $\leq n-1$, we know its $n$-point Newton–Cotes approximation is exact, i.e.

$$\int_a^b r(x)w(x)\,\mathrm{d}x = \sum_{i=1}^{n} w_i r(x_i)\,,$$

so rearranging (1.6) gives

$$\int_a^b f(x)w(x)\,\mathrm{d}x = \sum_{i=1}^{n} w_i r(x_i) = \sum_{i=1}^{n} w_i[f(x_i) - q(x_k)\omega_n(x_k)] = \sum_{i=1}^{n} w_i f(x_i)\,.$$

□

Finding the precise $\{x_1, \ldots, x_n\}$ which satisfy the requirements of Theorem 1.7 is easier than it seems. Equation (1.4) shows us that we just need to set $\{x_1, \ldots, x_n\}$ to be the zeros of the node polynomial, and the second condition (1.5) tells us that the node polynomial $\omega_n(x)$ is the $n$-th orthogonal polynomial with respect to the weight function $w(x)$ over $[a, b]$.

So, the procedure for Gaussian quadrature is to find the $n$-th orthogonal polynomial for the given weight function $w(x)$ and range of integration $[a, b]$ (either by looking in the literature or by Gram–Schmidt orthogonalisation), set $\{x_1, \ldots, x_n\}$ to be zeros of this polynomial, and set the $w_k$ by (1.2).

There is a large literature on quadrature (cf. [78] and references), however these algorithms are subject to the ominously-named *curse of dimensionality*. This phrase refers to the fact that, when trying to achieve some fixed accuracy using the approximation

$$\mathbb{E}[g(\boldsymbol{X})] \approx \sum_{i=1}^{n} w_i g(\boldsymbol{x}_i) f_{\boldsymbol{X}}(\boldsymbol{x}_i)\,,$$

the number of evaluation points $n$ required increases exponentially in the dimension $d$. Thus, the rule of thumb is to use quadrature with caution for $d$ in the range of about 1–5, and consider other approaches for larger $d$ (unless the integrand is particularly smooth).

### 1.3.2 Laplace transform inversion

To discuss Laplace transform inversion, we must first define the Laplace transform.

**Definition 1.8.** *For a function $f : \mathbb{R}_+ \to \mathbb{R}_+$, we define*

$$\mathscr{L}\{f\}(t) := \int_0^\infty \mathrm{e}^{-tx} f(x)\,\mathrm{d}x\,, \quad \textit{for } t \in \mathbb{C} \textit{ with } \Re(t) \geq 0\,,$$

*to be the corresponding* Laplace transform. *For a positive random variable $X$ with probability density function (pdf) $f_X$, we write $\mathscr{L}_X(t) := \mathscr{L}\{f_X\}(t) = \mathbb{E}[\mathrm{e}^{-tX}]$.* ⋄

Some useful relations for Laplace transforms include, for $t > 0$

$$\mathscr{L}\{F_X\}(t) = \frac{\mathscr{L}\{f_X\}(t)}{t} = \frac{\mathscr{L}_X(t)}{t}\,, \text{ and}$$

$$\mathscr{L}\{\overline{F}_X\}(t) = \frac{1}{t} - \mathscr{L}\{F_X(x)\}(t) = \frac{1 - \mathscr{L}_X(t)}{t}\,.$$

A function $f$ can be recovered from its Laplace transform by a standard Bromwich integral. We assume $f : \mathbb{R}_+ \to \mathbb{R}_+$, is a measurable function with locally bounded variation. To

define the Bromwich integral, first select a $\gamma > 0$ then

$$f(x) = \frac{2\mathrm{e}^{\gamma x}}{\pi} \int_0^\infty \cos(xs) \Re\left[\mathscr{L}\{f\}(\gamma + \mathrm{i}s)\right] \mathrm{d}s.$$

We apply a basic quadrature rule to the Bromwich integral by first *discretizing* the integral and then *truncating* the resulting infinite sum. In both steps, we follow the steps of Abate and Whitt [2].

### Discretisation

We will use a semi-infinite trapezoidal rule, despite the apparent simplicity of the method. With a grid size $h > 0$, this discretisation yields

$$f(x) \approx f_{\mathrm{disc}}(x) := \frac{2\mathrm{e}^{\gamma x}}{\pi} \cdot h \left\{ \tfrac{1}{2} \mathscr{L}\{f\}(\gamma) + \sum_{j=1}^{\infty} \cos(x \cdot hj) \Re\left[\mathscr{L}\{f\}(\gamma + \mathrm{i}hj)\right] \right\},$$

since $\Re\left[\mathscr{L}\{f\}(\gamma)\right] = \mathscr{L}\{f\}(\gamma)$. We simplify this by choosing $h = \pi/(2x)$ and $\gamma = a/(2x)$ for an $a > 0$, achieving

$$f_{\mathrm{disc}}(x) = \frac{\mathrm{e}^{a/2}}{2x} \mathscr{L}\{f\}\left(\frac{a}{2x}\right) + \frac{\mathrm{e}^{a/2}}{x} \sum_{k=1}^{\infty} (-1)^k \Re\left[\mathscr{L}\{f\}\left(\frac{a + 2\pi\mathrm{i}k}{2x}\right)\right]. \tag{1.7}$$

From Theorem 5.5.1 of [147] we have that the *discretisation error* (also called *sampling error*) is simply

$$f_{\mathrm{disc}}(x) - f(x) = \sum_{k=1}^{\infty} \mathrm{e}^{-ak} f[(2k+1)x]. \tag{1.8}$$

In particular, if $0 \leq f(x) \leq 1$, then

$$f_{\mathrm{disc}}(x) - f(x) \leq \frac{\mathrm{e}^{-a}}{1 - \mathrm{e}^{-a}}. \tag{1.9}$$

There are no absolute value signs here — the discretisation introduces a systematic overestimate of the true value. Also, (1.8) implies $a$ should be as large as possible (limited eventually by finite-precision computation). The benefit of knowing this result is slightly offset by the requirement that $h$ and $\gamma$ now be functions of $x$ rather than constants.

**Truncation**

Due to the infinite series, the expression in (1.7) cannot be directly computed, thus it has to be truncated. The arbitrary-seeming choice of $h$ and $\gamma$ in Section 1.3.2 not only allows for calculation of the discretisation error, but also benefits the truncation step. This is because the sum in (1.7) is (nearly) of alternating sign, and thus *Euler series acceleration* can be applied to decrease the truncation error. Define for $\ell = 1, 2, \ldots$

$$s_\ell(x) := \frac{\mathrm{e}^{a/2}}{2x}\mathscr{L}\{f\}\left(\frac{a}{2x}\right) + \frac{\mathrm{e}^{a/2}}{x}\sum_{k=1}^{\ell}(-1)^k \Re\left[\mathscr{L}\{f\}\left(\frac{a+2\pi\mathrm{i}k}{2x}\right)\right].$$

Then, for some positive integers $M_1$ and $M_2$,

$$f(x) \approx f_{\mathrm{disc}}(x) \approx f_{\mathrm{approx}}(x) := \sum_{k=0}^{M_1}\binom{M_1}{k}2^{-M_1}s_{M_2+k}(x)\,. \tag{1.10}$$

### 1.3.3 Orthogonal polynomials

We begin by a discussion of orthogonality of functions and generalised Fourier series expansions. This exposition draws on [158] and [78]. Firstly, we will define some vector space notation for a space over $\mathbb{R}$ with respect to a *weight function* $w : \mathbb{R} \to \mathbb{R}_+$. For the purposes of this thesis, we have that $w$ is a pdf, i.e. $\int_{\mathbb{R}} w(x)\,\mathrm{d}x = 1$.

**Definition 1.9.** *We define the weighted inner product of functions $f$ and $g$ to be*

$$\langle f, g\rangle_w := \int_{\mathbb{R}} f(x)g(x)w(x)\,\mathrm{d}x$$

*and say that a function's weighted 2-norm is*

$$\|f\|_w := \sqrt{\langle f, f\rangle_w}\,.$$

*This defines a vector space we denote $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$, where $f \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ means that $\|f\|_w < \infty$.*

**Definition 1.10.** *We say the functions $f_1, \ldots f_n \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ are* linearly independent *if an affine combination of the functions is zero only if all coefficients are zero,*

*i.e.*

$$\sum_{i=1}^{n} c_i f_i \equiv 0 \Rightarrow c_1 = \cdots = c_n = 0\,.$$

**Definition 1.11.** *We say a set of functions* $\phi_1, \ldots \phi_n \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ *are* orthogonal *if*

$$\langle \phi_i, \phi_j \rangle_w = \int_{\mathbb{R}} \phi_i(x)\phi_j(x)w(x)\,\mathrm{d}x = \begin{cases} 0 & i \neq j \\ c_i & i = j \end{cases}\,. \tag{1.11}$$

*If the* $c_i$ *are all equal to 1, then we say the functions are* orthonormal*. If not, one can construct* normalised *versions of them,* $\overline{\phi}_i(x) = \phi_i(x)/\|\phi_i\|_w$*, which are orthonormal.*

Any set of linearly independent functions can be used to construct a set of orthogonal functions (and hence a set of orthonormal functions), a process which is called *orthogonalisation*. Algorithm 1 shows the well-known *Gram–Schmidt* procedure for orthogonalisation.

**Algorithm 1** Gram–Schmidt orthogonalisation

```
1: function GRAM–SCHMIDT(f_1,...,f_n, w)
2:     for i ← 0,...,n do
3:         φ_i ← f_i
4:         for j = 0,...,i − 1 do
5:             φ_i ← φ_i − ⟨f_i, φ_j⟩_w φ_j
6:         end for
7:         φ̄_i ← φ_i/‖φ_i‖_w
8:     end for
9:     return {φ̄_0,...,φ̄_n}
10: end function
```

**Definition 1.12.** *Given a family of orthonormal functions* $\phi_0, \phi_1, \cdots \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$*, and some real function* $f$*, we define the* generalised Fourier expansion *of* $f$ *to be* $\sum_{i=0}^{\infty} f_i\phi_i(x)$ *where the Fourier coefficients are* $f_i = \langle f, \phi_i \rangle_w$*.*

The following theorem (adapted from Theorem 2.1.2 of [158]) considers how Fourier expansions perform as function approximations:

**Theorem 1.13.** *Given* $f \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ *and the orthonormal functions* $\phi_0, \ldots, \phi_n$*, consider all functions of the form* $g_n(x) = \sum_{i=0}^{n} c_i\phi_i(x)$*. The specific* $g_n$ *which minimises* $\|f - g_n\|_w$*, equivalently which minimises* $\|f - g_n\|_w^2$*, is* $g_n^* = \sum_{i=0}^{n} f_i\phi_i = \sum_{i=0}^{n}\langle f, \phi_i\rangle_w\phi_i$*.*

*Proof.* We minimise the error by setting the derivative to zero; the resulting stationary point will be the unique minimiser of the error since it is of quadratic form. The error is

$$\begin{aligned}\|f-g_n\|_w^2 &= \langle f-g_n, f-g_n\rangle_w = \|f\|_w^2 - 2\langle f, g_n\rangle_w + \|g_n\|_w^2 \\ &= \|f\|_w^2 - 2\int_{\mathbb{R}} f(x)[\sum_{i=0}^{n} c_i\phi_i(x)]w(x)\,\mathrm{d}x + \int_{\mathbb{R}}\sum_{i,j=0}^{n} c_ic_j\phi_i(x)\phi_j(x)w(x)\,\mathrm{d}x\,.\end{aligned}$$

Taking derivatives inside the integrals yields

$$\begin{aligned}\frac{\mathrm{d}\|f-g_n\|_w^2}{\mathrm{d}c_k} &= \int_{\mathbb{R}}\sum_{i=0}^{n} 2c_i\phi_i(x)\phi_k(x)w(x)\,\mathrm{d}x - 2\int_{\mathbb{R}} f(x)\phi_k(x)w(x)\,\mathrm{d}x \\ &= 2\Big[\sum_{i=0}^{n} c_i\langle\phi_i,\phi_k\rangle_w - \langle f,\phi_k\rangle_w\Big] = 2\Big[c_k - \langle f,\phi_k\rangle_w\Big]\end{aligned}$$

where the last equality follows from the orthonormality of the $\phi_i$. Setting each derivate to zero, $\frac{\mathrm{d}}{\mathrm{d}c_k}\|f-g_n\|_w^2 \equiv 0$, yields the stated minimiser as $c_k = \langle f,\phi_k\rangle_w$. □

This theorem motivates using approximations of the form $g_n^*(x) := \sum_{i=0}^{n}\langle f,\phi_i\rangle_w\phi_i(x)$. The error for the $g_n^*$ approximation is

$$\|f - \sum_{i=1}^{n} f_i\phi_i\|_w^2 = \int_{\mathbb{R}} f(x)^2 w(x)\,\mathrm{d}x - \sum_{i=1}^{n} f_i^2 = \|f\|_w^2 - \sum_{i=1}^{n} f_i^2\,.$$

From this, we can see that increasing $n$ will not increase the error, so the sequence of non-negative errors

$$\|f-g_1^*\|_w^2 \geq \|f-g_2^*\|_w^2 \geq \|f-g_3^*\|_w^2 \geq \dots$$

must converge to a limiting value. It is desirable for this limit to be 0.

**Definition 1.14.** *If for every $f \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$, with $g_n^* = \sum_{i=1}^{n}\langle f,\phi_i\rangle_w\phi_i$, we satisfy $\|f-g_n^*\|_w^2 \to 0$ as $n\to\infty$, we say that the orthonormal sequence $\{\phi_i\}_{i\in\mathbb{N}_0}$ is* complete *in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$.*

When we work with a complete orthonormal sequence $\{\phi_i\}_{i\in\mathbb{N}_0}$, then for any function

$f \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ and $\varepsilon > 0$, we can find an $N$ so that

$$\|f - g_N^*\|_w^2 \leq \varepsilon\,.$$

We return to the question of completeness after first specifying the orthonormal sequences we will use.

Now, consider the case where the orthonormal sequence $\{\phi_i\}_{i\in\mathbb{N}_0}$ are polynomials.

**Definition 1.15.** *We say $\{p_k\}_{k\in\mathbb{N}_0}$ are the orthonormal polynomials for $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ if: i) $p_k$ is a polynomial of order $k$, and ii) the polynomials are orthonormal as per Definition 1.11.*

These can be constructed from the sequence $1, x, x^2, \ldots$ of monomials. The monomials are linearly independent for all choices of $w$, so one can orthogonalise them by using the iterative Gram–Schmidt outlined in Algorithm 1.

Alternatively, there exists a somewhat complicated direct method for constructing them. We denote the moments of the weight pdf as $m_i = \int_{\mathbb{R}} x^i w(x)\,\mathrm{d}x$, and construct Hankel matrices of them as

$$\boldsymbol{H}_n = \begin{pmatrix} m_0 & m_1 & m_2 & \cdots & m_n \\ m_1 & m_2 & m_3 & \cdots & m_{n+1} \\ & & \vdots & & \\ m_{n-1} & m_n & m_{n+1} & \cdots & m_{2n-1} \\ m_n & m_{n+1} & m_{n+2} & \cdots & m_{2n} \end{pmatrix}, \quad n \in \mathbb{N}_+\,.$$

Lastly, if we denote $\widetilde{\boldsymbol{H}}_n(x)$ to be the Hankel matrix $\boldsymbol{H}_n$ where the last row is replaced by $(1, x, x^2, \ldots, x^n)$, we can write

$$p_n(x) = \frac{1}{\sqrt{\det(\boldsymbol{H}_{n-1})\det(\boldsymbol{H}_n)}} \det(\widetilde{\boldsymbol{H}}_n(x))\,, \quad n \in \mathbb{N}_+\,. \tag{1.12}$$

For more details, see [158, pp. 26–27]. This method is performed in Appendix 3.A to describe the orthonormal polynomials w.r.t. the pdf of a lognormal distribution.

For orthonormal polynomial systems, there is a simple condition on the weight function

$w$ which implies completeness:

**Proposition 1.16.** *If there exists an $\alpha > 0$ such that*

$$\int_{\mathbb{R}} \mathrm{e}^{\alpha|x|} w(x)\,\mathrm{d}x < \infty\,,$$

*then the orthonormal polynomials $\{p_k\}_{k\in\mathbb{N}_0}$ are complete in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$.*

See [159, p. 333] for the proof.

### 1.3.4 Monte Carlo techniques

When analytic integration and quadrature both fail, we can attempt *Monte Carlo integration (MCI)*. Say we want to estimate $\ell = \mathbb{E}[g(\boldsymbol{X})]$. If we can sample values $R \in \mathbb{N}_+$ independent and identically distributed (iid) vectors from the distribution $f_{\boldsymbol{X}}$, denoted $\boldsymbol{X}^{[r]} \overset{\text{iid}}{\sim} f_{\boldsymbol{X}}$, then we have the *crude Monte Carlo (CMC)* estimator

$$\widehat{\ell}_{\text{CMC}} = \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{X}^{[r]})\,. \tag{1.13}$$

The estimator is unbiased, meaning $\mathbb{E}[\widehat{\ell}_{\text{CMC}}] = \ell$, and by the law of large numbers we know $\widehat{\ell}_{\text{CMC}} \xrightarrow{\text{a.s.}} \ell$ as $R \to \infty$.

When $\sigma^2 := \mathbb{V}\text{ar}[g(\boldsymbol{X})] < \infty$, the central limit theorem applies, so

$$\sqrt{R}\,\widehat{\ell}_{\text{CMC}} \xrightarrow{\mathcal{D}} \mathsf{Normal}(\ell, \sigma^2) \qquad \text{as } R \to \infty\,. \tag{1.14}$$

If we choose $R$ to be large, then we can say $\sqrt{R}(\widehat{\ell}_{\text{CMC}} - \ell)/\sigma \mathrel{\dot\sim} \mathsf{Normal}(0,1)$ and generate approximate confidence intervals at significant level $\alpha \in (0,1)$,

$$\mathbb{P}\Big(\ell - q_{1-\alpha/2}\frac{\sigma}{\sqrt{R}} \le \widehat{\ell}_{\text{CMC}} \le \ell + q_{1-\alpha/2}\frac{\sigma}{\sqrt{R}}\Big) \approx 1-\alpha\,. \tag{1.15}$$

To actually evaluate this, we substitute the standard estimate $\widehat{\sigma}$ for $\sigma$,

$$\widehat{\sigma}^2 = \frac{1}{R-1}\sum_{r=1}^{R}(g(\boldsymbol{X}^{[r]}) - \widehat{\ell}_{\text{CMC}})^2\,. \tag{1.16}$$

Alternatively, we could use bootstrapping to produce approximate confidence intervals without assuming that the asymptotic normality of $\widehat{\ell}_{\text{CMC}}$ has already kicked in.

Considering these confidence intervals, we can see this method allows us to escape the event-horizon of the curse of dimensionality. Increasing the accuracy of the estimator by 1 significant figure (for a fixed significance level $\alpha$) corresponds to increasing $R$ by a factor of 100, since (1.14) tells us the error is of order $\mathcal{O}(R^{-1/2})$. The amazing conclusion is that this factor of 100 is unaffected by the dimension $d$ of the integral $\ell$ which is being approximated.

What is also amazing is that increasing $R$ by 100, which is equivalent to increasing the computational burden by 100, is an enormous cost to pay for a paltry significant figure of accuracy. We return to this problem later when we discuss quasi-Monte Carlo.

Next, we consider the common algorithms for generating random variables from a specified distribution, then consider *variance reduction techniques* which can greatly improve the efficiency of the Monte Carlo method.

### Sampling uniforms

Sampling from any probability distribution relies upon a sequence of random numbers from the $\mathsf{Uniform}(0,1)$ distribution. Typically, we do not use truly random numbers, but pseudorandom numbers which mimic the behaviour of uniform random variables. This means that the $n$-th pseudorandom number $u_n$ is a deterministic function of the previous $u_{n-1}$; cf. Chapter 1 of [110] for more on pseudorandom number generation, and the famous generator called the *Mersenne twister*.

### Inverse transform method

Say that we want to generate $X^{[r]} \overset{\text{iid}}{\sim} f_X$, we can set

$$X^{[r]} = F_X^{\leftarrow}(U^{[r]}) \text{ where } U^{[r]} \overset{\text{iid}}{\sim} \mathsf{Uniform}(0,1)\,,$$

where $F_X^{\leftarrow}(u)$ is the (quasi-)inverse $F_X^{\leftarrow}(u) := \inf\{x : F_X(x) \geq u\}$ for $0 \leq u \leq 1$. How can we evaluate $F_X^{\leftarrow}$? It is sometimes known analytically, the canonical example being $X \sim \mathsf{Exponential}(\lambda)$, so $F_X(x) = 1 - \mathrm{e}^{-\lambda x}$, and hence $F^{\leftarrow}(u) = -\log(1-u)/\lambda$.

In most other cases, we need to perform a root-finding step, for example using the Newton–Raphson method, to invert $F_X$ approximately. If we cannot even (analytically) integrate the pdf $f_X$ to find the cdf $F_X$, then this method is almost hopeless. Similarly, if we only know $f_X$ up to a constant of proportionality, or if we want to simulate $d > 1$ dimensions, then we must turn to other methods.

One useful aspect of the inverse transform method is that it allows one to simulate $X$ conditional on $X > \gamma$, by

$$X^{[r]} = F_X^{\leftarrow}(\widetilde{U}^{[r]}) \text{ where } \widetilde{U}^{[r]} \overset{\text{iid}}{\sim} \mathsf{Uniform}(F_X(\gamma), 1)\,.$$

As $\gamma$ becomes very large ($\mathbb{P}(X > \gamma)$ becomes very small), then we are evaluating $F_X^{\leftarrow}$ near one. This can lead to numerical instability, so care must be taken that this technique does not return $X^{[r]} = \infty$ or $X^{[r]} = \mathrm{NaN}$.

### Acceptance–rejection

Assume we have a distribution $f_Y$, called the *proposal distribution*, and a $C > 0$ such that

$$f_X(x) \leq C f_Y(x) \quad \forall x \in \mathbb{R}\,. \tag{1.17}$$

If we can sample from the proposal distribution, then we can sample $f_X$ by Algorithm 2.

**Algorithm 2** Acceptance–rejection

```
function AcceptanceRejection(f_X, f_Y, C)
    while True do
        Y ~ f_Y, U ~ Uniform(0, 1)
        if U ≤ f_X(Y)/C f_Y(Y) then
            return Y
        end if
    end while
end function
```

Acceptance–rejection is quite general, as it works if $f_X(x)$ is substituted for $\widetilde{f}_X(x) \propto f_X(x)$ in (1.17) and Algorithm 2. Figure 1.4 illustrates an example of this.

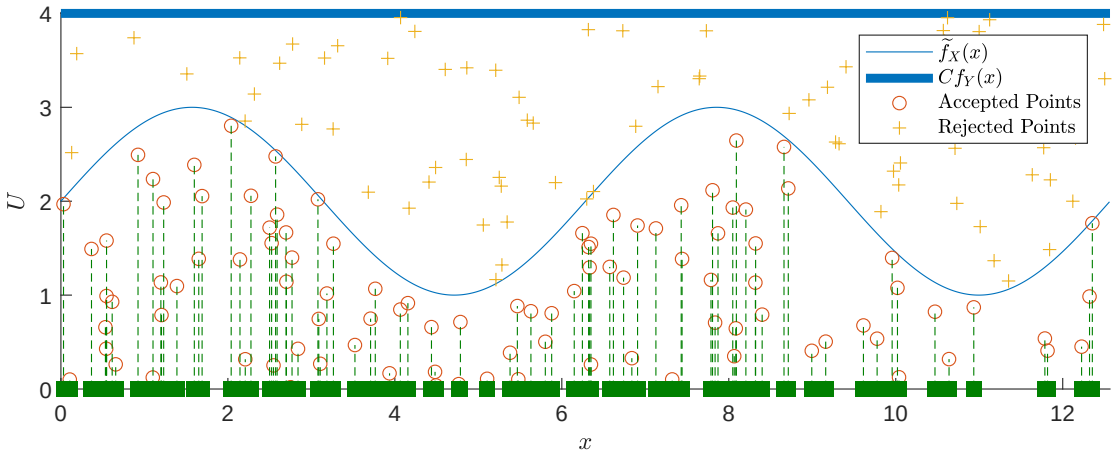



Figure 1.4: Acceptance–rejection sampling from $\widetilde{f}_X(x) = 2+\sin(x)$ for $x \in [0, 4\pi]$ using the proposal distribution $\mathsf{Uniform}[0, 4\pi]$, $C = 16\pi$, and sampling $R = 100$ random variables.

This efficiency of acceptance–rejection depends heavily upon the constant $C$. If we use $f_X$ and not $\widetilde{f}_X$ above (i.e. we know the normalising constant of the desired distribution) then $C$ can be easily interpreted. That is, the expected number of samples from $f_Y$ needed to accept one sample from $f_X$ is $1/C$. So the optimal value of $C$ is 1; then, we never generate proposal samples which are wastefully discarded.

Determining $C$ can be laboriously done by hand, or it can be the result of a root-finding algorithm. Extending to $d > 1$ dimensions is possible, but the determination of $C$ becomes even more difficult.

### Markov chain Monte Carlo

The most general, and most complicated, form of sampling is *Markov chain Monte Carlo (MCMC)*. While the algorithms listed above sample exactly from the desired target distribution $f_X$, MCMC samples will only be approximately from this distribution. Furthermore, the samples produced will not be independent and will not be identically distributed.

The main idea is to construct a Markov chain, $\{X_n\}_{n\in\mathbb{N}_0}$, which has a stationary distribution which is equal to the target distribution $f_X$, then the LLN and the CLT both apply to the sequence $\frac{1}{j}\sum_{i=0}^{j} g(X_j)$. We will skip the Markov chain details as they are not relevant to this thesis, cf. [124], or for a general treatment on MCMC see any detailed textbook on Monte Carlo such as [15, 79, 110].

The crucial step is to provide a *transition kernel* $q(x \hookrightarrow y)$, which for every fixed $x$ in the support of $f_X$ is a pdf in $y$, i.e. $q(x \hookrightarrow \cdot) \geq 0$, and $\int_{\mathbb{R}} q(x \hookrightarrow y)\,\mathrm{d}y = 1$. Also, we need to be able to simulate from $q(x \hookrightarrow \cdot)$. If this is satisfied, then sampling becomes a sequence of *Metropolis–Hasting* steps:

**Algorithm 3** Markov chain Monte Carlo

**function** MCMC($f_X$, $R$, $q$, $X_0$)
  **for** $r = 1$ to $R$ **do**
    $X_r^* \sim q(X_{r-1} \hookrightarrow \cdot)$
    $U \sim \mathsf{Uniform}(0, 1)$
    **if** $U \leq [f_X(X_r^*)q(X_r^* \hookrightarrow X_{r-1})]/[f_X(X_{r-1})q(X_{r-1} \hookrightarrow X_r^*)]$ **then**
      $X_r \leftarrow X_r^*$
    **else**
      $X_r \leftarrow X_{r-1}$
    **end if**
  **end for**
  **return** $(X_1, \ldots, X_R)$
**end function**

Sampling from a complicated distribution using MCMC can be more of an art than a science. The MCMC practitioner's toolbox includes a vast array of tricks: to assess whether the Markov chain has reached stationarity (often the samples from the start of the chain, called the *burn in* period, are discarded), to see the effective number of samples (which is $\leq R$, considering that many samples will be duplicated), to choose the arbitrary starting point(s) $X_0$ (it is common to run many chains from multiple starting points).

As alluded to earlier, the most important choice for the MCMC practitioner is to set the transition kernel. Being supplied Algorithm 3 without a specific transition kernel, is like being given a car without an engine. The MCMC literature supplies a baffling array of kernel options, each with varying degrees of hyperparameters which may need tuning. Modern MCMC research and software tools have thankfully moved in the direction of

automatically choosing a good transition kernel, either by allowing the transition kernel to adapt to the target function $f_X$ while MCMC is in progress (called adaptive methods), or by using a technique called Hamiltonian MCMC.

MCMC is increasingly the only possible option for evaluating high-dimensional complicated integrals, especially if they are fit into the framework of Bayesian statistics. Yet if of the other integration techniques mentioned in this chapter were able to be used instead of MCMC, then they almost certainly should be used. In this sense, MCMC is the worst form of integration, except for all the others.

### Common random numbers

We now turn to the problem of variance reduction. Consider the case where we have a Monte Carlo estimator $\widehat{f}$ of a pdf $f$, and we wish to estimate this density at many points, $\widehat{f}(x_1) \approx f(x_1), \ldots, \widehat{f}(x_n) \approx f(x_n)$. In this case, we can treat each problem separately, using $R$ random variables to construct an estimator $\widehat{f}(x_1)$ then another $R$ random variables to construct $\widehat{f}(x_2)$, and so on. Alternatively, we can generate $R$ random variables and share these between all $n$ estimation problems. This is called using *common random numbers (CRN)*.

This produces a smoothing effect, illustrated by Figure 1.5 where the pdf of the sum of thirty iid $\mathsf{Gamma}(3, 2)$ random variables is estimated with and without CRN using the Monte Carlo estimator in [112].[1] While using CRN creates a more realistic result for miniminal effort, the effect diminishes as $R$ becomes larger.

[1] This is a toy problem as the sum is Erlang distributed.

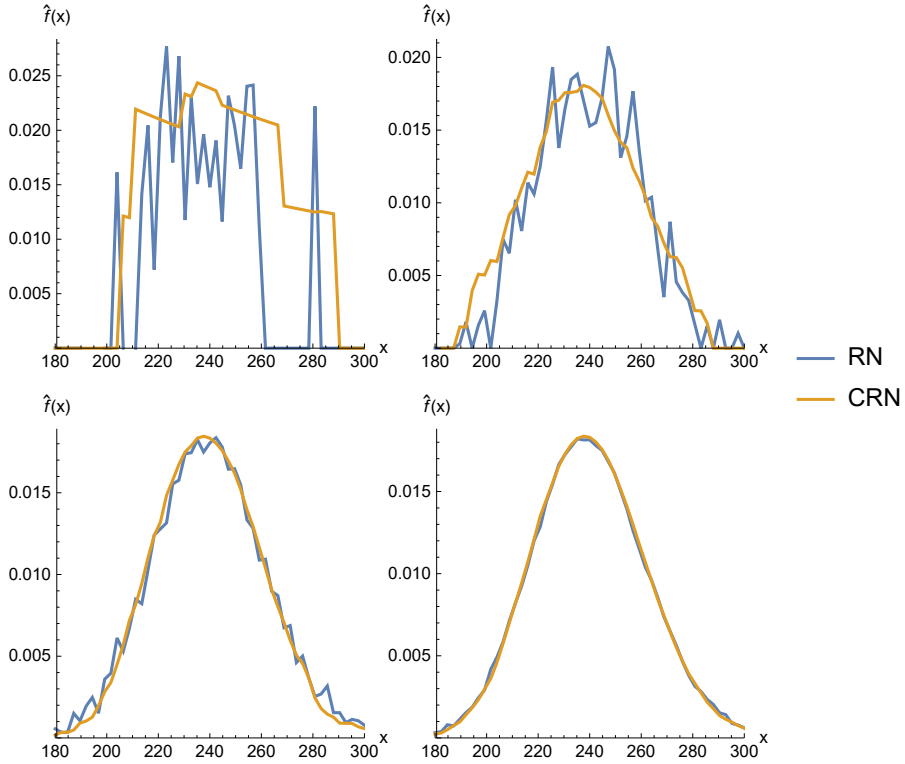



Figure 1.5: Estimating a pdf with and without CRN, for $R = 10$, $10^2$, $10^3$, $10^4$ resp.

### Importance sampling

One of the most useful variance reduction techniques follows from the old trick of multiplying by one. The expectation being evaluated, as an integral, is

$$\ell = \mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x = \int_{\mathbb{R}} g(x) f_X(x) \frac{f_Y(x)}{f_Y(x)}\,\mathrm{d}x = \mathbb{E}[g(Y)\frac{f_X(Y)}{f_Y(Y)}]$$

where $X \sim f_X$ and $Y \sim f_Y$, and so the *importance sampling (IS)* Monte Carlo estimator is

$$\widehat{\ell}_{\mathrm{IS}} = \frac{1}{R}\sum_{r=1}^{R} g(Y^{[r]}) \frac{f_X(Y^{[r]})}{f_Y(Y^{[r]})}\,, \quad \text{for } Y^{[r]} \overset{\text{iid}}{\sim} f_Y\,.$$

This approach holds so long as there are $x$ where $f_Y(x) = 0$ but $f_X(x) > 0$, or in words, the support of $Y$ includes the support of $X$ (the jargon for this is that $f_X$ is absolutely continuous w.r.t. $f_Y$). The pdf $f_Y$ is called the *proposal density*, and the

fraction $f_X(Y)/f_Y(Y)$ is called the *likelihood ratio* (or the Radon–Nikodým derivative) and can be seen as correcting for the fact that we have changed the distribution which we are sampling from.

It is slightly misleading to call this technique a variance reduction algorithm, since with a poor choice of proposal pdf we can easily create estimators with *more variance* than CMC. Yet for a well-chosen density, the variance reduction can be many orders of magnitude.

A well-known fact of importance sampling for estimating $\ell = \mathbb{P}(X < x)$ is that

$$f_Y(x) = \frac{\mathbb{I}\{X < x\}}{\mathbb{P}(X < x)} f_X(x)$$

is the proposal density which minimises the variance of $\widehat{\ell}_{\mathrm{IS}}$. It is easy to see that this proposal yields an unbiased estimator (all IS estimators are unbiased) which has zero variance!

$$\widehat{\ell}_{\mathrm{IS}} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\{Y^{[r]} < x\} \frac{f_X(Y^{[r]})}{\frac{\mathbb{I}\{Y^{[r]}<x\}}{\mathbb{P}(X<x)} f_X(Y^{[r]})} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{P}(X < x) = \ell\,.$$

Obviously the result is without immediate practical value, since to create this proposal density we need to normalise by the unknown probability which is being estimated in the first place. However, it does provide the intuition for selecting proposal densities: choose proposals which increases the probability of interesting events.

If the problem is to estimate $\ell = \mathbb{P}(X > \gamma)$ for a large $\gamma$, then proposals should be taken which increase the probability of large samples. One scheme, called *exponential tilting* (or *twisting*), is frequently used, in which

$$f_Y(x) = \frac{\mathrm{e}^{\theta x}}{\mathbb{E}[\mathrm{e}^{\theta X}]} f_X(x) \tag{1.18}$$

where $\theta$ is a positive constant (it can be negative if we wish to induce more small samples). The optimal choice of $\theta$ can be calculated, and it is the value which satisfies $\mathbb{E}[Y] = \gamma$.

Exponential tilting (with a positive $\theta$) is only applicable when the original distribution $f_X$ has a well-defined moment generating function (otherwise $\mathbb{E}[\mathrm{e}^{\theta X}] = \infty$). A similar

technique which can be employed (if $X$ is a positive random variable) in this scenario is *hazard rate twisting*,

$$f_Y(x) = (1-\theta) f_X(x) \mathrm{e}^{\theta \Lambda_X(x)}$$

in which $\Lambda_X(x) = \int_0^x \lambda_X(x)\,\mathrm{d}x$ where $\lambda_X$ is the hazard rate of $X$, i.e., the ratio of its pdf to its survival function [105].

One limitation for any IS scheme is that the proposal density must be chosen to reduce variance, but must also allow us to simulate from it. Luckily there are many cases where the exponentially-tilted pdf describes a distribution within the same family as the original pdf (this is true for distributions in the *natural exponential family*, including the normal, gamma, Poisson and Weibull distributions), and so the difficulty in simulation is not increased over CMC.

### Conditional Monte Carlo

Sometimes we can introduce extra theoretical knowledge to a Monte Carlo problem to decrease the variance. This is the main idea of *conditional Monte Carlo*, in which we simplify certain problems by using properties of conditional probability. This is easiest seen with an example: say we wish to estimate $\ell = \mathbb{P}(X_1 + X_2 > \gamma)$, then the CMC estimator is

$$\widehat{\ell}_{\mathrm{CMC}} = \frac{1}{R} \sum_{r=1}^{R} \mathbb{I}\{X_1^{[r]} + X_2^{[r]} > \gamma\}\,, \quad (X_1, X_2) \overset{\text{iid}}{\sim} F_{\boldsymbol{X}}\,.$$

Yet we also know that $\mathbb{P}(X_1 + X_2 > \gamma \mid X_2 = x_2) = \mathbb{P}(X_1 > \gamma - x_2) = \overline{F}_{X_1}(\gamma - x_2)$, which motivates a conditional MC estimator

$$\widehat{\ell}_{\mathrm{Cond}} = \frac{1}{R} \sum_{r=1}^{R} \overline{F}_{X_1}(\gamma - X_2^{[r]})\,, \quad X_2 \overset{\text{iid}}{\sim} F_{X_2}\,.$$

An advantage of conditional MC is that the variance will either decrease or stay the same relative to the CMC estimator. This technique can achieve a remarkable variance reduction, as exemplified by the Asmussen–Kroese estimator [21]; also, see [13] for a recent review of conditional MC for sums of random variables.

### Quasi-Monte Carlo

The main ingredient of CMC is randomness, and though we know that on average CMC's results are accurate, we can simply be unlucky and see an estimation which is very innaccurate. How does this happen, if for example, we are trying to estimate $\ell = \int_{[0,1]^d} f(\boldsymbol{u})\,\mathrm{d}u$ by $\widehat{\ell} = \frac{1}{R}\sum_{r=1}^{R} f(\boldsymbol{U}^{[r]})$ for $\boldsymbol{U}^{[r]} \overset{\text{iid}}{\sim} \mathsf{Uniform}([0,1]^d)$? It can occur if the $\boldsymbol{U}^{[r]}$ cluster together, and the integrand's behaviour is not fully considered over its range $[0,1]^d$. This leads to the question, can we constrain the $\boldsymbol{U}^{[r]}$ so that they do not overly cluster together?

This reasoning leads us to *quasi-Monte Carlo (QMC)*, where random $(\boldsymbol{U}^{[1]}, \ldots, \boldsymbol{U}^{[R]})$ sampling points are replaced by deterministic $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R)$ which are designed to be evenly spread across the $d$-dimensional hypercube. Figure 1.6 shows scatterplots of 2 dimensional uniform points and from a QMC *low-discrepency sequence* called Sobol's sequence.

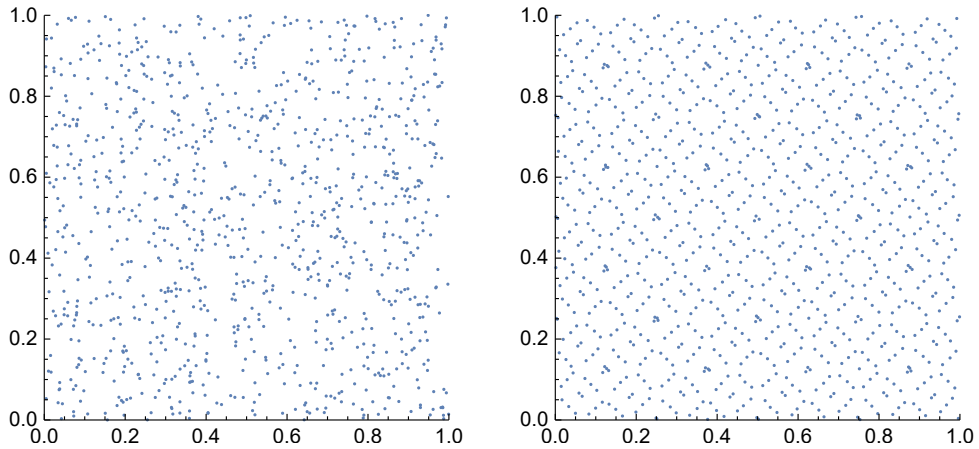



Figure 1.6: Scatterplot of $R = 1024$ iid two-dimensional $\mathsf{Uniform}(0,1)$ random variables, and 1024 points of the two-dimensional Sobol sequence resp.

Designing and analysing new QMC schemes is a field of ongoing research, with a large active community. An allure of QMC is that there are bounds on the integration error, e.g. for a smooth integrand $f$ the *Koksma–Hlawka inequality* [15, 61, 79] is

$$\left| \int_{[0,1]^d} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} - \frac{1}{R}\sum_{r=1}^{R} f(\boldsymbol{u}_r) \right| \leq V(f) D^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R)$$

where $V(f)$ is the Hardy–Krause variation, and $D^*$ is the star discrepency $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R)$.

As $D^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R) = \mathcal{O}((\log R)^d/R) = \mathcal{O}(R^{-(1-\varepsilon)})$ for all $\varepsilon > 0$, we can see that QMC techniques offer an error which can be close to $\mathcal{O}(R^{-1})$. The comparable CMC error guarantee, arrived at by we rearranging the approximate confidence intervals result in (1.15) is

$$\left| \int_{[0,1]^d} f(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} - \frac{1}{R} \sum_{r=1}^{R} f(\boldsymbol{U}^{[r]}) \right| \leq q_{1-\alpha/2} \frac{\sigma}{\sqrt{R}}$$

with probability $1 - \alpha$. While the two statements are not directly comparable (the CMC claim is a probabilistic bound), it seems that for $R$ large enough QMC allows us to break out of unimpressive $\mathcal{O}(R^{-1/2})$ rate up to something approaching $\mathcal{O}(R^{-1})$, if the dimension of the problem is not too high.

### Rare events

One common problem in financial and insurance applications is estimating the likelihood of rare events (e.g. market crashes, or extreme climate events). In this context, using CMC can be difficult. For example, let us say that we want to estimate $\ell(\gamma) = \mathbb{P}(X > \gamma)$ with

$$\widehat{\ell}_{\mathrm{CMC}} = \frac{1}{R} \sum_{r=1}^{R} \mathbb{I}\{X^{[r]} > \gamma)\} \quad \text{where } \boldsymbol{X}^{[r]} \overset{\text{iid}}{\sim} f_X \,. \tag{1.19}$$

If the true probability is $\ell(\gamma) = 10^{-10}$, and we set $R = 10^6$ then about 99.99% of the time we simply get the estimate $\widehat{\ell}_{\mathrm{CMC}} = 0$. Since every indicator function returned 0, then the variance estimate for $\widehat{\ell}_{\mathrm{CMC}}$ is also 0, and hence the confidence intervals are also $[0, 0]$.

A simple solution to this problem is to employ importance sampling, though this can lead to likelihood degeneration for large enough $\gamma$. More complicated iterative methods can be employed, such as the *cross-entropy method* [57], or *multi-level splitting* [80, 81, 42]. These methods are similar to MCMC in that they require great care in choosing parameters and can significantly increase the computing time necessary to allow for the increased generality. For more details, see [148] or [110].

Given an IS estimator, we can categorise its rare-event performance into different categories.

**Definition 1.17.** *An estimator $\widehat{p}_\gamma$ of some rare probability $p_\gamma$ which satisfies $\forall \varepsilon > 0$*

$$\limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{p}_\gamma}{p_\gamma^{2-\varepsilon}} = 0 \qquad \limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{p}_\gamma}{p_\gamma^{2}} < \infty \qquad \limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{p}_\gamma}{p_\gamma^{2}} = 0$$

*has* logarithmic efficiency*,* bounded relative error*, or* vanishing relative error *respectively. These are in increasing order of strength, that is,*

$$\textit{Vanishing relative error} \;\Rightarrow\; \textit{Bounded relative error} \;\Rightarrow\; \textit{Logarithmic efficiency}.$$

These allow us to compare two competing methods on their theoretical properties. Of course, just as the seemingly inefficient $\mathcal{O}(n^2)$ *quicksort* algorithm is usually faster than the $\mathcal{O}(n \log n)$ *mergesort* algorithm, the fact that an estimator satisfies vanishing relative error does not always translate into better variance reduction.

### 1.3.5 Dependence and copulas

A theme of modern probability research, and of this thesis, is to relax the independence assumptions which constrict older models. *Copulas* allow us to analyse just the dependence structure of a random vector, without considerations of the marginal distributions. A copula is a joint cdf for a random vector of $\mathsf{Uniform}(0,1)$ random variables. *Sklar's theorem* which is shown below (adapted from [132]) is the foundational result which shows the generality of copulas.

**Theorem 1.18** (Sklar's theorem)**.** *Consider a random vector $\boldsymbol{X} = (X_1, X_2)$ which has the joint cdf $F_{\boldsymbol{X}}$ and marginal cdfs $F_{X_1}$ and $F_{X_2}$. We can write*

$$F_{\boldsymbol{X}}(x_1, x_2) = C(F_{X_1}(x_1), F_{X_2}(x_2)) \tag{1.21}$$

*where $C$ is a copula, and if $F_{X_1}$ and $F_{X_2}$ are both continuous then $C$ is unique. Conversely, if $C$ is a copula, and $F_{X_1}$ and $F_{X_2}$ are cdfs, then the function $C(F_{X_1}(\cdot), F_{X_2}(\cdot))$ is a joint cdf with marginals $F_{X_1}$ and $F_{X_2}$.*

The definition (1.21) can be extended to any number of dimensions, and the definition can be reversed to get

$$C(u_1, \ldots, u_d) = \mathbb{P}(X_1 \leq F^{\leftarrow}_{X_1}(u_1), \ldots, X_d \leq F^{\leftarrow}_{X_d}(u_d)) \,. \tag{1.22}$$

Also, taking derivatives of (1.21) allows us to write

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{d} f_{X_i}(x_i) \times c(F_1(x_1), \ldots, F_d(x_d))$$

where $c$ is the *copula density*,

$$c(u_1, \ldots, u_d) = \frac{\mathrm{d}^d}{\mathrm{d}u_1 \ldots \mathrm{d}u_d} = C(u_1, \ldots, u_d) \,.$$

In general, the copula density can be difficult to calculate.

The relation (1.22) is useful for extracting the copula given a random vector $F_{\boldsymbol{X}}$, for example, we can define the *Gaussian copula* to be

$$C(u_1, \ldots, u_d) = \mathbb{P}(\Phi(X_1) \leq u_1, \ldots, \Phi(X_d) \leq u_d) \quad \text{where } \boldsymbol{X} \sim \mathsf{Normal}(\mathbf{0}, \boldsymbol{\Sigma}) \,.$$

Similarly, the $t$-copula can be extracted from a multivariate $t$-distribution. McNeil et al. [123] call these copulas, which are determined by well-known multivariate distributions, *implicit copulas*.

As the implicit copulas do not allow for a great variety of dependence behaviours, it is useful to consider another class of copulas well the *Archimidean copulas*. We say $\boldsymbol{X}$'s dependence structure is given by an Archimedean copula with generator $\psi$ if its cdf is

$$C(u_1, \ldots, u_d) = \phi\Big(\sum_{i=1}^{d} \psi(u_i)\Big)$$

where $\phi := \psi^{[-1]}$ is the (pseudo-)inverse of $\psi$, defined as

$$\psi^{[-1]}(t) = \begin{cases} \psi(t) & 0 \leq t \leq \psi(0) \\ 0 & \psi(0) \leq t \leq \infty \end{cases} .$$

These copulas are exchangeable, i.e. $C(u_1, u_2) = C(u_2, u_1)$, and have the property that if $\boldsymbol{X}$'s dependence is specified by the generator $\psi$, then $\boldsymbol{X}_{-i}$ is also specified by this generator (we have closure of dependence structure under variable subsets). One useful property is that the Archimedean copula density can be written somewhat explicitly as

$$c(u_1, \ldots, u_d) = \phi^{(d)}\Big[\sum_{i=1}^{d} \psi(u_i)\Big] \prod_{i=1}^{d} \psi'(u_i)\,.$$

One general problem when using copulas is that any arbitrary copula may be difficult to simulate from. As many software packages currently have limited or no support for copulas, users must write their own code for simulating them If we can simulate from the copula, then simulating any vector with this dependence structure is simply done by

$$\boldsymbol{X} = (F_1^{\leftarrow}(U_1), \ldots, F_d^{\leftarrow}(U_d))\,, \quad \text{where } \boldsymbol{U} \sim C(\cdot)\,.$$

The standard approach for copula simulation, the *conditional distribution method*, dictates that we use the inverse transform method on the conditional distributions:

$$U_1 \sim \mathsf{Uniform}(0,1)\,, \quad U_2 \sim C^{\leftarrow}(\,\cdot \mid U_1)\,, \quad \ldots \quad , \quad U_d \sim C^{\leftarrow}(\,\cdot \mid U_1, \ldots, U_{d-1})\,,$$

where $C^{\leftarrow}(\,\cdot \mid u_1, \ldots, u_{i-1})$ is the inverse of $\mathbb{P}(U_i = \cdot \mid U_1 = u_1, \ldots, U_{i-1} = u_{i-1})$. For Archimedean copulas this has the somewhat formidable form (cf. Cambou et [41])

$$C^{\leftarrow}(u_i \mid u_1, \ldots, u_{i-1} = \phi\Big\{\phi^{(i-1)^{\leftarrow}}\Big[u_i \phi^{(i-1)}\Big(\sum_{j=1}^{i-1} \psi(u_i)\Big)\Big] - \phi^{(i-1)}\Big(\sum_{j=1}^{i-1} \psi(u_i)\Big)\Big\}\,.$$

For Archimedean copulas, there is an alternative approach which uses the *Marshall–Olkin form* of the copula. That is, if we can write the generator inverse as $\phi(s) = \mathbb{E}[\mathrm{e}^{-sZ}]$ for some positive random variable $Z$ with cdf $F_Z$, then an $\boldsymbol{X}$ with this dependence structure can be simulated via

$$\boldsymbol{X} = \Big(F_{X_1}^{-1}\Big(\phi\Big(\frac{E_1}{Z}\Big)\Big), \ldots, F_{X_n}^{-1}\Big(\phi\Big(\frac{E_n}{Z}\Big)\Big)\Big), \quad E_i \overset{\text{iid}}{\sim} \mathsf{Exponential}(1),\ Z \sim F_Z\,.$$

This form is utilised in Asmussen [13], [112], and in Chapter 5 below.

Nelsen [132] and Joe [103] are the common references on copulas, though I would recommend McNeil et al. [123] especially for the application of copulas to financial modelling. Also, see Mikosch [125] and the many responses for a lively debate around the limitations of copula analysis.

### 1.3.6 Asymptotic analysis and extreme value theory

A common theme in this thesis is to consider complicated expressions, and see how they behave with extremely large or extremely small inputs. This is called asymptotic analysis, and it can lead to very interesting insights. A prime example of this, is the behaviour of sums of *subexponential random variables*.

A distribution $F_X$ belongs to the subexponential class if

$$\lim_{x\to\infty} \frac{\mathbb{P}(X_1 + \cdots + X_d > x)}{d\,\mathbb{P}(X_1 > x)} = 1\,.$$

A convenient notation for $\lim_{x\to\infty} f(x)/g(x) = 1$ is $f(x) \sim g(x)$ as $x \to \infty$. With this notation, the subexponential property is

$$\mathbb{P}(X_1 + \cdots + X_d > x) \sim d\,\mathbb{P}(X_1 > x)\,, \quad \text{for } x \to \infty\,.$$

This is a case where notation improves comprehension, as we can read the $\sim$ sign as a $\approx$ sign, at least when $x$ is large. So, the story of sums of subexponential variables is this: if the sum is an extremely large value, then it is probably because one of the summands is extremely large (as opposed to all of the summands being large together). This behaviour is aptly named the *principle of the single big jump*, cf. [73]. Chapter 5 heavily uses asymptotic properties of this kind to construct efficient Monte Carlo estimators.

One collection of asymptotic results relating to maxima of random variables is called *extreme-value theory* [58]. A key result in extreme-value theory, which is remarkable as it is so general and so unexpected, is:

**Theorem 1.19** (Fisher–Tippet theorem)**.** *Assume that, as $n \to \infty$ the random variables*

*$(M_n - a_n)/b_n$ converge in distribution to a non-degenerate distribution $G$, where*

$$M_n = \max_{i=1,\dots,n} \{X_i\}\,, \quad \textit{for } X_i \overset{\text{iid}}{\sim} F_X\,,$$

*and $\{a_n\}_{n\in\mathbb{N}_+}$, $\{b_n\}_{n\in\mathbb{N}_+}$ are real-valued sequences. Then, $G$ is either: i) Gumbel distributed, ii) Fréchet distributed, or iii) Weibull distributed.*

Both extreme-value theory, and the related field of *large deviations theory* [60], consider the limiting behaviour of the maximum or sum of random variables where the limit takes the number of the underlying random variables to infinity. These fields are not relevant to this thesis since, in the cases where we consider rare asymptotic probabilities, we look at the probability of a sum or maximum of random variables exceeding a large threshold where the number of random variables is fixed.

Only in some chapters do we use the Fisher–Tippet result to categorise distributions according to their limiting distribution, also called their *maximum domain of attraction (MDA)*. For example, we write $F \in \mathsf{MDA(Fréchet)}$ if the limit in the Fisher–Tippet theorem for the distribution $F$ is Fréchet distributed.

## 1.4 Existing methods and contributions

Since each chapter covers different problems, a specialised literature review is included in each. This chapter outlines some general methods which are used for sums of random variables, and describes the contributions of this thesis.

### 1.4.1 The normal approximation

I will begin this overview of the various methods for approximating sums of random variables with the simplest approximation suggested by the *central limit theorem*. The standard CLT suggests that we approximate the pdf $f$ of a sum $S = X_1 + \cdots + X_d$ by

$$\widehat{f}_{\text{CLT}}(x) = \phi_{\mu,\sigma^2}(x)$$

where $\mu = \mathbb{E}[S]$, $\sigma^2 = \mathbb{V}\text{ar}[X]$, and $\phi_{\mu,\sigma^2}$ is the pdf of $\mathsf{Normal}(\mu, \sigma^2)$. This approximation has many excellent qualities: the approximation matches $S$'s first two moments, the approximation will improve if $d \to \infty$ (specifically, we have uniform convergence of the cdfs), the CLT is a famous and simple theorem, the approximating distribution (the normal distribution) is known explicitly and it is analytically tractable, and the approximation only requires us to find the mean (as easy as $\mu = \sum_{i=1}^{n} \mathbb{E}[X_i]$) and variance of $S$.

The approach is obviously not universally applicable; the CLT does not apply if any $X_i$ has an infinite variance or mean, or if the summands exhibited a strong dependence structure. And though $S$ can certainly exhibit a normal-like behaviour for small values of $d$ — the example in Figure 1.5 of $d = 30$ gamma summands appears very normal-like — one only expects the approximation to be accurate if $d$ is large. Some properties of the $X_i$, such as a high skewness or that they exhibit heavy tails, can also induce non-normality in $S$ when $d$ is small.

Another fundamental problem with the CLT approximation is that it is not adjustable. Other methods allow the user to increase either the computational time and/or the complexity of the approximation to increase accuracy. The CLT is has no such options; the approximation will always be light-tailed, symmetric, and have support over all of $\mathbb{R}$.

### 1.4.2 Beyond the central limit theorem

It wasn't long after the CLT was originally proved that mathematicians set about trying to generalise it to create more accurate approximations [88].[1] They kept the normal distribution in a central role in the approximation, and showed that $\widehat{f}_{\text{CLT}}(x)$ can be seen as the one-term truncation of an asymptotic expansion

$$f(x) = \sum_{i=0}^{\infty} \mathbb{E}[h_i(S)] h_i(x) \phi_{\mu,\sigma^2}(x)\,,$$

where and $\{h_i\}_{i \in \mathbb{N}_0}$ are the *Hermite polynomials* which are orthonormal w.r.t. $\phi_{\mu,\sigma^2}$.[2] This is called the *Gram–Charlier (type A)* expansion or the *Edgeworth* expansion, cf. [26, 108].

[1] The names of the mathematicians who contributed to this effort read like a roll call of famous 19th century mathematicians: Laplace, Poisson, Bessel, Chebyshev, Hermite, Fourier, et cetera.

[2] Note, $h_0(x) = 1$, so $\mathbb{E}[h_0(S)] h_0(x) \phi_{\mu,\sigma^2}(x) = \phi_{\mu,\sigma^2}(x) = \widehat{f}_{\text{CLT}}(x)$.

It can be rewritten as

$$f(x) = \phi_{\mu,\sigma^2}(x) - \frac{\kappa_3}{3!}\phi^{(3)}_{\mu,\sigma^2}(x) + \frac{\kappa_4}{4!}\phi^{(4)}_{\mu,\sigma^2}(x) - \frac{\kappa_5}{5!}\phi^{(5)}_{\mu,\sigma^2}(x) + \frac{1}{6!}(10\kappa_3^2 + \kappa_6)\phi^{(6)}_{\mu,\sigma^2}(x) + \dots$$

where $\kappa_i$ is the $i$-th cumulant of $S$, cf. [30, 51]

Taking inspiration from Fourier's expansion of periodic functions in terms of trigonometric basis functions, mathematicians then considered approximations where $\phi_{\mu,\sigma^2}$ was replaced by an arbitrary pdf $w$, and the Hermite polynomials by the $\{p_i\}_{i\in\mathbb{N}_0}$ polynomials which are orthonormal w.r.t. $w$. We call this more general approach the *orthogonal polynomial expansion*, however the names Gram–Charlier expansion and Edgeworth expansion still exist in recent literature. They took the generalised Fourier expansion of $f/w$

$$\frac{f(x)}{w(x)} = \sum_{i=0}^{\infty} \left\langle \frac{f_S}{w}, p_i \right\rangle_w p_i(x) = \sum_{i=0}^{K} \mathbb{E}[p_i(S)] p_i(x)\,,$$

and truncated it to $K+1$ terms to achieve

$$\widehat{f}_{\mathrm{OP}}(x) = \sum_{i=0}^{K} \mathbb{E}[p_i(S)] p_i(x) w(x)\,.$$

Overall, the results using the orthogonal polynomial approach can be impressive, cf. [83, 84] for some applications in insurance. As $\widehat{f}_S$ is in such a simple form, integrals involving it can often be solved analytically — for example, Dufresne and Li [67] give an explicit form for the price of a (discrete) Asian option using the orthogonal polynomial pdf approximation. Also, a recent R package `PDQUtils` helps to automate this procedure in that language [139].

However, care must be taken to ensure the approximation is theoretically valid. For $\widehat{f}_S$ to converge as $K \to \infty$, we need the check that the integrability condition $f_S/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ is satisfied. Oftentimes this fails, and instead we have to approximate $f_{\widetilde{S}}$ where $\widetilde{S} = g(S)$ is some transformation of the sum (e.g. $\widetilde{S} = \log(S)$, or $\widetilde{S} = 1/S$), then write $f_S$ in terms of $f_{\widetilde{S}}$ by using the change of variables formula. The other theoretical requirement is that the polynomials $\{p_i\}_{i\in\mathbb{N}_0}$ are complete in the space $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$. If they are incomplete, as is the case for $w$ being a lognormal pdf, then $\widehat{f}_S$ will converge

as $K \to \infty$ but not to $f_S$ (in the lognormal case, it converges to a different pdf with the same sequence of moments as $w$). These extra requirements may explain why the the method is most commonly used where $w$ is taken as either a normal pdf or a gamma pdf. Here, the orthonormal polynomials (the Hermite and Laguerre polynomials) are classical and need not be constructed by an orthogonalisation procedure.

Even when the approach is theoretically valid, it can fail numerically if done incorrectly. The coefficients $a_i = \mathbb{E}[p_i(S)]$ can be solved algebraically if the $X_i$ are independent and their moments are known; this is because the moments of $S$ can be (somewhat laboriously) found by expanding $\mathbb{E}[S^n] = \mathbb{E}[(X_1 + \cdots + X_d)^n]$ with the binomial formula. Otherwise, one has to use quadrature or Monte Carlo integration to find the coefficients, and a small error in $a_i$ for a large $i$ can cause a noticeable overall error in $\widehat{f}_{\mathrm{OP}}$. Another practical concern is that the approximations seem to be very sensitive to parameters of $w$; for example, if we choose $w$ to be the pdf of a $\mathsf{Gamma}(r, m)$ distribution, then the specific $r$ and $m$ combination can be very important (one should not simply choose them to be arbitrarily placed in the region where the integrability condition is satisfied). The approximation $\widehat{f}_{\mathrm{OP}}$ can become negative if $K$ is too small, and therefore isn't a true pdf. Lastly, the convergence of $\widehat{f}_{\mathrm{OP}}$ is in terms of absolute error — the approximation's relative error can be significant, especially in the tails.

### 1.4.3 Other approaches

Two algorithms which are commonly used for sums of random variables are *integral transform inversion* and *Panjer's algorithm*. There are many integral transform inversion (here, I am collectively referring to Laplace transform inversion, Fourier transform inversion, and characteristic function inversion) approaches, all of which are variations on the method outlined in Section 1.3.2. Panjer's algorithm is a method which gives density estimates for compound sums, but as it only applies to discrete summands, it has not been considered here (cf. [69] for a comparison of Panjer's algorithm and a Fourier transform inversion algorithm).

Lastly, one can simply ignore the underlying summands and use any univariate approximation technique to fit the sum distribution directly (e.g. by sampling $S^{[1]}, \ldots, S^{[R]}$ with

Monte Carlo). Out of the plethora of statistical techniques which can be applied, we will briefly describe parametric approximations and kernel density estimation.

Parametric approximations assume that the sum distribution belongs to a specific family of distributions (e.g. $\mathsf{Normal}(\mu, \sigma^2)$) and find the specific parameters (e.g. the $\mu$ and $\sigma^2$) which best fit the samples $S^{[1]}, \ldots, S^{[R]}$. One can choose the parameters which maximise the likelihood function, using *maximum likelihood estimation*, or *expectation–maximisation*. Alternatively, if the approximating family has $p$ parameters to fit, one can set them so that the first $p$ moments of the approximation match the sample moments. More generally, we can set the parameters to fit quantities which are meaningful for our specific application — if it is more important for the left tail to be accurate than for the right, then we can select parameters which ensure left-tail accuracy (e.g. [95]). Parametric approximations are a popular technique for approximating the $\mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, where most authors choose an approximation which is $\mathsf{Lognormal}(\cdot, \cdot)$ distributed [71, 153, 4, 29, 72] or from related distributions [28, 95].

The accuracy of the parametric approximations are limited by the number of parameters which specify the approximating family. Thus, the families of distributions which allow an arbitrary number of parameters are common; for example, we can fit an approximation using a mixture distribution with $n$ components, check if the resulting accuracy is sufficient, and if not repeat with a larger $n$. Mixtures of exponential or Erlang distributions are common for positive random variables [164, 116, 163], though these are a special case of *phase-type distributions* which are also prevalent approximations in the literature [34, 93, 94, 12, 14]. Phase-type approximations are sometimes avoided as they cannot produce heavy-tailed approximations [69] but one can take the more general class of infinite-mixtures of phase-type distributions to create both heavy and light tailed approximations [146, 166].

All parametric approximations can be criticised for the arbitrariness of enforcing one particular family of distributions on the data, but the phase-type approximation can be somewhat justified by the fact that these distributions are dense in the class of all continuous distributions on $\mathbb{R}_+$. However, from my contribution to [18], I have found that fitting phase-type distributions using the standard expectation–maximisation algorithm [22, 135] can be a non-trivial numerical challenge which is remarkably slow if the number

of phases is moderately high.

The *kernel density estimator (KDE)* [161] takes the samples $S^{[1]}, \ldots, S^{[R]}$ (e.g. sampled using CMC) and gives the approximation

$$\widehat{f}_{\mathrm{KDE}}(s) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{h}K\Big(\frac{S^{[r]}-x}{h}\Big)\,,$$

for some *bandwidth* $h > 0$, and *kernel* $K$. The most common kernel used is the Gaussian kernel, $K(x) = \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$. On the relative importance of these two choices, Pagan and Ullah [136, p. 19] write that:

> "It is now well known that the choice of kernel is a minor issue, with any kernel being close to an optimal kernel for large samples. In contrast the selection of the window width [a.k.a. bandwidth] $h$ is crucial."

Accordingly, there are now various algorithms for choosing the value of $h$. Section 2.7.1 of [136] and Section 8.5 of [110] lists some methods, including *least squares cross-validation*, the *plug-in method*, and *likelihood cross-validation*, none of which appears to be categorically superior to the others. An alternative KDE method by Botev et al. [36] is available as a MATLAB library for easy use.

## 1.5 Contributions

As the work presented in the remainder of this thesis was done in collaboration with other authors, the pronouns will switch from 'I' to 'we' to reflect the multiple authors (or, more commonly, to refer to the reader and the authors).

In Chapter 2 we consider the $\mathsf{SumLognormal}(\cdot,\cdot)$ distribution, and consider its Laplace transform. We represent the Laplace transform $\mathscr{L}(\theta) = \mathbb{E}[\mathrm{e}^{-\theta S_n}] \propto \int \mathrm{e}^{-h_\theta(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}$ as $\widetilde{\mathscr{L}}(\theta)I(\theta)$, where $\widetilde{\mathscr{L}}(\theta)$ is given in a closed form and $I(\theta)$ is the error factor ($\approx 1$). We obtain $\widetilde{\mathscr{L}}(\theta)$ by replacing $h_\theta(\boldsymbol{x})$ with a second-order Taylor expansion around its minimiser $\boldsymbol{x}^*$. An algorithm for calculating the asymptotic expansion of $\boldsymbol{x}^*$ is presented, and it is shown that $I(\theta) \to 1$ as $\theta \to \infty$. A variety of numerical methods for evaluating $I(\theta)$

is discussed, including Monte Carlo with importance sampling and quasi-Monte Carlo. Numerical examples (including Laplace transform inversion for the density of $S_n$) are also given.

Next, in Chapters 3 and 4 we apply the orthogonal expansion approach to sums of random variables. Chapter 3 focuses on sums from the $S \sim \mathsf{SumLognormal}(\cdot, \cdot)$ distribution, where the summands may have different variances or be dependent. We consider orthogonal expansions for the pdf of $\log S$, and for the exponentially tilted sum density. The reference pdfs considered include the normal, gamma and lognormal densities. When $w$ is the lognormal pdf, we construct the orthonormal polynomials in closed form and show that they are not dense in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$, a result that is closely related to the lognormal distribution not being determined by its moments. This therefore warns against the most obvious choice of taking $w$ as lognormal. Numerical examples are presented and comparisons are made to an established approach, the Fenton–Wilkinson method, and a recent approach, the log skew normal approximation. Also, the extensions to density estimation for statistical data sets and non-Gaussian copulas are outlined.

Chapter 4 also focuses on orthogonal expansions to approach pdfs of sums of random variables, though it instead focuses on compound sums and gives applications for insurance. While the chapter provides new results for orthogonal expansions of compound sums, it can also be seen as an in-depth comparison (in the style of [69]) of orthogonal expansions to the Laplace transform inversion technique. The motivating application is to evaluate the stop-loss premium associated to a non-proportional global reinsurance treaty. The orthogonal expansion uses the gamma density as its reference pdf.

Orthogonal expansions aim to give pdf approximations which are accurate across the whole support of the random variable. Instead, it is sometimes more valuable to have the pdf or the cdf at a single point be approximated with high accuracy. I have included some work-in-progress in Chapter 5 for accurate approximation of the survival function of a sum of random variables. Here, the proposed estimator is based on an importance sampling scheme which incorporates our knowledge of the asymptotic form for the sum distribution. We compare this IS estimator against the Asmussen–Kroese estimator, exponential tilting, hazard-rate twisting, the cross-entropy method, and MCMC methods. It considers sums of independent variables, but in future work we expect to be able to incorporate copulas

which are asymptotically independent. We designed it as a rare-event estimator but with a focus on intermediate levels of rareness.

Lastly, Chapter 6 constructs a rare-event estimator for the probability of a union of dependent events. The central example given is calculating the probability that the maximum of a random vector exceeds a large threshold. We propose a flexible series of estimators for such probabilities, and describe variance reduction schemes applied to the proposed estimators. We derive efficiency results of the estimators in rare-event settings, in particular those associated with extremes. Finally, we examine the performance of our estimators in a numerical example.

## Chapter 2 Authorship Statement

**Citation:** Patrick J. Laub, Søren Asmussen, Jens Ledet Jensen, Leonardo Rojas-Nandayapa (2015), *Approximating the Laplace transform of the sum of dependent lognormals*, Advances in Applied Probability

The authors of this paper equally contributed to the following tasks:

1. conception and design of the project;
2. mathematical arguments, and interpretation of the results;
3. writing the publication.

In addition to this, I completed the majority of the computational work and of the editing (e.g. checking grammar and typographical details).

# Chapter 2

# Approximating the Laplace transform of the sum of dependent lognormals

## 2.1 Introduction

The lognormal distribution arises in a wide variety of disciplines such as engineering, economics, insurance and finance, and is often employed in modelling across the sciences [7, 56, 65, 104, 119]. It has a natural multivariate version, namely $(\mathrm{e}^{X_1}, \ldots, \mathrm{e}^{X_n}) \sim \mathsf{Lognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ when $(X_1, \ldots, X_n) \sim \mathsf{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We consider sums of lognormal random variables, $S_n := \mathrm{e}^{X_1} + \cdots + \mathrm{e}^{X_n}$, where the summands exhibit dependence ($\boldsymbol{\Sigma}$ is non-diagonal), using the notation that $S_n \sim \mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Such sums have many challenging properties. In particular, there are no closed-form expressions for the density $f(x)$ or Laplace transform $\mathscr{L}(\theta)$ of $S_n$.

Models using sums of dependent lognormals are widely applicable, though they are particularly important in telecommunications and finance [64, 65]. Indeed, many of the approximations for the Laplace transform of sums of independent lognormals originated from the wireless communications community [27]. This reflects the significance of the SLN distribution within many models, and also that the Laplace transform is of intrinsic

interest (engineers frequently work in the Laplace domain). In finance, the value of a portfolio (e.g. a collection of stocks) is SLN-distributed when using the assumptions of the common Black–Scholes framework. Thus the SLN distribution is central to the pricing of certain options (e.g., Asian and basket) [126]. Also, financial risk managers require estimates of $f(x)$ across $x \in (0, \mathbb{E}[S_n])$ to estimate risk measures such as value-at-risk or expected shortfall. Estimation of this kind has long been a legal requirement for many large banks, due to the Basel series of regulations (particularly, Basel II and Basel III), so in this context approximating $\mathscr{L}(\theta)$ is useful as a vehicle for computing the density $f(x)$ or the cumulative distribution function (cdf). These issues are carefully explained in [62], [70], and the new Chapter 1 in the recently revised volume of McNeil et al. [123]. Comprehensive surveys of applications and numerical methods for the LN and SLN distributions are in [85, 20, 19].

There exist many approximations to the density of the SLN distribution. Many approximations work from the premise [28] that a sum $S_n$ of lognormals can be accurately approximated by a single lognormal $L \sim \mathsf{Lognormal}(\mu_L, \sigma_L^2)$. We refer to this approach as the *SLN $\approx$ LN approximation*. Some well-known SLN $\approx$ LN approximations are the Fenton–Wilkinson [71] and Schwartz–Yeh [153] approaches. These were originally specified for sums of *independent* lognormals, but have since been generalised to the dependent case [4]. A more recent procedure (for the independent case) is the minimax approximation of Beaulieu and Xie [29], calculating the values of $\mu_L$ and $\sigma_L$ which minimise the maximum difference between the densities of $S_n$ and $L$. However, [29] concludes that the approach is inaccurate in large dimensions or when the $X_i$ have significantly different means or standard deviations. Finally, Beaulieu and Rajwani [28] describe a family of functions which mimic the characteristics of the SLN distribution function (in the independent case) with some success, i.e., high accuracy and closed-form expressions.

Another related avenue of research focuses on the asymptotic behaviour of $f(x)$ in the tails. First, Asmussen and Rojas-Nandayapa [23] characterised the right-tail asymptotics. Next, Gao et al. [77] gave the asymptotic form of the left tail for $n = 2$. Gulisashvili and Tankov [85] then provided the left-tail asymptotics for linear combinations of $n \geq 2$ lognormal variables. Yet these asymptotic forms cannot be used to approximate $f(x)$ with precision; to quote [85, p. 29], "these formulas are not valid for $x \geq 1$ and in practice have very poor accuracy unless $x$ is much smaller than one". Similar numerical experience

is reported in Asmussen et al. [19].

The approach taken here is via the Laplace transform. Accurate estimates for the Laplace transform can be numerically inverted to supply accurate density estimates. Asmussen et al. [20, 19] outline a framework to estimate $\mathscr{L}(\theta)$ for $n = 1$ using a modified saddlepoint approximation. In their work, the transform is decomposed into $\mathscr{L}(\theta) = \widetilde{\mathscr{L}}(\theta)I(\theta)$, where $\widetilde{\mathscr{L}}(\theta)$ has an explicit form and an efficient Monte Carlo estimator is given for $I(\theta)$.

This chapter generalises the approach of [20, 19] to arbitrary $n$ and dependence. The defining integral for the Laplace transform of $S_n$ is

$$\mathscr{L}(\theta) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \int_{\mathbb{R}^n} \exp\Big\{-\theta \sum_{i=1}^{n} \mathrm{e}^{\mu_i} \mathrm{e}^{x_i} - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}\Big\}\, \mathrm{d}\boldsymbol{x} \tag{2.1}$$

where $\boldsymbol{D} := \boldsymbol{\Sigma}^{-1}$ (assuming $\boldsymbol{\Sigma}$ to be positive definite so $\boldsymbol{D}$ is well defined). Write the integrand as $\exp\{-h_\theta(\boldsymbol{x})\}$. The idea is then to provide an approximation $\widetilde{\mathscr{L}}(\theta)$ by replacing $h_\theta(\boldsymbol{x})$ by a second-order Taylor expansion around its minimiser $\boldsymbol{x}^*$. Whereas the minimiser $x^*$ has a simple expression in terms of the Lambert W function when $n = 1$, as in [20, 19], the situation is much more complex when $n > 1$. As one of our main results we give a limit result for $\boldsymbol{x}^*$ as $\theta \to \infty$. Further, it is shown that the remainder $I(\theta)$ in the representation $\mathscr{L}(\theta) = \widetilde{\mathscr{L}}(\theta)I(\theta)$ goes to 1, a discussion of efficient Monte Carlo estimators of $I(\theta)$ follows, and numerical results showing the errors of our $\mathscr{L}(\theta)$ and (numerically inverted) $f(x)$ estimators are given.

## 2.2 Approximating the Laplace transform

Although the definition (2.1) makes sense for all $\theta \in \mathbb{C}$ with $\Re(\theta) > 0$ (we denote this set as $\mathbb{C}_+$), we will restrict the focus to $\theta \in (0, \infty)$. Of particular interest are the terms in the exponent, which in vector form (see Remark 2.1 below) are

$$h_\theta(\boldsymbol{x}) := \theta(\mathbf{e}^{\boldsymbol{\mu}})^\top \mathbf{e}^{\boldsymbol{x}} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}.$$

An approximation of simple form to $\mathscr{L}(\theta)$—written as $\widetilde{\mathscr{L}}(\theta)$—is available if $h_\theta(\boldsymbol{x})$ is replaced by a second-order Taylor expansion. The expansion is given in the proposition below.

**Remark 2.1.** *On vector notation.* All vectors are considered column vectors. Functions applied element-wise to vectors are written in boldface, such as $\mathbf{e}^{\boldsymbol{x}} := (\mathrm{e}^{x_1}, \ldots, \mathrm{e}^{x_n})^\top$ and $\mathbf{log}\, \boldsymbol{x} := (\log x_1, \ldots, \log x_n)^\top$. If a vector is to be raised element-wise to a common power, then the power will be boldface, as in $\boldsymbol{x}^{\boldsymbol{k}} := (x_1^k, \ldots, x_n^k)^\top$. The notation $\boldsymbol{x} \circ \boldsymbol{y}$ denotes element-wise multiplication of vectors. The $\mathrm{diag}(\cdot)$ function converts vectors to matrices and vice versa, like the MATLAB function. $\diamond$

**Proposition 2.2.** *The second-order Taylor expansion of $h_\theta(\boldsymbol{x})$ about its unique minimiser $\boldsymbol{x}^*$ is*

$$-\Big(\mathbf{1} - \frac{1}{2}\boldsymbol{x}^*\Big)^\top \boldsymbol{D}\boldsymbol{x}^* + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top(\boldsymbol{\Lambda} + \boldsymbol{D})(\boldsymbol{x} - \boldsymbol{x}^*)$$

*where $\boldsymbol{\Lambda} := \theta\mathrm{diag}(\mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*})$.*

*Proof.* As $h_\theta(\boldsymbol{x})$ is strictly convex, a unique minimum exists. Since $\nabla h_\theta(\boldsymbol{x}^*) = \mathbf{0}$, the linear term vanishes in the Taylor expansion, so we have

$$h_\theta(\boldsymbol{x}) \approx h_\theta(\boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top \boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}^*)$$

where $\boldsymbol{H}$ is defined as the Hessian $\partial^2 h_\theta(\boldsymbol{x})/(\partial x_i\, \partial x_j)$ evaluated at $\boldsymbol{x}^*$. To find the value of $\boldsymbol{H}$, we just take derivatives:

$$\nabla h_\theta(\boldsymbol{x}) = \theta\mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}} + \boldsymbol{D}\boldsymbol{x}, \quad \boldsymbol{H} = \boldsymbol{\Lambda} + \boldsymbol{D}.$$

Since $\boldsymbol{\Lambda}$ and $\boldsymbol{D}$ are both positive definite, so is $\boldsymbol{H}$. Also, $\nabla h_\theta(\boldsymbol{x}^*) = \mathbf{0}$ gives

$$-\theta\mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*} = \boldsymbol{D}\boldsymbol{x}^* \text{ which implies } -\theta(\mathbf{e}^{\boldsymbol{\mu}})^\top \mathbf{e}^{\boldsymbol{x}^*} = \mathbf{1}^\top \boldsymbol{D}\boldsymbol{x}^*. \tag{2.2}$$

Therefore the expansion becomes

$$\begin{aligned} h_\theta(\boldsymbol{x}) &\approx -\mathbf{1}^\top \boldsymbol{D}\boldsymbol{x}^* + \frac{1}{2}(\boldsymbol{x}^*)^\top \boldsymbol{D}\boldsymbol{x}^* + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top(\boldsymbol{\Lambda} + \boldsymbol{D})(\boldsymbol{x} - \boldsymbol{x}^*) \\ &= -\Big(\mathbf{1} - \frac{1}{2}\boldsymbol{x}^*\Big)^\top \boldsymbol{D}\boldsymbol{x}^* + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top(\boldsymbol{\Lambda} + \boldsymbol{D})(\boldsymbol{x} - \boldsymbol{x}^*). \end{aligned}$$

□ □

This expansion allows $\mathscr{L}(\theta)$ to be approximated as a constant factor $\exp\{-h_\theta(\boldsymbol{x}^*)\}$ times the integral over a normal density (with inverse covariance $\boldsymbol{\Lambda} + \boldsymbol{D}$), which leads to

$$\mathscr{L}(\theta) \approx \widetilde{\mathscr{L}}(\theta) := \frac{1}{\sqrt{\det(\boldsymbol{\Sigma}\boldsymbol{H})}} \exp\left\{\left(\mathbf{1} - \frac{1}{2}\boldsymbol{x}^*\right)^\top \boldsymbol{D}\boldsymbol{x}^*\right\}.$$

We need a suitable error or correction term in order to assess the accuracy of this approximation, so we will decompose the original integral (2.1) into $\mathscr{L}(\theta) = \widetilde{\mathscr{L}}(\theta)I(\theta)$. In the integral of (2.1) change variables such that $\boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{H}^{-1/2}\boldsymbol{y}$. Then by applying (2.2), multiplying by $\exp\{\mathbf{1}^\top \boldsymbol{D}\boldsymbol{x}^* - \mathbf{1}^\top \boldsymbol{D}\boldsymbol{x}^*\}$, and rearranging, we arrive at

$$\begin{aligned}\mathscr{L}(\theta) &= \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma}\boldsymbol{H})}} \int_{\mathbb{R}^n} \exp\Big\{-\theta(\mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*})^\top \mathbf{e}^{\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y}} \\ &\qquad - \frac{1}{2}(\boldsymbol{x}^* + \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y})^\top \boldsymbol{D}(\boldsymbol{x}^* + \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y})\Big\}\,\mathrm{d}\boldsymbol{y} \\ &= \widetilde{\mathscr{L}}(\theta)I(\theta)\end{aligned}$$

where

$$I(\theta) := \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n}} \exp\Big\{(\boldsymbol{x}^*)^\top \boldsymbol{D}\Big(\mathbf{e}^{\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y}} - \mathbf{1} - \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y}\Big) - \frac{1}{2}\boldsymbol{y}^\top(\boldsymbol{\Sigma}\boldsymbol{H})^{-1}\boldsymbol{y}\Big\}\,\mathrm{d}\boldsymbol{y}. \tag{2.3}$$

This equation can be rewritten in ways more convenient for Monte Carlo estimation.

**Proposition 2.3.** *We have that*

$$I(\theta) = \mathbb{E}\big[g(\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{Z})\big] = \sqrt{\det(\boldsymbol{\Sigma}\boldsymbol{H})}\ \mathbb{E}\big[v(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z})\big] \tag{2.4}$$

*where*

$$\begin{aligned} g(\boldsymbol{u}) &:= \exp\left\{(\boldsymbol{x}^*)^\top \boldsymbol{D}(\mathbf{e}^{\boldsymbol{u}} - \mathbf{1} - \boldsymbol{u}) + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{u}\right\}, \\ v(\boldsymbol{u}) &:= \exp\Big\{(\boldsymbol{x}^*)^\top \boldsymbol{D}(\mathbf{e}^{\boldsymbol{u}} - \mathbf{1} - \boldsymbol{u})\Big\}, \end{aligned}$$

*and* $\boldsymbol{Z} \sim \mathsf{Normal}(\mathbf{0}, \boldsymbol{I})$.

*Proof.* To show that $I(\theta)$ can be written as the first expectation in (2.4), use $\boldsymbol{H} = \boldsymbol{\Lambda} + \boldsymbol{D}$, then add and subtract a term, to get

$$(\boldsymbol{\Sigma H})^{-1} = [\boldsymbol{\Sigma}(\boldsymbol{D} + \boldsymbol{\Lambda})]^{-1} = (\boldsymbol{I} + \boldsymbol{\Sigma\Lambda})^{-1}\boldsymbol{I} \pm (\boldsymbol{I} + \boldsymbol{\Sigma\Lambda})^{-1}(\boldsymbol{\Sigma\Lambda}) = \boldsymbol{I} - \boldsymbol{H}^{-1}\boldsymbol{\Lambda}$$

and substitute this into the $-\frac{1}{2}\boldsymbol{y}^\top(\boldsymbol{\Sigma H})^{-1}\boldsymbol{y}$ term in (2.3).

To prove $I(\theta)$ equals the second expectation of (2.4), change variables in (2.3) so that $\boldsymbol{y} = (\boldsymbol{\Sigma H})^{1/2}\boldsymbol{z}$, giving

$$I(\theta) = \sqrt{\det(\boldsymbol{\Sigma H})} \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n}} \exp\Big\{(\boldsymbol{x}^*)^\top \boldsymbol{D}\big(\mathrm{e}^{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z}} - \boldsymbol{1} - \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z}\big) - \frac{1}{2}\boldsymbol{z}^\top \boldsymbol{I}\boldsymbol{z}\Big\}\,\mathrm{d}\boldsymbol{z}. \tag{2.5}$$

□ □

**Remark 2.4.** When $n = 1$, $\boldsymbol{\Sigma} = \sigma^2$ and $\mu = 0$, (2.5) becomes

$$I(\theta) = \sqrt{1 + \theta\sigma^2\mathrm{e}^{x^*}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\Big\{\frac{x^*}{\sigma^2}(\mathrm{e}^{\sigma z} - 1 - \sigma z) - \frac{1}{2}z^2\Big\}\,\mathrm{d}z.$$

This can be simplified using the *Lambert W* function, denoted $\mathcal{W}$, which is defined [53] as the solution to the equation $\mathcal{W}(z)\mathrm{e}^{\mathcal{W}(z)} = z$. With this we have $x^* = -\mathcal{W}(\theta\sigma^2)$. Also, we can manipulate $\sqrt{1 + \theta\sigma^2\mathrm{e}^{x^*}} = \sqrt{1 - x^*} = \sqrt{1 + \mathcal{W}(\theta\sigma^2)}$, so $I(\theta)$ becomes

$$I(\theta) = \sqrt{1 + \mathcal{W}(\theta\sigma^2)} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\Big\{-\frac{\mathcal{W}(\theta\sigma^2)}{\sigma^2}(\mathrm{e}^{\sigma z} - 1 - \sigma z) - \frac{1}{2}z^2\Big\}\,\mathrm{d}z,$$

which coincides with the original result of [20] equation (2.3). ◇

**Remark 2.5.** The simplistic representation above hides a more subtle understanding. The following is based off Senaratne and Tellambura [154] who considered numerically evaluating $\mathscr{L}(\theta)$ for $n = 1$. Consider $\theta \in \mathbb{C}_+$, and note that $\exp\{h_\theta(\cdot)\}$ is analytic everywhere (that is, it satisfies the Cauchy–Riemann conditions). Then consider contour integrals of the form

$$\int_{\mathcal{C}} \exp\{h_\theta(\boldsymbol{z})\}\,\mathrm{d}\boldsymbol{z}\,,$$

where $\mathcal{C}$ is some contour. Cauchy's integral theorem implies that $\mathcal{C}$ can be deformed continuously through the domain of analyticity (here, $\mathbb{C}_+$) in any way while keeping the

endpoints constant without affecting the result.

A preferable contour for numerical evaluation is one which clusters most of the integrand's mass together, the *steepest descent contour* $\mathcal{C}^*$, which also passes through the point which maximises the integrand, the *saddlepoint* $\boldsymbol{z}^*$. Analyticity implies orthogonality of $u(\boldsymbol{z}) := \Re(h_\theta(\boldsymbol{z}))$ and $v(\boldsymbol{z}) := \Im(h_\theta(\boldsymbol{z}))$, so the steepest descent contour is equivalent to a contour where $v$ is constant. That is,

$$\boldsymbol{z}^* = \arg\max_{\boldsymbol{z}\in\mathbb{C}} \exp\{h_\theta(\boldsymbol{z})\}\,, \text{ and } \mathcal{C}^* = \{\boldsymbol{z} = \boldsymbol{x} + \mathrm{i}\boldsymbol{y} \,|\, v(\boldsymbol{z}) = v(\boldsymbol{z}^*)\}\,. \tag{2.6}$$

Say that $\boldsymbol{y}(\boldsymbol{x})$ is the solution for $\boldsymbol{y}$ given an $\boldsymbol{x}$ from (2.6), and for $\boldsymbol{z} = \boldsymbol{x} + \mathrm{i}\boldsymbol{y}$ then $\mathrm{d}\boldsymbol{z} = \mathrm{d}\boldsymbol{x} + \mathrm{i}\,\mathrm{d}\boldsymbol{y} = (1 + \mathrm{i}\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}\boldsymbol{x}})\,\mathrm{d}\boldsymbol{x}$. Then, in theory, we could evaluate

$$\mathscr{L}(\theta) = \int_{\mathcal{C}^*} \exp\{h_\theta(\boldsymbol{z})\}\,\mathrm{d}\boldsymbol{z} = \int_{\mathbb{R}^n} \exp\{u(\boldsymbol{x} + \mathrm{i}\boldsymbol{y}(\boldsymbol{x})) + \mathrm{i}v(\boldsymbol{x} + \mathrm{i}\boldsymbol{y}(\boldsymbol{x}))\}\Big(1 + \mathrm{i}\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}\boldsymbol{x}}\Big)\,\mathrm{d}\boldsymbol{x}$$

which by the definition of $\mathcal{C}^*$ simplifies to

$$\mathscr{L}(\theta) = \exp\{\mathrm{i}v(\boldsymbol{z}^*)\}\int_{\mathbb{R}^n} \exp\{u(\boldsymbol{x} + \mathrm{i}\boldsymbol{y}(\boldsymbol{x}))\}\Big(1 + \mathrm{i}\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}\boldsymbol{x}}\Big)\,\mathrm{d}\boldsymbol{x}\,. \tag{2.7}$$

For $\theta \in \mathbb{R}$ then $\mathcal{C}^*$ simply returns to $\mathbb{R}^n$ and $\boldsymbol{z}^* \to \boldsymbol{x}^*$ above. For $\Im(\theta) \neq 0$ then (2.7) can no longer be simplified by transformation into a Gaussian integral, and $\boldsymbol{z}^*$ is no longer unique. Therefore, we continue to restrict our attention to the $\theta \in \mathbb{R}$ case. ◇

## 2.3 Asymptotic behaviour of the minimiser $\boldsymbol{x}^*$

We first introduce some notation. For a matrix $\boldsymbol{X}$, we write $\boldsymbol{X}_{i,\bullet}$ and $\boldsymbol{X}_{\bullet,i}$ for the $i$-th row and column. Denote the row sums of $\boldsymbol{D}$ as $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$, that is, $a_i = \boldsymbol{D}_{i,\bullet}\boldsymbol{1}$. For sets of indices $\Omega_1$ and $\Omega_2$, then $\boldsymbol{X}_{\Omega_1,\Omega_2}$ denotes the submatrix of $\boldsymbol{X}$ containing row/column pairs in $\{(u, v) : u \in \Omega_1, v \in \Omega_2\}$. A shorthand is used for iterated logarithms: $\log_1\theta := \log\theta$ and $\log_n\theta := \log\log_{n-1}\theta$ for $n \geq 2$ (note that $\log_k\theta$ is undefined for small or negative $\theta$, but this is no problem as we are considering the case $\theta \to \infty$).

The approach taken to find $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)^\top$ is to set the gradient of $h_\theta(\boldsymbol{x})$ to $\boldsymbol{0}$, that

is, to solve

$$\theta \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*} + \boldsymbol{D}\boldsymbol{x}^* = \mathbf{0}. \tag{2.8}$$

We will show that the asymptotics of the $x_i^*$ are of the form

$$x_i^* = \sum_{j=1}^{n} \beta_{i,j} \log_j \theta - \mu_i + c_i + r_i(\theta) \tag{2.9}$$

for some $\boldsymbol{\beta} = (\beta_{i,j}) \in \mathbb{R}^{n\times n}$, $\boldsymbol{c} = (c_1, \ldots, c_n)^\top \in \mathbb{R}^n$ and $\boldsymbol{r}(\theta) = (r_1(\theta), \ldots, r_n(\theta))^\top$ where each $r_i(\theta) = \mathrm{o}(1)$. Before giving the general result, we consider the special case where all $a_i > 0$ since this result and its proof are much simpler.

**Proposition 2.6.** *If all row sums of $\boldsymbol{D}$ are positive then the minimiser $\boldsymbol{x}^*$ takes the form*

$$x_i^* = -\log\theta + \log_2\theta - \mu_i + \log a_i + r_i(\theta) \tag{2.10}$$

*where $r_i(\theta) = \mathcal{O}(\log_2\theta/\log\theta) = \mathrm{o}(1)$ for $1 \le i \le n$, as $\theta \to \infty$.*

*Proof.* Inserting (2.10) in (2.8) we find

$$\theta \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*} + \boldsymbol{D}\boldsymbol{x}^* = (\boldsymbol{a}\log\theta) \circ \mathbf{e}^{\boldsymbol{r}(\theta)} - \boldsymbol{a}\log\theta + \boldsymbol{a}\log_2\theta - \boldsymbol{D}\boldsymbol{\mu} + \boldsymbol{D}\mathbf{log}\,\boldsymbol{a} + \boldsymbol{D}\boldsymbol{r}(\theta) = \mathbf{0}.$$

Looking at these equations we see that we must have

$$\limsup_{\theta} \max_i r_i(\theta) = \liminf_{\theta} \min_i r_i(\theta) = 0,$$

and to remove the $\log_2\theta$ term the main term of $r_i(\theta)$ has to be $-\log_2\theta/\log\theta$. This gives the result of the proposition. □ □

In the general case where some $a_i \le 0$, the asymptotic form of $\boldsymbol{x}^*$ is different from (2.10) and its derivation is much more intricate.

**Theorem 2.7.** *There exists a partition of $\{1, \ldots, n\}$ into $\mathcal{F}_+$ and $\mathcal{F}_-$ such that for $i \in \mathcal{F}_+$,*

$$x_i^* = -\log\theta + \log_{k_i}\theta - \mu_i + c_i + \mathrm{o}(1)$$

*for some* $1 < k_i \leq n$. *All* $x_i^*$ *in* $\mathcal{F}_-$ *follow the general form of* (2.9). *In more detail, there exists a partition of* $\mathcal{F}_-$ *into* $\mathcal{F}_-(1)$ *and* $\mathcal{F}_- \setminus \mathcal{F}_-(1)$, *such that if* $i \in \mathcal{F}_-(1)$ *then* $\beta_{i,1} < -1$, *and if* $i \in \mathcal{F}_- \setminus \mathcal{F}_-(1)$ *then*

$$\beta_{i,1} = -1,\ \beta_{i,2} = \cdots = \beta_{i,k_i-1} = 0,\ \beta_{i,k_i} < 0$$

*for some* $1 < k_i \leq n$. *Finally we have, writing subscripts* $+$ *and* $-$ *for* $\mathcal{F}_+$ *and* $\mathcal{F}_-$, *that* $\boldsymbol{x}_- = \boldsymbol{C}\boldsymbol{x}_+ + \mathrm{o}(1)$ *where* $\boldsymbol{C} = -\boldsymbol{D}_{-,-}^{-1}\boldsymbol{D}_{-,+}$. *The sets* $\mathcal{F}_+$, $\mathcal{F}_-$, $\mathcal{F}_-(1)$ *and the constants* $\beta_{i,j}$, $c_i$, $k_i$ *are determined by Algorithm 2.8 below.*

See Remark 2.12 for some further remarks on the role of the signs of the row sums.

**Algorithm 2.8.**

1. *Let* $\boldsymbol{\beta}_{\bullet,1}$ *be the value of* $\boldsymbol{w}$ *that minimises* $\boldsymbol{w}^\top \boldsymbol{D}\boldsymbol{w}$ *over the set* $\{\boldsymbol{w} : w_i \leq -1\}$. *It will be proved in the appendix that the solution has* $\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,1} \leq 0$ *when* $\beta_{i,1} = -1$ *and* $\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,1} = 0$ *when* $\beta_{i,1} < -1$. *Accordingly, we can partition* $\{1, \ldots, n\}$ *into the disjoint sets*

$$\mathcal{F}_+(1) = \emptyset, \quad \mathcal{F}_*(1) = \{i : \boldsymbol{D}_{j,\bullet}\boldsymbol{\beta}_{\bullet,1} < 0\},$$
$$\mathcal{F}_0(1) = \{i : \beta_{i,1} = -1, \boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,1} = 0\}, \quad \mathcal{F}_-(1) = \{i : \beta_{i,1} < -1\}.$$

2. *For* $k = 2, \ldots, n$ *recursively calculate* $\boldsymbol{\beta}_{\bullet,k}$ *as the value of* $\boldsymbol{w}$ *that minimises* $\boldsymbol{w}^\top \boldsymbol{D}\boldsymbol{w}$ *whilst satisfying*

$$w_i = 0 \text{ for } i \in \mathcal{F}_+(k-1), \quad w_i = 1 \text{ for } i \in \mathcal{F}_*(k-1),$$
$$w_i \leq 0 \text{ for } i \in \mathcal{F}_0(k-1), \quad \boldsymbol{D}_{i,\bullet}\boldsymbol{w} = 0 \text{ for } i \in \mathcal{F}_-(k-1).$$

*It will be proved in the appendix that the solution has* $\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,k} \leq 0$ *for* $i \in \mathcal{F}_0(k-1)$, $\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,k} = 0$ *when* $\beta_{i,k} < 0$ *for* $i \in \mathcal{F}_0(k-1)$, *and at least one element of* $\mathcal{F}_0(k-1)$

*has* $\boldsymbol{D}_{i,\bullet}\beta_{\bullet,k} < 0$. *This allows us to create a new partition by*

$$\begin{aligned}
\mathcal{F}_+(k) &= \mathcal{F}_+(k-1) \cup \mathcal{F}_*(k-1),\\
\mathcal{F}_*(k) &= \{i \in \mathcal{F}_0(k-1) : \beta_{i,k} = 0, \boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,k} < 0\},\\
\mathcal{F}_0(k) &= \{i \in \mathcal{F}_0(k-1) : \beta_{i,k} = 0, \boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,k} = 0\},\\
\mathcal{F}_-(k) &= \mathcal{F}_-(k-1) \cup \{i \in \mathcal{F}_0(k-1) : \beta_{i,k} < 0\}.
\end{aligned}$$

*Terminate the loop early if* $\mathcal{F}_0(k-1) = \emptyset$.

3. *Say* $\mathcal{F}_+ = \mathcal{F}_+(k)$ *and* $\mathcal{F}_- = \mathcal{F}_-(k)$. *For each* $i \in \mathcal{F}_+$, *let* $\ell_i$ *be the index of the first element of* $\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}$ *which is negative, and we have* $c_i = \log(-\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,\ell_i})$. *Determine the remaining elements (using the same subscript shorthand introduced above) by*

$$\boldsymbol{c}_- = -\boldsymbol{D}_{-,-}^{-1}\boldsymbol{D}_{-,+}(\boldsymbol{c}_+ - \boldsymbol{\mu}_+) + \boldsymbol{\mu}_-. \tag{2.11}$$

*Proof of Theorem* 2.7. We propose a solution of the form (2.9) and show that when the $\beta_{i,j}$ are constructed from Algorithm 2.8, the remainder term $r_i$ is o(1).

The construction allows us to draw the following conclusions for the $x_i^*$. Let $\mathcal{F}_+$ and $\mathcal{F}_-$ be the sets as defined in Step 3 above. Consider individually the indices which terminated in the $\mathcal{F}_+$ and in the $\mathcal{F}_-$ sets. In the first case, there exists a $k_i$ with $1 < k_i \le n$ such that

$$\beta_{i,j} = \begin{cases} -1, & j = 1,\\ 1, & j = k_i,\\ 0, & \text{otherwise,}\end{cases} \qquad \text{and} \quad \boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,j} = \begin{cases} 0, & 1 \le j < k_i - 1,\\ < 0, & j = k_i - 1.\end{cases}$$

Insertion in (2.8) gives

$$\begin{aligned}
0 &= \theta \mathrm{e}^{\mu_i + x_i^*} + \boldsymbol{D}_{i,\bullet}\boldsymbol{x}^*\\
&= -\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,k_i-1}\mathrm{e}^{r_i(\theta)}\log_{k_i-1}\theta + \boldsymbol{D}_{i,\bullet}\left(\sum_{j=k_i-1}^{n}\boldsymbol{\beta}_{\bullet,j}\log_j\theta - \boldsymbol{\mu} + \boldsymbol{c} + \boldsymbol{r}(\theta)\right),
\end{aligned}$$

showing that the remainder is o(1).

In the second case, with $i \in \mathcal{F}_+$,

$$\beta_{i,1} < -1 \quad \text{and} \quad \boldsymbol{D}_{i,\boldsymbol{\cdot}}\boldsymbol{\beta}_{\boldsymbol{\cdot},j} = 0,\ 1 \le j \le n,$$

or there exists $1 < k_i \le n$ such that

$$\beta_{i,j} = \begin{cases} -1, & j = 1, \\ 0, & 2 \le j < k_i, \\ < 0, & j = k_i, \end{cases} \quad \text{and} \quad \boldsymbol{D}_{i,\boldsymbol{\cdot}}\boldsymbol{\beta}_{\boldsymbol{\cdot},j} = 0 \text{ for } 1 \le j \le n.$$

For this case we find $\theta \mathrm{e}^{\mu_i + x_i^*} + \boldsymbol{D}_{i,\boldsymbol{\cdot}}\boldsymbol{x}^* = \mathrm{o}(1) + \boldsymbol{D}_{i,\boldsymbol{\cdot}}\boldsymbol{r}(\theta)$, again showing that the remainder is o(1). Lastly, to show $\boldsymbol{x}_-$ in terms of $\boldsymbol{x}_+$, consider $\theta \mathbf{e}^{\boldsymbol{\mu}_- + \boldsymbol{x}_-^*} + \boldsymbol{D}_{-,+}\boldsymbol{x}_+ + \boldsymbol{D}_{-,-}\boldsymbol{x}_- = \mathbf{0}$. As $\theta \mathbf{e}^{\boldsymbol{\mu}_- + \boldsymbol{x}_-^*} = \mathrm{o}(1)$, we see that $\boldsymbol{x}_- = -\boldsymbol{D}_{-,-}^{-1}\boldsymbol{D}_{-,+}\boldsymbol{x}_+ + \mathrm{o}(1) = \boldsymbol{C}\boldsymbol{x}_+ + \mathrm{o}(1)$. □

In some cases above, we have been able to write the constant $c_i$ as an expression involving $\boldsymbol{D}$ and $\boldsymbol{\mu}$. For example, in Proposition 2.6 we have $c_i = \log a_i$, and in Theorem 2.7 (2.11) gives the value of $c_i$ for $i \in \mathcal{F}_-$. We can show a similar result in the general case for all $i \in \mathcal{F}_*(1)$, that is, for all $i$ where $x_i^* = -\log\theta + \log_2\theta - \mu_i + c_i + \mathrm{o}(1)$.

Say $\mathcal{F}_* := \mathcal{F}_*(1)$ and $\mathcal{F}_\sim := \mathcal{F}_*^{\mathsf{c}}$; in the subscripts below, $*$ and $\sim$ refer to these sets. Since $\boldsymbol{D}$ is regular, so is $\boldsymbol{D}_{\sim,\sim}$. Say that $\overline{\boldsymbol{D}} := \boldsymbol{D}_{*,*} - \boldsymbol{D}_{*,\sim}\boldsymbol{D}_{\sim,\sim}^{-1}\boldsymbol{D}_{\sim,*}$, and denote the corresponding row sums by $\overline{\boldsymbol{a}} = (\overline{a}_i, i \in \mathcal{F}_*)$.

**Corollary 2.9.** *For all $i \in \mathcal{F}_*$*

$$x_i^* = -\log\theta + \log_2\theta - \mu_i + \log\overline{a}_i + r_i(\theta)$$

*where $r_i(\theta) = \mathrm{o}(1)$ and $\overline{a}_i > 0$ as $\theta \to \infty$.*

*Proof.* Let $\boldsymbol{b} = -\boldsymbol{\beta}_{\boldsymbol{\cdot},1}$. We have

$$b_i = \begin{cases} 1, & i \in \mathcal{F}_*(1) \cup \mathcal{F}_0(1), \\ > 1, & i \in \mathcal{F}_-(1), \end{cases} \qquad \boldsymbol{D}_{i,\boldsymbol{\cdot}}\boldsymbol{b} = \begin{cases} \mathrm{e}^{c_i}, & i \in \mathcal{F}_*(1) = \mathcal{F}_*, \\ 0, & i \in \mathcal{F}_0(1) \cup \mathcal{F}_-(1) = \mathcal{F}_\sim. \end{cases}$$

Split $\boldsymbol{D}$ according to indices in $\mathcal{F}_*$ and $\mathcal{F}_\sim$, then

$$\boldsymbol{D}_{\sim,*}\boldsymbol{b}_* + \boldsymbol{D}_{\sim,\sim}\boldsymbol{b}_\sim = \mathbf{0} \quad \text{and} \quad \boldsymbol{D}_{*,*}\boldsymbol{b}_* + \boldsymbol{D}_{*,\sim}\boldsymbol{b}_\sim = \mathbf{e}^{\boldsymbol{c}_*} > \mathbf{0}.$$

The first equation gives $\boldsymbol{b}_\sim = -\boldsymbol{D}_{\sim,\sim}^{-1}\boldsymbol{D}_{\sim,*}\boldsymbol{b}_*$, and this with the second equation shows that $\overline{\boldsymbol{D}}\boldsymbol{b}_* = \overline{\boldsymbol{D}}\mathbf{1} = \overline{\boldsymbol{a}} = \mathbf{e}^{\boldsymbol{c}_*} > \mathbf{0}$; thus $\overline{\boldsymbol{D}}$ has all row sums positive and $\boldsymbol{c}_* = \mathbf{log}\,(\overline{\boldsymbol{D}}\boldsymbol{b}_*) = \mathbf{log}\,\overline{\boldsymbol{a}}$. □

There are some simple forms of $\boldsymbol{\Sigma}$ which fall into the case where all $a_i > 0$. These include the case where all diagonal elements of $\boldsymbol{\Sigma}$ are identical, and all non-diagonal elements are identical. Note, by positive definiteness of $\boldsymbol{\Sigma}$ we must have at least one row-sum of $\boldsymbol{D}$ positive. Also, if $X_1, \ldots, X_n$ is an AR(1) process, the resulting covariance matrix will have all $a_i > 0$. Meanwhile, cases where there exist $a_i \leq 0$ are not difficult to find. For the case $n = 2$ with variances $\sigma_1^2 \leq \sigma_2^2$ and correlation $\rho$, a simple calculation gives that both row sums are positive when $\rho < \sigma_1/\sigma_2$, and one is negative when $\rho > \sigma_1/\sigma_2$ (see Gao et al. [77] for the expansion of $f(x)$ as $x \downarrow 0$ for these cases). We now list a couple of examples of asymptotic forms of $\boldsymbol{x}^*$ for specific $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which have some non-positive row sums of $\boldsymbol{\Sigma}^{-1}$.

**Example 2.10.** Consider $\boldsymbol{\mu} = (-10, 0, 10)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 1 & 2 \\ 1 & 3 & 4 \\ 2 & 4 & 10 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 14 & -2 & -2 \\ -2 & 1 & 0 \\ -2 & 0 & 0.5 \end{pmatrix}.$$

Implementing the algorithm gives that

$$\begin{aligned} x_1^* &= -\log\theta + \log_2\theta + (10 + \log 2) + \mathrm{o}(1), \\ x_2^* &= -2\log\theta + 2\log_2\theta + (20 + 2\log 2) + \mathrm{o}(1), \\ x_3^* &= -4\log\theta + 4\log_2\theta + (40 + 4\log 2) + \mathrm{o}(1), \end{aligned}$$

and

$$(\boldsymbol{\beta} \mid \boldsymbol{c} - \boldsymbol{\mu}) = \left(\begin{array}{ccc|c} -1 & 1 & 0 & 10.69 \\ -2 & 2 & 0 & 21.39 \\ -4 & 4 & 0 & 42.77 \end{array}\right), \quad \boldsymbol{D}(\boldsymbol{\beta} \mid \boldsymbol{c} - \boldsymbol{\mu}) = \left(\begin{array}{ccc|c} -2 & * & * & * \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

(where unimportant values of $\boldsymbol{D}(\boldsymbol{\beta} \mid \boldsymbol{c} - \boldsymbol{\mu})$ are replaced by stars). □

**Example 2.11.** Consider $\boldsymbol{\mu} = (1, 2, 3)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.4545 & 0.4545 & 0.4545 \\ 0.4545 & 1.7204 & 1.8470 \\ 0.4545 & 1.8470 & 2.9862 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 3 & -0.9 & 0.1 \\ -0.9 & 2 & -1.1 \\ 0.1 & -1.1 & 1 \end{pmatrix}.$$

Implementing the algorithm gives that

$$\begin{aligned} x_1^* &= -\log\theta + \log_2\theta - 1 + \log 2.2 + \mathrm{o}(1), \\ x_2^* &= -\log\theta + \log_3\theta - 2 + \log 0.79 + \mathrm{o}(1), \\ x_3^* &= -\log\theta - 0.1\log_2\theta + 1.1\log_3\theta - 3 + c_3 + \mathrm{o}(1), \end{aligned}$$

where $c_3 = 0.9 - 0.1\log 2.2 + 1.1\log 0.79$, and

$$(\boldsymbol{\beta} \mid \boldsymbol{c} - \boldsymbol{\mu}) = \left(\begin{array}{ccc|c} -1 & 1 & 0 & -0.2 \\ -1 & 0 & 1 & -2.2 \\ -1 & -0.1 & 1.1 & -2.4 \end{array}\right), \quad \boldsymbol{D}(\boldsymbol{\beta} \mid \boldsymbol{c} - \boldsymbol{\mu}) = \left(\begin{array}{ccc|c} -2.2 & * & * & * \\ 0 & -0.79 & * & * \\ 0 & 0 & 0 & 0 \end{array}\right).$$

□

**Remark 2.12.** The importance of the sign of the row sums of $\boldsymbol{D}$, as illustrated by Proposition 2.6, perplexed us for quite some time. However Gulisashvili and Tankov [85] describe an interesting link between the row sums and the *minimum variance portfolio*. They show that the leading asymptotic term of $\mathbb{P}(S_n < x)$ as $x \downarrow 0$ depends upon

$$\overline{\boldsymbol{w}}^\top \boldsymbol{\Sigma}\, \overline{\boldsymbol{w}} = \min_{\boldsymbol{w} \in \Delta} \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}, \text{ where } \Delta := \Big\{ \boldsymbol{w} : \sum_i w_i = 1, w_i \geq 0 \Big\}.$$

The $i$ for which $\overline{w}_i > 0$ indicate which summands in $S_n$ have the 'least variance'. These summands are asymptotically important in the left tail, as they will struggle the most to take very small values. Seen from the viewpoint of modern portfolio theory [122], the solution $\overline{\boldsymbol{w}}$ is viewed as the optimal portfolio weights to create the minimum-variance portfolio. When all $a_i > 0$ then $\overline{w}_i = a_i / \sum_{j=1}^n a_j$ which represents full diversification. However when assets become highly correlated (meaning that some $\boldsymbol{D}$ row sums are non-positive) then there exist $\overline{w}_i = 0$, i.e., some assets are ignored. Thus the asymptotics are

qualitatively different when the signs of the row sums change. The exact point where an asset's optimal weight becomes 0 occurs when $a_i = 0$, and this phase change produces a unique and convoluted asymptotic form. As $\mathscr{L}(\theta)$ as $\theta \to \infty$ is related to $\mathbb{P}(S_n < x)$ as $x \downarrow 0$, the behaviour of $\boldsymbol{x}^*$ is explained. $\diamond$

For applications we will need to find $\boldsymbol{x}^*$ for a large number of $\theta$ numerically. The results above give a sensible starting point for an iterative solver, such as Newton–Raphson.

## 2.4 Asymptotic behaviour of $I(\theta)$

In order to discuss $I(\theta)$ as $\theta \to \infty$ we will consider it in a form different from Section 2.2. Define $\boldsymbol{\sigma} := \operatorname{diag}(\boldsymbol{H})^{-1/2} \in (0,\infty)^n$ and $\boldsymbol{B} := \operatorname{diag}(\boldsymbol{\sigma})\boldsymbol{H}\operatorname{diag}(\boldsymbol{\sigma}) \in \mathbb{R}^{n\times n}$. In (2.3), substitute $\boldsymbol{H}^{-1/2}\boldsymbol{y} = \boldsymbol{\sigma} \circ \boldsymbol{z}$, so

$$I(\theta) = \int_{\mathbb{R}^n} \frac{\exp(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z})}{\sqrt{(2\pi)^n \det(\boldsymbol{B}^{-1})}} \exp\Big\{-\theta(\mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*})^\top \Big[\mathbf{e}^{\boldsymbol{\sigma}\circ\boldsymbol{z}} - \mathbf{1} - \boldsymbol{\sigma}\circ\boldsymbol{z} - \frac{1}{2}(\boldsymbol{\sigma}\circ\boldsymbol{z})^{\mathbf{2}}\Big]\Big\}\,\mathrm{d}\boldsymbol{z}. \quad (2.12)$$

The limit of this integrand is the density of a multivariate normal distribution, which when integrated is 1. To see this, consider the following. As $\theta \to \infty$ we have $\sigma_i \to 0$ or $\sigma_i \to D_{i,i}^{-1/2} > 0$, so taking $\ell \in (2,\infty)$ means

$$\theta \mathrm{e}^{\mu_i + x_i^*}\sigma_i^\ell = \theta \mathrm{e}^{\mu_i+x_i^*}(\theta \mathrm{e}^{\mu_i+x_i^*} + D_{i,i})^{-\frac{\ell}{2}} = \mathrm{o}(1). \quad (2.13)$$

Consider the second exponent of (2.12). For fixed $\boldsymbol{z}$, $\mathrm{e}^{\sigma_i z_i} - 1 - \sigma_i z_i - \frac{1}{2}\sigma_i^2 z_i^2 = \mathcal{O}(\sigma_i^3)$, and since $\theta \mathrm{e}^{\mu_i+x_i^*}\sigma_i^3 = \mathrm{o}(1)$ by (2.13) we have

$$\theta(\mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*})^\top \Big[\mathbf{e}^{\boldsymbol{\sigma}\circ\boldsymbol{z}} - \mathbf{1} - \boldsymbol{\sigma}\circ\boldsymbol{z} - \frac{1}{2}(\boldsymbol{\sigma}\circ\boldsymbol{z})^{\mathbf{2}}\Big] = \mathrm{o}(1). \quad (2.14)$$

Finally, we consider $\boldsymbol{B}$ as $\theta \to \infty$. Say that $n_+ := |\mathcal{F}_+|$ and assume that these are the first $n_+$ indices. We can then write that $\boldsymbol{B} \to \boldsymbol{B}^* := \operatorname{diag}(\boldsymbol{I}_{n_+}, \boldsymbol{F})$ where this $\boldsymbol{F}$ is the bottom-right submatrix of size $(n-n_+)\times(n-n_+)$ of the inverted correlation matrix implied by $\boldsymbol{\Sigma}$. The $\boldsymbol{B}$ matrices are positive definite for all $\theta \in (0,\infty]$; thus the limiting

form of the integrand in (2.12) is a non-degenerate multivariate normal density.

**Proposition 2.13.** $\lim_{\theta\to\infty} I(\theta) = 1$.

*Proof.* We use the dominated convergence theorem. By (2.14) and the paragraph which follows that equation, the exponent of the integrand is bounded by a constant $g_1$ for $\|\boldsymbol{z}\| < 1$, say, and the exponent is below $-g_2\|\boldsymbol{z}\|$ otherwise ($g_2 > 0$), for $\theta > \theta_0$, say. The latter comes from the positive definiteness of $\boldsymbol{B}^*$, the convergence of $\boldsymbol{B}$ to $\boldsymbol{B}^*$ and the convergence of (2.14). Next, convexity implies that the exponent is bounded by $-g_2\|\boldsymbol{z}\|$ for $\|\boldsymbol{z}\| > 1$. In total we have the bound

$$\exp(g_1\mathbb{I}\{\|\boldsymbol{z}\| \le 1\} - g_2\|\boldsymbol{z}\|\mathbb{I}\{\|\boldsymbol{z}\| > 1\}),$$

which is an integrable function. Thus the conditions for dominated convergence are satisfied and we can safely switch the limit and integral to obtain $I(\theta) \to 1$. □

## 2.5 Estimators of $\mathscr{L}(\theta)$ and $I(\theta)$

The simplest approach is to use quadrature to integrate the original expression in (2.1). This approach is used as a baseline against which the following estimators are compared (the approach can, however, be slow or impossible for large $n$). The next naïve approach is to estimate the expectation $\mathbb{E}[\mathrm{e}^{-\theta S_n}]$ by crude Monte Carlo (CMC). This would involve simulating random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_R \overset{\text{iid}}{\sim} \mathsf{Lognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{X}_r = (X_{r,1}, \ldots, X_{r,n})$, and computing

$$\widehat{\mathscr{L}}_{\text{CMC}}(\theta) := \frac{1}{R}\sum_{r=1}^{R} \exp\Big\{-\theta \sum_{i=1}^{n} X_{r,i}\Big\}.$$

However this estimator is not efficient for large $\theta$, and rare-event simulation techniques are required.

Given the decomposition of $\mathscr{L}(\theta) = \widetilde{\mathscr{L}}(\theta)I(\theta)$, some more accurate estimators can be assessed. Simply using $\widetilde{\mathscr{L}}(\theta)$ gives a biased estimator (which is fast and deterministic) for the transform, however the bias is decreased by estimating $I(\theta)$ with Monte Carlo integration. Proposition 2.3 gives two probabilistic representations of $I(\theta)$. We expect

the CMC estimator of the first—$\mathbb{E}[g(\boldsymbol{H}^{-1/2}\boldsymbol{Z})]$—to exhibit infinite variance as $\theta \to \infty$ as this has been proven for $n = 1$ in [20]. Therefore this estimator does not seem promising. The second estimator—$\sqrt{\det(\boldsymbol{\Sigma}\boldsymbol{H})}\ \mathbb{E}[v(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z})]$—can be viewed as the first estimator after importance sampling has been applied, so we focus upon this. Taking $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_R \overset{\text{iid}}{\sim} \mathsf{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$,

$$\widehat{\mathscr{L}_{\text{IS}}}(\theta) := \frac{1}{R} \exp\left\{ \left( \mathbf{1} - \frac{1}{2}\boldsymbol{x}^* \right)^\top \boldsymbol{D}\boldsymbol{x}^* \right\} \sum_{r=1}^{R} \exp\left\{ (\boldsymbol{x}^*)^\top \boldsymbol{D}(\mathrm{e}^{\boldsymbol{Z}_r} - \mathbf{1} - \boldsymbol{Z}_r) \right\}.$$

Many variance-reduction techniques can be applied to increase the efficiency of these estimators. The effect of including control variates into $\widehat{\mathscr{L}_{\text{IS}}}(\theta)$ was considered, using the control variate $(\boldsymbol{x}^*)^\top \boldsymbol{D}\boldsymbol{Z}_r^{\boldsymbol{2}}$ (note the element-wise square). The variance reduction achieved was small considering the large overhead of computing the variates (and their expectations) so these results have been omitted. Lastly, we considered an estimator based on the Gumbel distribution. Say that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is a vector of iid standard Gumbel random variables, that is, $\mathbb{P}(Y_r < x) = \exp\{-\mathrm{e}^{-x}\}$ for $x \in \mathbb{R}$. Then $\mathscr{L}(\theta)$ can be rewritten as an integral over the density of a vector of standard Gumbel random variables. This estimator was quite accurate, though it had higher relative error and variance than the estimators based on $\widehat{\mathscr{L}_{\text{IS}}}(\theta)$ so it too has been excluded from the results.

The final two variance reduction techniques investigated were *common random numbers* and *quasi-Monte Carlo* applied to $\widehat{\mathscr{L}_{\text{IS}}}(\theta)$; for a detailed explanation of these techniques see [79] or [15]. Both individually achieved significant variance reduction, and together provided the best estimator. Specifically,

$$\widehat{\mathscr{L}_{\text{Q}}}(\theta) := \frac{1}{R} \exp\left\{ \left( \mathbf{1} - \frac{1}{2}\boldsymbol{x}^* \right)^\top \boldsymbol{D}\boldsymbol{x}^* \right\} \sum_{r=1}^{R} \exp\left\{ (\boldsymbol{x}^*)^\top \boldsymbol{D}(\mathrm{e}^{\boldsymbol{q}_r} - \mathbf{1} - \boldsymbol{q}_r) \right\},$$

where $\boldsymbol{q}_r := \boldsymbol{\Sigma}^{1/2}\Phi^{-1}(\boldsymbol{u}_r)$, using $\Phi^{-1}$ as the (element-wise) standard normal inverse cdf, and where $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots\}$ is the $n$-dimensional Sobol sequence started at the same point for every $\theta$. Therefore, $\widehat{\mathscr{L}_{\text{Q}}}(\theta)$ is deterministic (for a fixed $R$ and $\theta$), and using this scheme is therefore a kind of numerical quadrature. More sophisticated adaptive quadrature methods could possibly be applied.

## 2.6 Numerical Results

Relative errors are given for the main estimators of $\mathscr{L}(\theta)$ in the table below. In all estimators the smoothing technique of using common random variables is employed, and all estimators are compared against numerical integration of the relevant integrals to 15 significant digits. See [111] for the software implementation used to create these results.

Table 2.1: Relative error for various approximations of $\mathscr{L}(\theta)$ for $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = [1, 0.5; 0.5, 1]$. The number of Monte Carlo replications $R$ used is $10^6$. Note: $*$ indicates that the CMC estimator simply gave an estimate of 0.

| $\theta$ | 100 | 2,500 | 5,000 | 7,500 | 10,000 |
|---|---|---|---|---|---|
| $\widetilde{\mathscr{L}}$ | -9.89e-3 | -1.27e-2 | -1.28e-2 | -1.27e-2 | -1.27e-2 |
| $\widehat{\mathscr{L}}_{\mathrm{CMC}}$ | 1.29e-2 | $*$ | $*$ | $*$ | $*$ |
| $\widehat{\mathscr{L}}_{\mathrm{IS}}$ | 3.36e-4 | 2.96e-4 | 2.57e-4 | 2.31e-4 | 2.11e-4 |
| $\widehat{\mathscr{L}}_{\mathrm{Q}}$ | -3.19e-6 | -5.03e-6 | -5.31e-6 | -5.56e-6 | -5.98e-6 |

Also, the pdf of $S_n$ can be estimated by numerical inversion of the Laplace transform. As the approximations of $\mathscr{L}(\theta)$ above are valid only for $\theta \in (0, \infty)$, not $\theta \in \mathbb{C}_+$, this restricts the options for Laplace transform inversion algorithms. The Gaver–Stehfest algorithm [156] and so-called power algorithms [24] can be used. We report on the results of using the Gaver–Stehfest algorithm as implemented by Mallet [121].

Other options for estimating $f(x)$ include numerically integrating the convolution equation (typically this is viable only for small $n$), the conditional Monte Carlo method (as in Example 4.3 on page 146 of [15]), and kernel density estimation. The following estimators are reported: the conditional Monte Carlo estimator $\widehat{f}_{\mathrm{Cond}}$, $\widetilde{f} := \mathscr{L}^{-1} \circ \widetilde{\mathscr{L}}$, $\widehat{f}_{\mathrm{IS}} := \mathscr{L}^{-1} \circ \widehat{\mathscr{L}}_{\mathrm{IS}}$ and $\widehat{f}_{\mathrm{Q}} := \mathscr{L}^{-1} \circ \widehat{\mathscr{L}}_{\mathrm{Q}}$.

The numerically inverted Laplace transforms are surprisingly accurate. Using common random numbers for the $\mathscr{L}(\theta)$ estimators was necessary, otherwise the inversion algorithms became confused by the non-smooth input. The precision of the inversion algorithms cannot be arbitrarily increased when using standard double-floating-point arithmetic [3], so the software suite MATHEMATICA was used. Yet this did not solve the problem of the Gaver–Stehfest algorithm becoming unstable (and very slow) when trying to increase the desired precision. Also, the inversion results became markedly poorer

Table 2.2: Relative errors for estimators of $f(x)$ for $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = [1, 0.5; 0.5, 1]$. The number of Monte Carlo repetitions for each $x$ is $R = 10^4$ for $\widehat{f}_{\text{Cond}}$, $\widehat{f}_{\text{IS}}$ and $\widehat{f}_{\text{Q}}$.

| $x$ | 0.01 | 1 | 1.5 | 2 | 3 |
|---|---|---|---|---|---|
| $\widehat{f}_{\text{Cond}}$ | -1.17e-1 | 2.20e-2 | 3.72e-3 | 5.21e-3 | -4.60e-3 |
| $\widetilde{f}$ | -7.03e-3 | 2.56e-2 | 1.79e-2 | 6.00e-2 | 3.82e-2 |
| $\widehat{f}_{\text{IS}}$ | 1.94e-3 | 1.43e-2 | -6.13e-3 | 4.00e-2 | 3.68e-3 |
| $\widehat{f}_{\text{Q}}$ | 2.90e-4 | 1.11e-2 | -9.04e-3 | 3.70e-2 | 2.44e-3 |

when $f(x)$ exhibited high kurtosis (i.e., when $\det(\boldsymbol{\Sigma})$ became small).

## 2.7 Closing Remarks

The estimators above give an accurate, relatively simple, and computationally swift method of computing the Laplace transform of the sum of dependent lognormals. We have shown that the approximation's error diminishes to zero ($I(\theta) \to 1$) as $\theta \to \infty$, and that it is still accurate for small values of $\theta$. One can find $\boldsymbol{x}^*$—for each $\theta$ examined—using a Newton–Raphson scheme, and Section 2.3 gives an accurate starting value for the iterations.

## 2.A Remaining steps in the proof of Theorem 2.7

First we note that all the minimisations are convex problems and therefore have unique solutions.

For the initial step of the algorithm let $\overline{\boldsymbol{w}}$ be the solution of the minimisation problem and let $\boldsymbol{e}_i$ be the vector with 1 at coordinate $i$ and zero at the other coordinates. Then $g_i(\varepsilon) = (\overline{\boldsymbol{w}} + \varepsilon\boldsymbol{e}_i)^\top \boldsymbol{D}(\overline{\boldsymbol{w}} + \varepsilon\boldsymbol{e}_i)$ is minimised at $\varepsilon = 0$. When $\overline{w}_i < -1$ the vector $\overline{\boldsymbol{w}} + \varepsilon\boldsymbol{e}_i$ is in the search set for all $\varepsilon$ small. We therefore have $g_i'(0) = 0$ which gives $\boldsymbol{D}_{i,\bullet}\overline{\boldsymbol{w}} = 0$. When $\overline{w}_i = -1$ the vector $\overline{\boldsymbol{w}} + \varepsilon\boldsymbol{e}_i$ is in the search set only for non-positive values of $\varepsilon$. This implies $g_i'(0) \leq 0$ giving $\boldsymbol{D}_{i,\bullet}\overline{\boldsymbol{w}} \leq 0$.

For the general recursive step we let $\boldsymbol{u} = \boldsymbol{w}_{\mathcal{F}_0(k-1)}$ and express $\boldsymbol{w}_{\mathcal{F}_-(k-1)}$ in terms of $\boldsymbol{u}$

from the equations $\boldsymbol{D}_{i,\bullet}\boldsymbol{w} = 0$, $i \in \mathcal{F}_-(k-1)$. The derivative of $\boldsymbol{w}^\top \boldsymbol{D}\boldsymbol{w}$ with respect to $u_i$ ($i$ being the index inherited from $\boldsymbol{w}$) is then

$$2\boldsymbol{D}_{i,\bullet}\boldsymbol{w} + 2\frac{\partial \boldsymbol{w}_{\mathcal{F}_-(k-1)}}{\partial u_i}\boldsymbol{D}_{\mathcal{F}_-(k-1)}\boldsymbol{w} = 2\boldsymbol{D}_{i,\bullet}\boldsymbol{w}.$$

As above we find that the derivative of $\boldsymbol{w}^\top \boldsymbol{D}\boldsymbol{w}$ with respect to $u_i$ at the minimising point is zero when $u_i < 0$ and less than or equal to zero when $u_i = 0$.

What is left to prove is that $\mathcal{F}_0(k)$ always has at least one element with $\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,k+1} < 0$. To this end define $d_1 = -\boldsymbol{\beta}_{\bullet,1}$ and $d_k = d_{k-1} - \boldsymbol{\beta}_{\bullet,k}$ for $k > 1$. From the properties of $\boldsymbol{\beta}$ we find

$$\begin{aligned} &d_{\mathcal{F}_+(k),k} = 0; \quad d_{\mathcal{F}_*(k),k} = 1 \text{ and } \boldsymbol{D}_{\mathcal{F}_*(k)}d_k > 0; \\ &d_{\mathcal{F}_0(k),k} = 1 \text{ and } \boldsymbol{D}_{\mathcal{F}_0(k)}d_k = 0; \quad \boldsymbol{D}_{\mathcal{F}_-(k)}d_k = 0. \end{aligned}$$

Assume now that $\boldsymbol{D}_{i,\bullet}\boldsymbol{\beta}_{\bullet,k+1} = 0$ for all $i \in \mathcal{F}_0(k)$. We show that this leads to a contradiction. Using the assumption, $\boldsymbol{\beta}_{\bullet,k+1}$ has the properties

$$\begin{aligned} &\boldsymbol{\beta}_{\mathcal{F}_+(k),k+1} = 0; \quad \boldsymbol{\beta}_{\mathcal{F}_*(k),k+1} = 1; \\ &\boldsymbol{\beta}_{\mathcal{F}_0(k),k+1} \leq 0 \text{ and } \boldsymbol{D}_{\mathcal{F}_0(k)}\boldsymbol{\beta}_{\bullet,k+1} = 0; \quad \boldsymbol{D}_{\mathcal{F}_-(k)}\boldsymbol{\beta}_{\bullet,k+1} = 0. \end{aligned}$$

Combining the two displays we have

$$\boldsymbol{D}_{\mathcal{F}_0(k)}d_k = \boldsymbol{D}_{\mathcal{F}_0(k)}\boldsymbol{\beta}_{\bullet,k+1}, \quad \boldsymbol{D}_{\mathcal{F}_-(k)}d_k = \boldsymbol{D}_{\mathcal{F}_-(k)}\boldsymbol{\beta}_{\bullet,k+1}.$$

Since $d_k$ and $\boldsymbol{\beta}_{\bullet,k+1}$ are identical on $\mathcal{F}_+(k-1)$ and $\mathcal{F}_*(k-1)$ the equations reduce to

$$\boldsymbol{D}_0 \begin{pmatrix} d_{\mathcal{F}_0(k),k} \\ d_{\mathcal{F}_-(k),k} \end{pmatrix} = \boldsymbol{D}_0 \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{F}_0(k),k} \\ \boldsymbol{\beta}_{\mathcal{F}_-(k),k} \end{pmatrix}, \text{ where } \boldsymbol{D}_0 = \begin{pmatrix} \boldsymbol{D}_{\mathcal{F}_0(k),\mathcal{F}_0(k)} & \boldsymbol{D}_{\mathcal{F}_0(k),\mathcal{F}_-(k)} \\ \boldsymbol{D}_{\mathcal{F}_-(k),\mathcal{F}_0(k)} & \boldsymbol{D}_{\mathcal{F}_-(k),\mathcal{F}_-(k)} \end{pmatrix}.$$

Since the matrix $\boldsymbol{D}_0$ is positive definite and $d_{\mathcal{F}_0(k),k} \neq \boldsymbol{\beta}_{\mathcal{F}_0(k),k}$, we have reached a contradiction. □

## Chapter 3 Authorship Statement

**Citation:** Søren Asmussen, Pierre-Olivier Goffard, Patrick J. Laub (2015), *Orthonormal polynomial expansions and lognormal sum densities*, Risk and Stochastics: Ragnar Norberg at 70 (Mathematical Finance Economics), World Scientific

The authors of this paper equally contributed to the following tasks:

1. conception and design of the project;
2. mathematical arguments, and interpretation of the results;
3. writing the publication.

In addition to this, I completed the majority of the computational work and of the editing (e.g. checking grammar and typographical details).

# Chapter 3

# Orthonormal polynomial expansions and densities of sums of lognormals

## 3.1 Introduction

The previous chapter considered approximating the SLN Laplace transform and used this to approximate distribution's pdf. This chapter discusses a different method where one approximates the SLN pdf $f$ using polynomials $\{p_k\}_{k\in\mathbb{N}_0}$ which are orthonormal w.r.t. some reference pdf $w$. In the general formulation, one is interested in approximating a target density $g$ using the pdf $w$ as reference. One then finds a series representation of $g/w$ of the form $\sum_{k=0}^{\infty} a_k p_k$, and then the approximation of $g$ is

$$\widehat{g}(x) \;=\; w(x)\sum_{k=0}^{K} a_k p_k(x), \tag{3.1}$$

for some suitable $K$. The most obvious connection to the sum of lognormals problem is $g = f$, but for some choices of $w$ we must take a different target $g$. In one case we set $g$ as the density of $\log S$ and transform back to get the approximation $\widehat{f}(x) = \widehat{g}(\log x)/x$. In another case we set $g$ as the exponentially tilted SLN pdf. The choice of $w$ is a crucial step, and three candidates for $w$ are investigated, the pdfs of the normal, gamma, and lognormal distributions.

The form of the $p_k$ is classical for the normal distribution where it is the Hermite polynomials and for the gamma where it is the generalised Laguerre polynomials, but for the lognormal distributions it does not appear to be in the literature and we give here the functional expression (Theorem 3.3). The Fenton–Wilkinson method may be seen as the $K = 2$ case of $w$ being lognormal (with $g = f$), and this choice of $w$ may be the most obvious one. However, we show that in the lognormal case the orthonormal polynomials are not dense in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$. This result is closely related to the lognormal distribution not being determined by its moments [98, 31] and indicates that a lognormal $w$ is potentially dangerous. For this reason, the rest of the chapter concentrates on taking the reference distribution as normal (using the logarithmic transformation) or gamma (using exponential tilting).

Applying orthogonal polynomial expansions to sums of lognormals is not a new idea; many papers, whose motivation relates to pricing Asian options, have contributed to this task. The earliest relevant work on this is from Turnbull and Wakeman [160], who constructed an orthogonal polynomial expansion for the sum of correlated lognormals using a lognormal reference distribution. Dufresne and Li refer state of this: "the convergence of those ... series has never been proved, and their theoretical convergence is highly unlikely" [67, p. 1]. This conjecture is precisely what we have proved here, showing the incompleteness of the lognormal orthogonal polynomials in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$.

The next development was Dufresne [63] who tackled the related problem of pricing continuous Asian options — here the underlying average is $A = \frac{1}{T}\int_0^T \mathrm{e}^{X_t}\,\mathrm{d}t$ where $\{X_t\}_{t\in[0,T}$ is a Brownian motion. They use an orthogonal polynomial expansion using a gamma reference distribution, and sidestep the integrability problem by approximation the pdf of $1/A$ rather than $A$. The expansion's coefficient are constructed from the moments of $1/A$, which are found using a recursive scheme of symbolic integrations in MATHEMATICA.

Popovic and Goldsman [140] consider the problem pricing discrete Asian options — that is, the underlying average is $A = \frac{1}{n}\sum_{i=1}^n \mathrm{e}^{X_i}$ where $\{X_i\}_{i\in\mathbb{N}_0}$ is a discretely observed Brownian motion (they consider stock prices driven by other Lévy processes, e.g., the variance-gamma process). They construct an orthogonal expansion of $\log(A)$ using a normal reference distribution, and evaluate the coefficients using Monte Carlo (with some variance reduction techniques). Dufresne and Li [67] take the same approach, but add in

the theory which was missing by giving the conditions so that the orthogonal expansion will converge. They also write the explicit form of an Asian option price given this orthogonal polynomial approximation for $\log(A)$. Lastly, the paper from Chateau and Dufresne [45], which is more recent than the contents of this chapter, ought to be noted to complete this review.

As noted above, we are the first to derive the orthogonal polynomials w.r.t. the lognormal reference, and prove their incompleteness in the relevant space. Our normal reference approximation is applied to the logarithm of the sum, akin to [140] and [67]. The gamma reference approximation is not applied to the reciprocal random variable, like [63], but to the exponentially tilted sum of lognormal distribution.

After discussing the details of the orthonormal polynomials expansions in Sections 3.2 and 3.3, we proceed in Section 3.4 to show a number of numerical examples. The polynomial expansions are compared to existing methods as Fenton–Wilkinson and a more recent approximation in terms of log skew normal distributions [95], as well as to exact values obtained by numerical quadrature in cases where this is possible or by Monte Carlo density estimation. Section 3.4 also outlines an extension to statistical data sets and non-Gaussian copulas. Appendix A contains a technical proof and Appendix B some new material on the SLN Laplace transform.

## 3.2 Orthogonal polynomial representation of probability density functions

Let $X$ be a random variable which has a density $f$. If $f$ is unknown but the distribution of $X$ is expected to be close to some pdf $w$, one may use $w$ as a first approximation to $f$ and next try to improve by invoking suitable correction terms.

In the setting of this chapter $X$ is the sum of lognormal random variables and the correction terms are obtained by expansions in terms of orthonormal polynomials. Before going into the details of the lognormal example, let us consider the general case.

Assuming all moments of $w$ to be finite, the standard Gram–Schmidt orthogonalisation technique shows the existence of a set of polynomials $\{p_k\}_{k\in\mathbb{N}_0}$ which are orthonormal in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ equipped with the usual inner product $\langle g, h\rangle_w$ and the corresponding norm $\|g\|^2$. From Proposition 1.16, we know that if there is an $\alpha > 0$ such that

$$\int_{\mathbb{R}} \mathrm{e}^{\alpha|x|} w(x)\,\mathrm{d}x < \infty\,, \tag{3.2}$$

then the polynomials $\{p_k\}_{k\in\mathbb{N}_0}$ are complete in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$. The implication is that if $f/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$, that is, if

$$\int_{\mathbb{R}} \frac{f(x)^2}{w(x)^2} w(x)\,\mathrm{d}x \;=\; \int_{\mathbb{R}} \frac{f(x)^2}{w(x)}\,\mathrm{d}x \;<\; \infty\,, \tag{3.3}$$

we may expand $f/w$ as $\sum_{k=0}^{\infty} a_k p_k$ where

$$a_k \;= \langle f/w, p_k\rangle_w \;=\; \int_{\mathbb{R}} f(x) p_k(x)\,\mathrm{d}x \;=\; \mathbb{E}\left[p_k(X)\right]\,. \tag{3.4}$$

This suggests that we use the form of (3.1) as an approximation of $f$ in situations where the pdf of $X$ is unknown but the moments are accessible.

**Remark 3.1.** If the first $m$ moments of $X$ and $w$ coincide, one has $a_k = 0$ for $k = 1, \ldots, m$, which is a consequence of the linear independence of the orthogonal polynomials. When choosing $w$, a possible guideline is therefore to match as many moments as possible. ◇

Due to the Parseval relationship $\sum_{k=0}^{\infty} a_k^2 = \|f/w\|^2$, the coefficients of the polynomial expansion, $\{a_k\}_{k\in\mathbb{N}_0}$, tend toward 0 as $k \to \infty$. The accuracy of the approximation (3.1), for a given order of truncation $K$, depends upon how swiftly the coefficients decay; note that the $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ loss of the approximation of $f/w$ is $\sum_{K+1}^{\infty} a_k^2$. Note also that the orthogonal polynomials can be specified recursively (see Thm. 3.2.1 of [158]) which allows a reduction of the computing time required for the coefficients' evaluation and makes it feasible to consider rather large $K$.

### 3.2.1 Normal reference distribution

A common choice as a reference distribution is the normal $\mathsf{Normal}(\mu, \sigma^2)$. The associated orthonormal polynomials are given by

$$p_k(x) = \frac{1}{k! 2^{k/2}} H_k\left(\frac{x-\mu}{\sigma\sqrt{2}}\right), \tag{3.5}$$

where $\{H_k\}_{k\in\mathbb{N}_0}$ are the (physicists') Hermite polynomials, defined in [158] for instance. If $f$ is continuous, a sufficient (and close to necessary) condition for $f/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ is

$$f(x) \;=\; \mathcal{O}(\mathrm{e}^{-ax^2}) \quad \text{as } x \to \pm\infty \quad \text{with } a > (4\sigma^2)^{-1}\,. \tag{3.6}$$

Indeed, we can write the integral in (3.3) as $I_1 + I_2 + I_3$, the integrals over $(-\infty, -A)$, $[-A, A]$, resp. $(A, \infty)$. Note that $I_2 < \infty$ follows since the integrand is finite by continuity, whereas the finiteness of $I_1, I_3$ is ensured by the integrands being $\mathcal{O}(\mathrm{e}^{-bx^2})$ where $b = 2a - 1/2\sigma^2 > 0$. Similar arguments apply to conditions (3.10) and (3.13) below.

**Example 3.2.** The classical example of this technique (though not usually put in this framework) is the *Edgeworth expansion*, which is almost identical to the *Gram–Charlier expansion of type A*. Consider $Y_1, \ldots Y_n$ iid random variables where $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}\mathrm{ar}[Y_i] = \omega^2$. The Edgeworth expansion covers $X = (S - n\mu)/(\sqrt{n}\,\omega)$ with the $\mathsf{Normal}(0, 1)$ density as the reference pdf.

The Edgeworth example illustrates well several aspects of the theory. In the scenario of Remark 3.1 with $m = 2$, condition (3.2) is trivially satisfied, but condition (3.3) requires more attention. It actually fails unless $f(x)$ decays very quickly as $x \to \pm\infty$, even in such a basic example as the $Y_i$ being standard exponential (then $f(x)$ is of order $x^{n-1}\mathrm{e}^{-xn^{1/2}}$ which multiplied by $\mathrm{e}^{x^2/2}$ does not integrate). Nevertheless, the Edgeworth approach has been observed to perform well even when condition (3.3) is not satisfied; in fact, the condition is only sufficient, not necessary (Cramér [54] showed it can be relaxed to $\mathbb{E}[\mathrm{e}^{X^2/4}] < \infty$). ◇

### 3.2.2 Gamma reference distribution

If $X$ has support $(0, \infty)$, it is natural to look for a $w$ with the same property. An obvious candidate is the gamma distribution, denoted $\mathsf{Gamma}(r, m)$ where $r$ is the shape parameter and $m$ the scale parameter. The pdf is

$$w(x) = \frac{x^{r-1}\mathrm{e}^{-x/m}}{m^r \Gamma(r)}, \quad x \in \mathbb{R}_+ \,. \tag{3.7}$$

The associated polynomials are given by

$$p_n(x) = (-1)^n \left[\frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)}\right]^{-1/2} L_n^{r-1}(x/m), \quad n \in \mathbb{N}_0, \tag{3.8}$$

where $\{L_n^{r-1}\}_{n\in\mathbb{N}_0}$ denote the *generalised Laguerre polynomials*, see [158]; in MATHEMATICA these are accessible via the `LaguerreL` function. The polynomials defined in (3.8) satisfy the recurrence relationship

$$\begin{aligned} np_n(x) &= \left(\frac{x}{m} - 2n - r + 2\right) p_{n-1}(x)\sqrt{\frac{n}{n+r-1}} \\ &- (n+r-2)p_{n-2}(x)\sqrt{\frac{n(n-1)}{(n+r-1)(n+r-2)}}. \end{aligned} \tag{3.9}$$

The recurrence relationship will be employed later to speed up the computation of the coefficients. A sufficient condition, similar to (3.6), for $f/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ is:

$$\begin{aligned} f(x) &= \mathcal{O}(\mathrm{e}^{-\delta x}) \quad \text{as } x \to \infty \quad \text{with } \delta > 1/2m\,, \text{ and} \\ f(x) &= \mathcal{O}(x^{\beta}) \quad\;\; \text{as } x \to 0 \quad\;\; \text{with } \beta > r/2 - 1\,. \end{aligned} \tag{3.10}$$

### 3.2.3 Lognormal reference distribution

We denote the lognormal distribution as $\mathrm{e}^Y \sim \mathsf{Lognormal}(\mu, \sigma^2)$ where $Y \sim \mathsf{Normal}(\mu, \sigma^2)$. It has support on $\mathbb{R}_+$. The polynomials orthogonal to the lognormal distribution are given in the following proposition, to be proved in the Appendix:

**Theorem 3.3.** *The polynomials orthonormal w.r.t. the* $\mathsf{Lognormal}(\mu, \sigma^2)$ *pdf are*

$$p_k(x) = \frac{\mathrm{e}^{-\frac{k^2\sigma^2}{2}}}{\sqrt{[\mathrm{e}^{-\sigma^2}; \mathrm{e}^{-\sigma^2}]_k}} \sum_{i=0}^{k} (-1)^{k+i} \mathrm{e}^{-i\mu - \frac{i^2\sigma^2}{2}} e_{k-i}(1, \ldots, \mathrm{e}^{(k-1)\sigma^2}) x^i, \tag{3.11}$$

*for* $k \in \mathbb{N}_0$ *where the* $e_i$ *are the elementary symmetric polynomials*

$$e_i(x_1, \ldots, x_k) = \begin{cases} 1 & \text{for } i = 0, \\ \sum_{1 \le j_1 < \ldots < j_i \le k} x_{j_1} \ldots x_{j_i}, & \text{for } 1 \le i \le k, \\ 0, & \text{for } i > k, \end{cases} \tag{3.12}$$

*and* $[x; q]_n = \prod_{i=0}^{n-1} (1 - xq^i)$ *is the q-Pochhammer symbol.*

**Remark 3.4.** The result of Theorem 3.3 does not appear to be in the literature; the closest reference seems to be a 1923 paper by Wigert [162] who considers the distribution with pdf $\ell \mathrm{e}^{-\ell^2 \ln^2(x)}/\sqrt{\pi}$, for $x > 0$, introduced by Stieljes [157, pp. 507–508] (later called the Stieltjes–Wigert distribution). These polynomials are also mentioned in [49, pp. 172–175]. ◇

The equivalent of condition (3.6) for $f/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ now becomes

$$f(x) \;=\; \mathcal{O}(\mathrm{e}^{-b \log^2 x}) \quad \text{for } x \to 0 \text{ and } \infty \quad \text{with } b > (4\sigma^2)^{-1}\,, \tag{3.13}$$

which is rather mild. However, a key difficulty in taking the reference distribution as lognormal is the following result related to the fact that the lognormal and the Stieltjes–Wigert distributions are not characterised by their moments, see [98, 31, 48, 50]. Hence, the orthogonal polynomials associated with the lognormal pdf and the Stieltjes–Wigert pdf are also the orthogonal polynomials for some other distribution.

**Proposition 3.5.** *The set of orthonormal polynomials in Theorem 3.3 is incomplete in* $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$. *That is,* $\mathrm{span}\{p_k\}_{k \in \mathbb{N}_0}$ *is a proper subset of* $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$.

*Proof.* Let $Y$ be a random variable whose distribution is the lognormal with pdf $w(x)$ and $X$ a random variable with a distribution different from $Y$ but with the same moments.

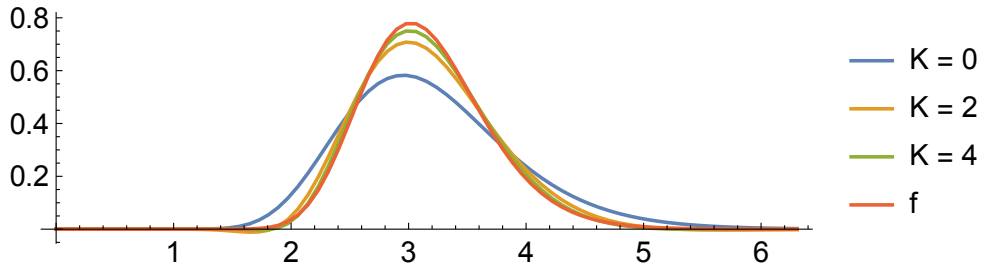



Figure 3.1: Examples of orthogonal polynomial approximations using a $\mathsf{Normal}(1.13, 0.23^2)$ reference converging to the target $f$ with increasing $K$.

According to [31, pp. 201–202] such an $X$ can be chosen such that $f_X/w$ is bounded and hence in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$. The projection of $f_X/w$ onto $\operatorname{span}\{p_k\}$ is then

$$\begin{aligned}\sum_{k=0}^{\infty} \langle f_X/w, p_k\rangle_w \, p_k \;&=\; \sum_{k=0}^{\infty} \mathbb{E}\left[p_k(X)\right] p_k \;=\; \sum_{k=0}^{\infty} \mathbb{E}\left[p_k(Y) p_0(Y)\right] p_k \\ &= \; p_0 \;=\; 1 \neq f_X/w,\end{aligned}$$

where the first step used (3.4), the second step that the moments are the same (and that $p_0 \equiv 1$), and the third follows by orthogonality of the polynomials. This implies $f_X/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x) \setminus \operatorname{span}\{p_k\}$ and the assertion. □

### 3.2.4 Convergence of the estimators w.r.t. $K$

Orthogonal polynomial approximations generally become more accurate as the order of the approximation $K$ increases. Figure 3.1 shows a specific orthogonal polynomial approximation, $\widehat{f}_{\mathrm{N}}$ (which involves a logarithmic transformation and is described in Section 3.3.2), converging to the true SLN density $f$ for increasing $K$. In this example, we take the SLN distribution with $\mu = (0, 0, 0)^\top$, $\Sigma_{ii} = 0.1$, and $\rho = -0.1$.

Proposition 3.5 implies that orthogonal polynomial approximations with a lognormal reference distribution cannot be relied upon to converge to the desired target density but may have a different limit (the orthogonal projection described there). The next plot, Figure 3.2, illustrates this phenomenon. The approximation appears to converge, but not to the target density. Our theoretical discussion suggests that this incorrect limit density has the same moments as the target lognormal distribution, and this was verified

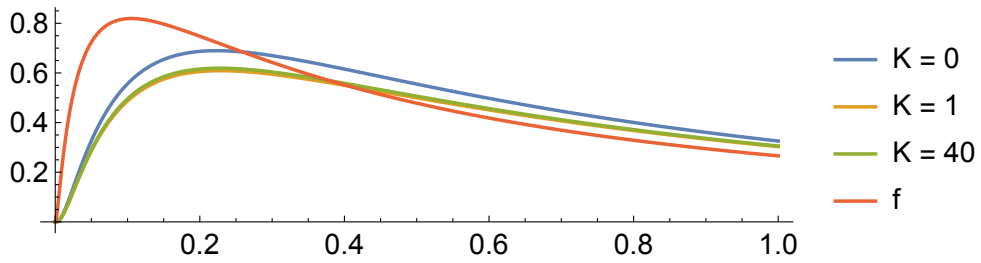



Figure 3.2: Example of orthogonal polynomial approximations of $f$ using a $\mathsf{Lognormal}(0, 1.22^2)$ reference not converging to the $\mathsf{Lognormal}(0, 1.50^2)$ target.

numerically for the first few moments.

Lastly, it must be noted that we cannot in practice take $K$ arbitrarily large, due to numerical errors incurred in calculating the $\{a_k\}$ coefficients. Obviously this can be overcome by using infinite precision operations, however this swiftly becomes prohibitively slow. Software tools like MATHEMATICA allow for arbitrarily large but finite precision, which gives on the flexibility to choose a desired accuracy/speed trade-off. We use this technology and select $K \leq 40$.

## 3.3 Application to sums of lognormals

Now we turn to our main case of interest where $X = S$ is a sum of lognormals. Specifically,

$$S = \mathrm{e}^{X_1} + \ldots + \mathrm{e}^{X_n}\,, \quad n \geq 2\,, \tag{3.14}$$

where the vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is governed by a multivariate normal distribution $\mathsf{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is the mean vector and $\boldsymbol{\Sigma} = (\sigma_{ij})$ the covariance matrix. This distribution is denoted $\mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and hereafter $f$ will be its pdf. We are interested in computing the pdf when the summands exhibit dependency ($\boldsymbol{\Sigma}$ is non-diagonal). This is an ambitious goal given that the pdf of the sum of two iid lognormally distributed random variables is already unknown. The validity of the polynomial approximations rely on the $\mathcal{L}^2$ integrability condition (3.3), which is difficult to check because the pdf of $S$ is not available. This challenge is solved by using asymptotic results describing the left and the right tail of the distribution of $S$, which are collected

in the following subsection.

### 3.3.1 Tail asymptotics of sums of lognormals

The tail asymptotics of $f(x)$ are given in the following lemma, which simply collects the results from Corollary 2 of [85] and Theorem 1 of [23].

**Lemma 3.6.** *We have*

$$f(x) = \mathcal{O}(\mathrm{e}^{-c_1 \ln(x)^2}) \text{ as } x \to 0 \text{ and} \tag{3.15}$$

$$f(x) = \mathcal{O}(\mathrm{e}^{-c_2 \ln(x)^2}) \text{ as } x \to \infty \tag{3.16}$$

*where*

$$c_1 = \left[2 \min_{\boldsymbol{w} \in \Delta} \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}\right]^{-1} \text{ and } \quad c_2 = \left[2 \max_{i=1,\ldots,n} \sigma_{ii}\right]^{-1},$$

*with the notation that* $\Delta = \{\boldsymbol{w} \,|\, w_i \in \mathbb{R}_+, \sum_{i=1}^n w_i = 1\}$.

We are also interested in the asymptotic behaviour of $Z = \ln(S)$ later in the chapter. L'Hôpital's rule gives us the asymptotic tails (extending [77]) of $f_Z(z) = \mathrm{e}^z f(\mathrm{e}^z)$ to be:

**Corollary 3.7.** *We have*

$$f_Z(z) = \mathcal{O}(\mathrm{e}^{-c_1 z^2}) \text{ as } z \to -\infty \text{ and} \tag{3.17}$$

$$f_Z(z) = \mathcal{O}(\mathrm{e}^{-c_2 z^2}) \text{ as } z \to +\infty \tag{3.18}$$

*where the constants are as in Lemma 3.6.*

### 3.3.2 Sums of lognormals with a normal reference distribution

Consider transforming the sum to $Z = \ln(S)$ and expanding this density with orthogonal polynomials using a normal distribution as reference. That is, our approximation to $f$ using a $\mathsf{Normal}(\mu, \sigma^2)$ reference is

$$\widehat{f}_{\mathrm{N}}(x) = \frac{1}{x} \widehat{f}_Z(\ln x) \quad \text{where} \quad \widehat{f}_Z(z) = \phi_{\mu,\sigma^2}(z) \sum_{i=1}^K a_i \, p_i(z)\,,$$

with the normal pdf $\phi_{\mu,\sigma^2} = w$. The following result tells us when the integrability condition $f_Z/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ is satisfied.

**Proposition 3.8.** *Consider $Z = \ln(S)$ where $S \sim \mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $w$ be the pdf of the $\mathsf{Normal}(\mu, \sigma^2)$ distribution. We have $f_Z/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ if*

$$2\sigma^2 > (2c_2)^{-1} = \max_{i=1,\ldots,n} \Sigma_{ii}\,. \tag{3.19}$$

*Proof.* It follows immediately by combining (3.6) and Corollary 3.7. □

Computing the $\{\widehat{a}_k\}_{k\in\mathbb{N}_0}$ coefficients can be done using Crude Monte Carlo (CMC), as in

$$\widehat{a}_k = \frac{1}{R}\sum_{r=1}^{R} p_n(\ln S_r)\,, \quad S_1, \ldots, S_R \overset{\text{iid}}{\sim} \mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,,$$

for $k = 0, \ldots, K$. We can use the same $S_1$, …, $S_R$ for all $\widehat{a}_k$ together with a smoothing technique called *common random numbers* [15, 79]. Note that a non-trivial amount of computational time is typically spent just constructing the Hermite polynomials. Incorporating the Hermite polynomial's recurrence relation in our calculations achieved a roughly 40× speed-up compared with using MATHEMATICA's `HermiteH` function.

### 3.3.3 Sums of lognormals with a gamma reference distribution

When $w$ is the pdf of the $\mathsf{Gamma}(r, m)$ distribution, it makes little sense to expand $f$ in terms of $\{p_k\}_{k\in\mathbb{N}_0}$ as the integrability condition (3.10) fails, $f/w \notin \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$. The workaround consists in using orthogonal polynomials to expand the *exponentially tilted* distribution, denoted $\mathsf{SumLognormal}_\theta(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This distribution's pdf is

$$f_\theta(x) = \frac{\mathrm{e}^{-\theta x} f(x)}{\mathscr{L}(\theta)}\,, \quad \theta \geq 0, \tag{3.20}$$

where $\mathscr{L}(\theta) = \mathbb{E}[\mathrm{e}^{-\theta S}]$ is the Laplace transform of $S$. Asmussen et al. [19] investigated the use of $f_\theta(x)$ in approximating the left tail of $S$, and developed asymptotic forms and Monte Carlo estimators of this density.

**Remark 3.9.** The use of gamma distribution and Laguerre polynomials links our approach to a well established technique called the *Laguerre method*. The expansion is an orthogonal projection onto the basis of Laguerre functions constructed by multiplying Laguerre polynomials and the square root of the exponential distribution with parameter 1. The method is described in [1]. Note also that the damping procedure employed when integrability problems arise is quite similar to considering the exponentially tilted distribution instead of the real one. The use of the gamma distribution as reference is applied to actuarial science in [84, 83]. ◇

Using (3.10), we immediately obtain the following result which sheds light on how to tune the parameters of the reference gamma distribution so the integrability condition $f_\theta/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ is satisfied.

**Proposition 3.10.** *Consider the random variable* $S_\theta \sim \mathsf{SumLognormal}_\theta(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, and let* $w$ *be the pdf of the* $\mathsf{Gamma}(r, m)$ *distribution. We have* $f_\theta/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ *if* $m > 1/2\theta$*.*

Hereafter we assume that the parameters $r$ and $m$ of the $\mathsf{Gamma}(r, m)$ reference distribution are chosen to satisfy Proposition 3.10's conditions.

Our approximation—based upon rearranging (3.20)—is of the form

$$\widehat{f}(x) = \mathrm{e}^{\theta x}\mathscr{L}(\theta)\widehat{f}_\theta(x) = \mathrm{e}^{\theta x}\mathscr{L}(\theta)\sum_{k=0}^{K} a_k p_k(x) w(x)\,. \tag{3.21}$$

The coefficients $a_k = \mathbb{E}[p_k(S_\theta)]$ can be estimated in (at least) three different ways: (i) using CMC, (ii) using MC while importance sampling from the original $\mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, or (iii) by directly computing the moments $\mathbb{E}[S_\theta^k]$. The first method is nontrivial, as simulating from $f_\theta$ likely requires using acceptance-rejection (as in [19]). Options (ii) and (iii) use

$$a_k = \mathbb{E}[p_k(S_\theta)] =: q_{k0} + q_{k1}\,\mathbb{E}[S_\theta] + \cdots + q_{kk}\,\mathbb{E}[S_\theta^k] \tag{3.22}$$

where $\{q_{ki}\}$ are the coefficients in $p_k$, and

$$\mathbb{E}[S_\theta^i] = \mathbb{E}[S^i \mathrm{e}^{-\theta S}]/\mathscr{L}(\theta) =: \mathscr{L}_i(\theta)/\mathscr{L}(\theta)\,.$$

The $\mathscr{L}_i(\theta)$ notation was selected to highlight the link between $\mathbb{E}[S_n^i \mathrm{e}^{-\theta S_n}]$ and the $i$-th

derivative of $\mathscr{L}(\theta)$. All three methods require access to the Laplace transform, and method (iii) requires $\mathscr{L}_i(\theta)$, however none of $\mathscr{L}(\theta)$ or $\mathscr{L}_i(\theta)$ are available in closed form. Our approach to circumvent these problems is presented in the Appendix.

## 3.4 Numerical illustrations

We take several approximations $\widehat{f}$ and compare them against the benchmark of numerical integration. One form of $f$ particularly useful for numerical integration, in terms of $f_{\boldsymbol{X}}$ the density of $\boldsymbol{X} \sim \mathsf{Lognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is as a surface integral, $f(s) = n^{-\frac{1}{2}} \int_{\Delta_n^s} f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$, where $\Delta_n^s = \{\boldsymbol{x} \in \mathbb{R}_+^n : ||\boldsymbol{x}||_1 = s\}$. MATHEMATICA integrates this within a reasonable time for $n = 2$ to 4 using `NIntegrate` and `ParametricRegion`). For $n > 4$ we qualitatively assess the performance of the estimators by plotting them.

The error measure used is the $\mathcal{L}^2([0, \mathbb{E}[S]])$ norm of $\widehat{f} - f$. We focus on this region as it is the hardest to approximate (indeed, Lemma 3.6 shows that just a single lognormal is a theoretically justified approximation of the SLN right tail) and due to its special relevance in applications, see for example the introduction of [19] and the references therein.

### 3.4.1 The estimators

We will compare the following approximations:

- the Fenton-Wilkinson approximation $\widehat{f}_{\mathrm{FW}}$, cf. [71], consists in approximating the distribution of $S$ by a single lognormal with the same first and second moment;
- the log skew normal approximation $\widehat{f}_{\mathrm{Sk}}$, cf. [95][1], is a refinement of Fenton–Wilkinson by using a log skew normal as approximation and fitting the left tail in addition to the first and second moment;
- the conditional Monte Carlo approximation $\widehat{f}_{\mathrm{Cond}}$ , cf. Example 4.3 on p. 146 of [15], uses the representation $f(x) = \mathbb{E}[\mathbb{P}(S \in \mathrm{d}x \mid Y)]$ for some suitable $Y$ (here chosen

[1]Note that in [95], the formula for $\varepsilon_{\mathrm{opt}}$ contains an typographic error.

as one of the normal random variables $X_i$ occurring in (3.14)) and simulates the conditional expectation;

- $\widehat{f}_{\mathrm{N}}$ is the approximation described in Section 3.3.2 using a logarithmic transformation and the Hermite polynomials with a normal reference distribution;

- $\widehat{f}_{\Gamma}$ is the approximation described in Section 3.3.3 using exponential tilting and the generalised Laguerre polynomials with a gamma reference distribution.

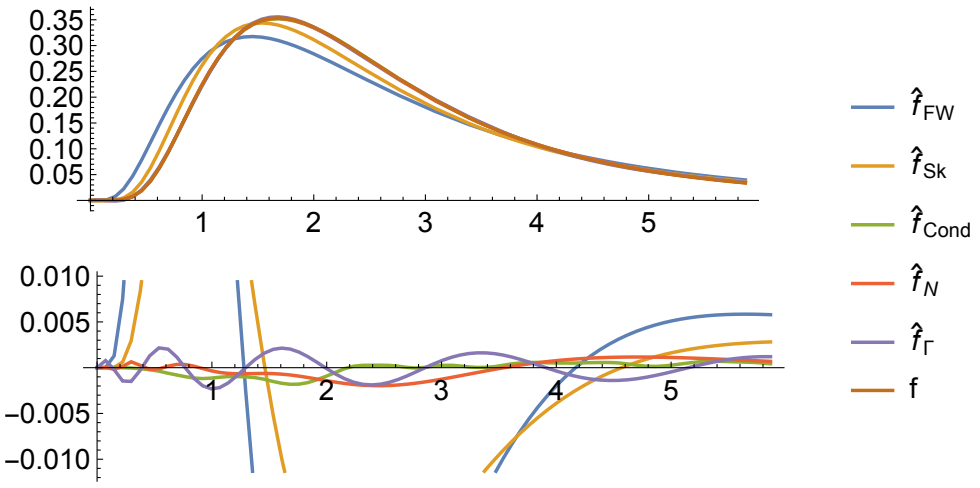
These approximations are all estimators of functions (i.e., not pointwise estimators, such as in Chapter 2) and they do not take excessive computational effort to construct. The first two, $\widehat{f}_{\mathrm{FW}}$ and $\widehat{f}_{\mathrm{Sk}}$, only need $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and do not have any Monte Carlo element. Similarly, the estimator $\widehat{f}_{\Gamma}$ when utilising the Gauss–Hermite quadrature described in (3.31) in the Appendix does not use Monte Carlo. For the remaining approximations we utilise the *common random numbers* technique, meaning that the same $R = 10^5$ iid $\mathsf{SumLognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ samples $\boldsymbol{S} = (S_1, \ldots, S_R)^{\top}$ are given to each algorithm. Lastly, all the estimators except $\widehat{f}_{\Gamma}$ satisfy $\int_{\mathbb{R}} \widehat{f}(x)\,\mathrm{d}x = 1$. One problem with the orthogonal polynomial estimators is that they can take negative values; this can easily be fixed, but we do not make that adjustment here.

For $\widehat{f}_{\mathrm{N}}$, we take $\mu = \mathbb{E}[Z]$ and $\sigma^2 = \mathbb{V}\mathrm{ar}[Z]$, calculated using numerical integration. The $\widehat{f}_{\Gamma}$ case is more difficult. Equation (3.21) shows that we must impose $\theta m < 1$ to ensure that $\widehat{f}_{\Gamma}(x) \to 0$ as $x \to \infty$. Exploring different parameter selections showed that fixing $\theta = 1$ worked reasonably well. Moment matching $f_\theta$ to $w$ leads to the selection of $m$ and $r$. The moments of $X_\theta \sim f_\theta$ are estimated by $\widehat{\mathbb{E}\,X_\theta} = \widehat{\mathscr{L}_1}(\theta)/\widehat{\mathscr{L}_0}(\theta)$ and $\widehat{\mathbb{V}\mathrm{ar}\,X_\theta} = \widehat{\mathscr{L}_2}(\theta)/\widehat{\mathscr{L}_0}(\theta) - \widehat{\mathbb{E}\,X_\theta}^2$ where the approximation uses Gauss–Hermite quadrature (3.31); for this we use $H = 64$, 32, 16 for $n = 2$, 3, 4 respectively (and CMC for $n > 4$).

With these regimes, parameter selection for the reference distributions is automatic, and the only choice the user must make is in selecting $K$. In these tests we examined various $K$ from 1 to 40, and show the best approximations found. The source code for these tests is available online at [16], and we invite readers to experiment with the effect of modifying $K$, $\theta$ and the parameters of the reference distributions.

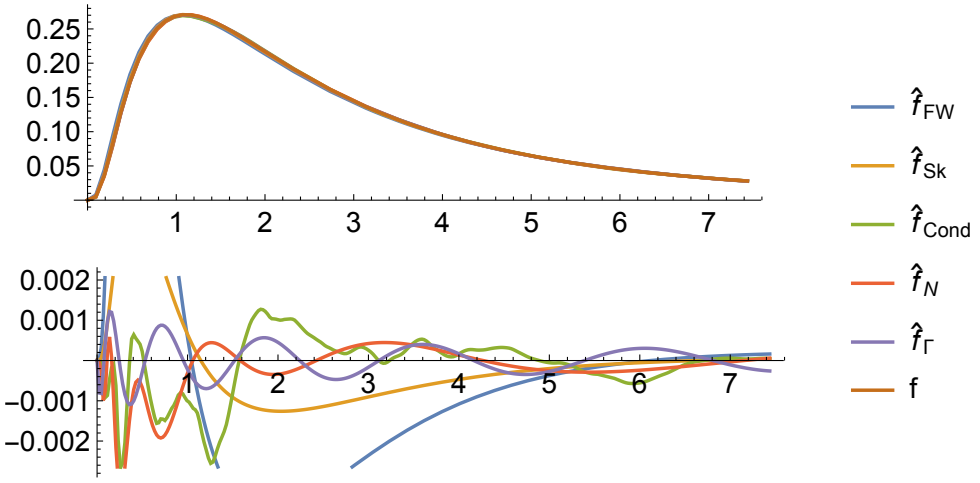
### 3.4.2 Results

For each test case with $n \leq 4$ we plot the $\widehat{f}(x)$ and $f(x)$ together and then $\widehat{f}(x) - f(x)$ over $x \in (0, 2\,\mathbb{E}[S])$. A table then shows the $\mathcal{L}^2$ errors over $(0, \mathbb{E}[S])$.



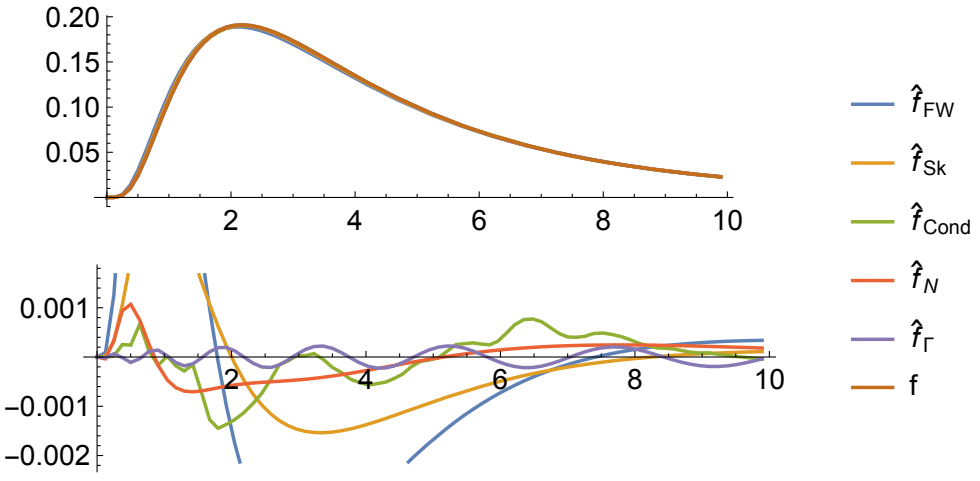

| | $\widehat{f}_{\mathrm{FW}}$ | $\widehat{f}_{\mathrm{Sk}}$ | $\widehat{f}_{\mathrm{Cond}}$ | $\widehat{f}_{\mathrm{N}}$ | $\widehat{f}_{\Gamma}$ |
|---|---|---|---|---|---|
| $\mathcal{L}^2$ | $8.01\times10^{-2}$ | $4.00\times10^{-2}$ | $1.56\times10^{-3}$ | $1.94\times10^{-3}$ | $2.28\times10^{-3}$ |

Test 1: $\boldsymbol{\mu} = (0, 0)$, $\mathrm{diag}(\boldsymbol{\Sigma}) = (0.5, 1)$, $\rho = -0.2$. Reference distributions used are $\mathsf{Normal}(0.88, 0.71^2)$ and $\mathsf{Gamma}(2.43, 0.51)$ with $K = 32$, 16 resp.



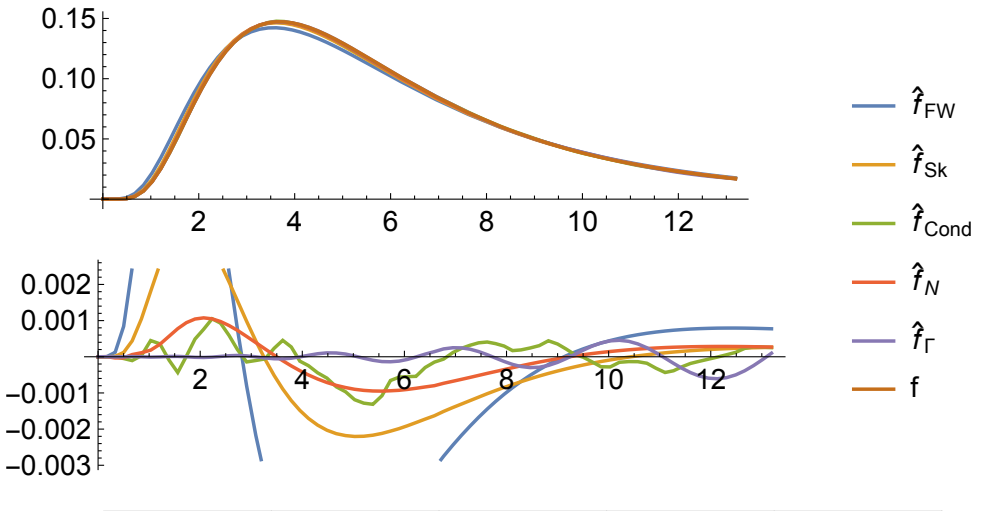

| | $\widehat{f}_{\mathrm{FW}}$ | $\widehat{f}_{\mathrm{Sk}}$ | $\widehat{f}_{\mathrm{Cond}}$ | $\widehat{f}_{\mathrm{N}}$ | $\widehat{f}_{\Gamma}$ |
|---|---|---|---|---|---|
| $\mathcal{L}^2$ | $1.02\times10^{-2}$ | $3.49\times10^{-3}$ | $1.78\times10^{-3}$ | $7.86\times10^{-4}$ | $7.24\times10^{-4}$ |

Test 2: $\boldsymbol{\mu} = (-0.5, 0.5)$, $\mathrm{diag}(\boldsymbol{\Sigma}) = (1, 1)$, $\rho = 0.5$. Reference distributions used are $\mathsf{Normal}(0.91, 0.90^2)$ and $\mathsf{Gamma}(2.35, 0.51)$ with $K = 32$, 16 resp.




| | $\widehat{f}_{\mathrm{FW}}$ | $\widehat{f}_{\mathrm{Sk}}$ | $\widehat{f}_{\mathrm{Cond}}$ | $\widehat{f}_{\mathrm{N}}$ | $\widehat{f}_{\Gamma}$ |
|---|---|---|---|---|---|
| $\mathcal{L}^2$ | $9.48\times10^{-3}$ | $3.71\times10^{-3}$ | $1.60\times10^{-3}$ | $1.18\times10^{-3}$ | $3.53\times10^{-4}$ |

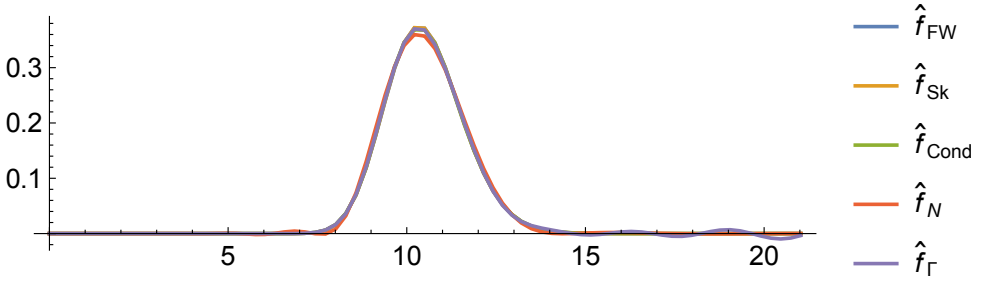
Test 3: $n = 3$, $\mu_i = 0$, $\Sigma_{ii} = 1$, $\rho = 0.25$. Reference distributions used are $\mathsf{Normal}(1.32, 0.74^2)$ and $\mathsf{Gamma}(3, 0.57)$ with $K = 7$, 25 resp.




| | $\widehat{f}_{\mathrm{FW}}$ | $\widehat{f}_{\mathrm{Sk}}$ | $\widehat{f}_{\mathrm{Cond}}$ | $\widehat{f}_{\mathrm{N}}$ | $\widehat{f}_{\Gamma}$ |
|---|---|---|---|---|---|
| $\mathcal{L}^2$ | $1.82\times10^{-2}$ | $6.60\times10^{-3}$ | $1.90\times10^{-3}$ | $1.80\times10^{-3}$ | $1.77\times10^{-4}$ |

Test 4: $n = 4$, $\mu_i = 0$, $\Sigma_{ii} = 1$, $\rho = 0.1$. Reference distributions used are $\mathsf{Normal}(1.32, 0.74^2)$ and $\mathsf{Gamma}(3.37, 0.51)$ with $K = 18$, 18 resp.

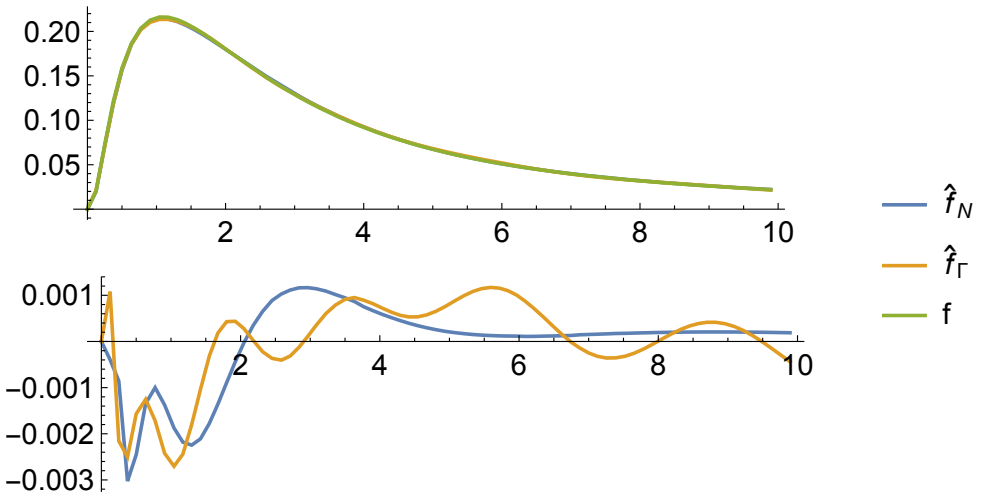
The following test case shows the density approximations for a large $n$.




Test 5: Sum of 10 iid $\mathsf{Lognormal}(0, 0.1)$ random variables. Reference distributions used are $\mathsf{Normal}(2.35, 0.23^2)$ and $\mathsf{Gamma}(12.61, 0.25)$ with $K = 18$, 35 resp.

Finally, we fit $\widehat{f}_{\mathrm{N}}$ and $\widehat{f}_{\Gamma}$ to simulated data ($10^5$ replications) for the sum of lognormals with a non-Gaussian dependence structure. Specifically, we take the sum of $n = 3$ standard lognormal random variables with a *Clayton copula*, $\mathsf{Clayton}(\theta)$, defined by its distribution function

$$C_\theta^{\mathrm{Cl}}(u_1, \ldots, u_n) = \left(1 - n + \sum_{i=1}^{n} u_i^{-\theta}\right)^{-1/\theta}, \qquad \text{for } \theta > 0\,.$$




Test 6: Sum of 3 Lognormal$(0,1)$ random variables with Clayton$(10)$ copula (i.e., $\tau = \frac{5}{6}$). Reference distributions used are Normal$(1.46, 0.71^2)$ and Gamma$(8.78, 0.25)$ with $K = 40$. The $\mathcal{L}^2$ errors of $\widehat{f}_{\mathrm{N}}$ and $\widehat{f}_{\Gamma}$ are $2.45 \times 10^{-3}$ and $2.04 \times 10^{-3}$ respectively.

The Kendall's tau correlation of the $C_\theta^{\mathrm{Cl}}$ copula is $\tau = \theta/(\theta+2)$ [123].

Our overall conclusion of the numerical examples is that no single method can be considered as universally superior. Of the methods in the literature, the log skew normal approximations is generally better than Fenton-Wilkinson, which is unsurprising given it is an extension introducing one more parameter. The estimators, $\widehat{f}_{\mathrm{N}}$ and $\widehat{f}_{\Gamma}$, based on orthogonal polynomial approximation techniques, are very flexible. They also display at least as good and sometimes better pdf estimates over the interval $(0, \mathbb{E}[S])$ and their periodic error indicates that they would supply even more accurate cdf estimates. One should note, however, that their performance relies on the tuning of parameters and that somewhat greater effort is involved in their computation (though this is mitigated through the availability of the software in [16]).

An interesting feature of $\widehat{f}_{\mathrm{N}}$ and $\widehat{f}_{\Gamma}$ is that the Clayton copula example indicates some robustness to the dependence structure used. In view of the current interest in financial applications of non-Gaussian dependence this seems a promising line for future research.

## 3.A Proof of Proposition 3.3

*Proof.* The polynomials which are orthogonal with respect to the lognormal distribution will be derived using the general formula (1.12). The moments of the $\mathsf{Lognormal}(\mu, \sigma^2)$ distribution are given by $m_n = p^n q^{n^2}$, where $p = \mathrm{e}^{\mu}$ and $q = \mathrm{e}^{\frac{\sigma^2}{2}}$. Consider

$$\boldsymbol{H}_n = \begin{pmatrix} 1 & pq & \cdots & p^n q^{n^2} \\ pq & p^2 q^4 & & p^{n+1} q^{(n+1)^2} \\ \vdots & \vdots & & \vdots \\ p^{n-1} q^{(n-1)^2} & p^n q^{n^2} & \cdots & p^{2n-1} q^{(2n-1)^2} \\ p^n q^{n^2} & p^{n+1} q^{(n+1)^2} & \cdots & p^{2n} q^{(2n)^2} \end{pmatrix}, \quad n \in \mathbb{N}_+ , \tag{3.23}$$

and denote by $R_k$ the $k$-th row and by $C_\ell$ the $\ell$-th column. Apply the elementary operations $R_{k+1} \to p^{-k} q^{-k^2} R_{k+1}$, and $C_{\ell+1} \to p^{-\ell} q^{-\ell^2} C_{\ell+1}$ for $k, \ell = 0, \ldots, n$ to get

$$\boldsymbol{A}_n = \begin{pmatrix} 1 & \alpha_0 & \cdots & \alpha_0^n \\ 1 & \alpha_1 & \cdots & \alpha_1^n \\ \vdots & \vdots & \vdots & \\ 1 & \alpha_{n-1} & \cdots & \alpha_{n-1}^n \\ 1 & \alpha_n & \cdots & \alpha_n^n \end{pmatrix}, \quad n \in \mathbb{N}_+ \tag{3.24}$$

where $\alpha_i = q^{2i}$. As $\boldsymbol{A}_n$ is a Vandermonde matrix, $\det(\boldsymbol{A}_n) = \prod_{i=0}^{n} \prod_{j=i+1}^{n} (\alpha_j - \alpha_i)$, and

$$\det(\boldsymbol{H}_n) = p^{n(n+1)} q^{\frac{n(n+1)(2n+1)}{3}} \det(\boldsymbol{A}_n) \tag{3.25}$$

$$= p^{n(n+1)} q^{\frac{n(n+1)(2n+1)}{3}} \prod_{i=1}^{n} (-1)^i q^{2(n-i)i} [q^2; q^2]_i \, . \tag{3.26}$$

Next, expand $\det(\widetilde{\boldsymbol{H}}_n(x))$ with respect to the last row to get

$$p_n(x) = \frac{1}{\sqrt{\det(\boldsymbol{H}_{n-1}) \det(\boldsymbol{H}_n)}} \sum_{k=0}^{n} (-1)^{n+k} \det(\boldsymbol{C}_{n,k}) x^k, \tag{3.27}$$

where $\boldsymbol{C}_{n,k}$ is $k$-th co-factor of $\boldsymbol{H}_n$ (i.e., $\boldsymbol{H}_n$ with the last row and the $(k+1)$-th column removed). Perform the same elementary operations as before on $\boldsymbol{C}_{n,k}$ to get

$$\det(\boldsymbol{C}_{n,k}) = p^{n^2-k} q^{\frac{2n^3+n}{3}-k^2} \det(\boldsymbol{A}_{n,k}) \tag{3.28}$$

where

$$\boldsymbol{A}_{n,k} = \begin{pmatrix} 1 & \alpha_0 & \cdots & \alpha_0^{k-1} & \alpha_0^{k+1} & \cdots & \alpha_0^n \\ 1 & \alpha_1 & \cdots & \alpha_1^{k-1} & \alpha_1^{k+1} & \cdots & \alpha_1^n \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \alpha_{n-1} & \cdots & \cdots & \cdots & \cdots & \alpha_{n-1}^n \end{pmatrix}.$$

Using the definition of *Schur polynomials*, it is clear that

$$\det(\boldsymbol{A}_{n,k}) = s_{\lambda(k)}(\alpha_0, \ldots, \alpha_{n-1}) \det(\boldsymbol{A}_{n-1})$$

where $\lambda(k) = (\mathbf{1}_{n-k}, \mathbf{0}_k)$. With these $\lambda(k)$, the Schur polynomials simplify to the *elementary symmetric polynomials*, so

$$\det(\boldsymbol{A}_{n,k}) = e_{n-k}(\alpha_0, \ldots, \alpha_{n-1}) \det(\boldsymbol{A}_{n-1}). \tag{3.29}$$

Combining (3.28), (3.29), and (3.25) yields

$$\begin{aligned} \det(\boldsymbol{C}_{n,k}) &= p^{n^2-k} q^{\frac{2n^3+n}{3}-k^2} e_{n-k}(\alpha_0, \ldots, \alpha_{n-1}) \det(\boldsymbol{A}_{n-1}) \\ &= p^{n^2-k} q^{\frac{2n^3+n}{3}-k^2} e_{n-k}(\alpha_0, \ldots, \alpha_{n-1}) p^{-n^2+n} q^{-\frac{(n-1)n(2n-1)}{3}} \det(\boldsymbol{H}_{n-1}) \\ &= p^{n-k} q^{n^2-k^2} e_{n-k}(\alpha_0, \ldots, \alpha_{n-1}) \det(\boldsymbol{H}_{n-1}). \end{aligned}$$

So, substituting this into the (3.27) gives

$$\begin{aligned} p_n(x) &= \frac{1}{\sqrt{\det(\boldsymbol{H}_{n-1}) \det(\boldsymbol{H}_n)}} \sum_{k=0}^{n} (-1)^{n+k} p^{n-k} q^{n^2-k^2} e_{n-k}(\alpha_0, \ldots, \alpha_{n-1}) \det(\boldsymbol{H}_{n-1}) x^k \\ &= \sqrt{\frac{\det(\boldsymbol{H}_{n-1})}{\det(\boldsymbol{H}_n)}} p^n q^{n^2} \sum_{k=0}^{n} (-1)^{n+k} p^{-k} q^{-k^2} e_{n-k}(\alpha_0, \ldots, \alpha_{n-1}) x^k. \end{aligned}$$

The constant $\det(\boldsymbol{H}_{n-1})/\det(\boldsymbol{H}_n)$ can be handled using (3.26)

$$\frac{\det(\boldsymbol{H}_{n-1})}{\det(\boldsymbol{H}_n)} = p^{-2n}q^{-2n^2}\frac{q^{-n(n-1)}}{(-1)^n[q^2;q^2]_n} = \frac{p^{-2n}q^{-3n^2+n}}{|[q^2;q^2]_n|}\,.$$

Finally, simplify this constant using $q^{n+n^2}/|[q^2;q^2]_n| = 1/[q^{-2};q^{-2}]_n$ to get (3.11). □

## 3.B Computing the coefficients of the expansion $\{a_k\}_{k\in\mathbb{N}_0}$ in the gamma case

We extend here the above techniques to construct an approximation for $\mathscr{L}_i(\theta)$. We note that $\mathscr{L}_i(\theta) \propto \int_{\mathbb{R}^n} \exp\{-h_{\theta,i}(\boldsymbol{x})\}\,\mathrm{d}\boldsymbol{x}$ where

$$h_{\theta,i}(\boldsymbol{x}) = -i\ln(\mathbf{1}^\top \mathrm{e}^{\boldsymbol{\mu}+\boldsymbol{x}}) + \theta\mathbf{1}^\top \mathrm{e}^{\boldsymbol{\mu}+\boldsymbol{x}} + \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\,, \quad i\in\mathbb{N}_0\,.$$

This uses the notation of the previous chapter, where for example, $\mathrm{e}^{\boldsymbol{x}} = (\mathrm{e}^{x_1},\dots,\mathrm{e}^{x_n})^\top$. Next, define $\boldsymbol{x}^*$ as the minimiser of $h_{\theta,i}$ (calculated numerically), and consider a second order Taylor expansion of $h_{\theta,i}$ about $\boldsymbol{x}^*$. Denote $\widetilde{\mathscr{L}_i}(\theta)$ as the approximation of $\mathscr{L}_i(\theta)$ in which $h_{\theta,i}$ is replaced by this Taylor expansion. Simplifying yields

$$\widetilde{\mathscr{L}_i}(\theta) = \frac{\exp\{-h_{\theta,i}(\boldsymbol{x}^*)\}}{\sqrt{|\det(\boldsymbol{\Sigma}\boldsymbol{H})|}} \tag{3.30}$$

where $\boldsymbol{H}$, the Hessian of $h_{\theta,i}$ evaluated at $\boldsymbol{x}^*$, is

$$\boldsymbol{H} = i\frac{\mathrm{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*}(\mathrm{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*})^\top}{(\mathbf{1}^\top\mathrm{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*})^2} + \boldsymbol{\Sigma}^{-1} - \mathrm{diag}(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^*)\,.$$

As $\theta\to\infty$ we have $\widetilde{\mathscr{L}_i}(\theta)\to\mathscr{L}_i(\theta)$. We rewrite $\mathscr{L}_i(\theta) = \widetilde{\mathscr{L}_i}(\theta)I_i(\theta)$ and estimate $I_i(\theta)$ as follows.

**Proposition 3.11.** *The moments of the* $\mathsf{SumLognormal}_\theta(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *distribution, denoted* $\mathscr{L}_i(\theta)$,

*can be written as* $\mathscr{L}_i(\theta) = \widetilde{\mathscr{L}_i}(\theta) I_i(\theta)$, *where* $\widetilde{\mathscr{L}_i}(\theta)$ *is in* (3.30), *and*

$$I_i(\theta) = \sqrt{|\det(\boldsymbol{\Sigma}\boldsymbol{H})|}\, v(\mathbf{0})^{-1}\, \mathbb{E}[v(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z})]$$

*for* $\boldsymbol{Z} \sim \mathsf{Normal}(\mathbf{0}, \boldsymbol{I})$, *and*

$$v(\boldsymbol{z}) = \exp\{i \ln(\mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*+\boldsymbol{z}}) - \theta \mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*+\boldsymbol{z}} - (\boldsymbol{x}^*)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{z}\}\,.$$

*Proof.* Substitute $\boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y}$ into $\mathscr{L}_i(\theta)$, then multiply by $\exp\{\pm \text{ some constants}\}$:

$$\begin{aligned}
\mathscr{L}_i(\theta) &= \int_{\mathbb{R}^n} \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{|\det(\boldsymbol{\Sigma})|}} \exp\{i \log(\mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}}) - \theta \mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}} - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}\}\, \mathrm{d}\boldsymbol{x} \\
&= \widetilde{\mathscr{L}_i}(\theta) \exp\{-i \log(\mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*}) + \theta \mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*}\} \\
&\quad \times \int_{\mathbb{R}^n} (2\pi)^{-\frac{n}{2}} \exp\{i \log(\mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*+\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y}}) - \theta \mathbf{1}^\top \mathbf{e}^{\boldsymbol{\mu}+\boldsymbol{x}^*+\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y}} \\
&\qquad - (\boldsymbol{x}^*)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^\top (\boldsymbol{\Sigma}\boldsymbol{H})^{-1}\boldsymbol{y}\}\, \mathrm{d}\boldsymbol{y}\,.
\end{aligned}$$

That is, $\mathscr{L}_i(\theta) = \widetilde{\mathscr{L}_i}(\theta) I_i(\theta)$. In $I_i(\theta)$, take the change of variable $\boldsymbol{y} = (\boldsymbol{\Sigma}\boldsymbol{H})^{\frac{1}{2}}\boldsymbol{z}$, and the result follows. □

**Remark 3.12.** The form of $I_i(\theta)$ naturally suggests evaluation using *Gauss–Hermite* quadrature:

$$\widehat{\mathscr{L}_i}(\theta) = \frac{\exp\{-h_{\theta,i}(\boldsymbol{x}^*)\}}{v(\mathbf{0})\, \pi^{n/2}} \sum_{i_1=1}^{H} \cdots \sum_{i_n=1}^{H} v(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z}) \prod_{j=1}^{n} w_{i_j} \tag{3.31}$$

where $\boldsymbol{z} = (z_{i_1}, \ldots, z_{i_n})^\top$, the set of weights and nodes $\{(w_i, z_i) : 1 \le i \le H\}$ is specified by the Gauss–Hermite quadrature algorithm, and $H \ge 1$ is the order of the approximation. This approximation is accurate, especially so when the $i$ in $\mathscr{L}_i$ becomes large. Even for $\mathscr{L}$ $(= \mathscr{L}_0)$ this method appears to outperform the quasi-Monte Carlo scheme outlined in Chapter 2. ◇

Thus, with $\widehat{\mathscr{L}_i}(\theta)$ given in (3.31), we can now estimate the coefficients. The three methods correspond to

1. $\widehat{a}_k = R^{-1} \sum_{r=1}^{R} p_k(S_r)$, for $S_1, \ldots, S_R \overset{\text{iid}}{\sim} f_\theta(x)$,

2. $\widehat{a}_k = \sum_{j=0}^{k} q_{kj} \widehat{\mathbb{E}[S_\theta^j]} = q_{k0} + (R\,\widehat{\mathscr{L}}(\theta))^{-1} \sum_{j=1}^{k} q_{kj} \sum_{r=1}^{R} S_r^j \mathrm{e}^{-\theta S_r}$, from (3.22), where $S_1, \ldots, S_R \overset{\text{iid}}{\sim} f(x)$,

3. $\widehat{a}_k = q_{k0} + \widehat{\mathscr{L}}(\theta)^{-1} \sum_{j=1}^{k} q_{kj}\, \widehat{\mathscr{L}_j}(\theta)$.

In the numerical illustrations, we switched between using methods (2) and (3) for large and small $n$ respectively. Algorithms for efficient simulation from $f_\theta$ is work in progress.

## Chapter 4 Authorship Statement

**Citation:** Pierre-Olivier Goffard, Patrick J. Laub (2017), *Two numerical methods to evaluate stop-loss premiums*, Scandinavian Actuarial Journal (submitted)

The authors of this paper equally contributed to the following tasks:

1. conception and design of the project;
2. mathematical arguments, and interpretation of the results;
3. writing the publication.

In addition to this, I completed the majority of the computational work and of the editing (e.g. checking grammar and typographical details).

# Chapter 4

# Two numerical methods to evaluate stop-loss premiums

## 4.1 Introduction

Consider the random variable

$$S_N = \sum_{k=1}^{N} U_k,$$

where $N$ is a counting random variable and $\{U_k\}_{k\in\mathbb{N}_0}$ is a sequence of random variables which are iid, non-negative, and independent of $N$. We denote the pdf of $S_N$ as $f_{S_N}$, and its survival function as

$$\overline{F}_{S_N}(x) = \mathbb{P}(S_N > x), \quad \text{for } x \geq 0.$$

This chapter concerns approximations of $f_{S_N}$ and $\overline{F}_{S_N}$ though we begin with a discussion of how $S_N$ is used in actuarial science.

Frequently $S_N$ models the aggregated losses of a non-life insurance portfolio over a given period of time—here $N$ represents the number of claims and $U_k$ the claim sizes—yet other applications also exist. Actuaries and risk managers typically want to quantify the risk of large losses by a single comprehensible number, a risk measure.

One popular risk measure is the VaR. In actuarial contexts, the VaR at level $\alpha \in (0,1)$ is defined such that the probability of (aggregated) losses exceeding the level VaR is at most $1-\alpha$. Following the European recommendation of the Solvency II directive, the standard value for $\alpha$ is 0.995, see [101]. It is used by risk managers in banks, insurance companies, and other financial institutions to allocate risk reserves and to determine solvency margins. Also, we have stop-loss premiums which are risk measures that are commonly used in reinsurance agreements.

A reinsurance agreement is a common risk management contract between insurance companies, one called the *cedant* and the other the *reinsurer*. Its aim is to keep the cedant's long-term earnings stable by protecting the cedant against large losses. The reinsurer absorbs part of the cedant's loss, say $f(S_N)$ where $0 \leq f(S_N) \leq S_N$, leaving the cedant with $I_f(S_N) = S_N - f(S_N)$. In return, the cedant pays a premium linked to

$$\Pi = \mathbb{E}[f(S_N)],$$

under the expected value premium principle.

In practice, there are a variety of reinsurance designs from which an insurer can choose. We focus in this work on the stop-loss reinsurance treaty associated with the following ceded loss function

$$f(S_N) = (S_N - a)_+, \ a \geq 0,$$

where $a$ is referred to as the retention level or priority. The ratemaking of the stop-loss reinsurance policy requires the evaluation of

$$\Pi_a(S_N) = \mathbb{E}\left[(S_N - a)_+\right], \tag{4.1}$$

also known as the usual stop loss premium.

One variation is the limited stop-loss function,

$$f(S_N) = \min[(S_N - a)_+, b], \ b \geq 0, \tag{4.2}$$

where $b$ is called the limit. The limited stop-loss function (4.2) is very appealing in practice because it prevents the cedant from over-estimating their losses and therefore

over-charging the reinsurer. Also, the change-loss function is defined as

$$f(S_N) = c(S_N - a)_+,\ 0 \leq c \leq 1,$$

which is in between stop-loss and quota-share reinsurance. The ratemaking in each case requires the expectation in (4.1).

From a practical point of view, a reinsurance treaty over the whole portfolio is less expensive to handle than one which involves claim-by-claim management. It also grants protection in the event of an unusual number of claims, triggered for instance by a natural disaster. From a theoretical point of view, it is well known that the stop-loss ceded function allows one to minimise the variance of the retained loss for a given premium level, see for instance the monograph of Denuit et al. [59]. Recently, it has been shown that stop-loss reinsurance is also optimal when trying to minimise the VaR and the expected shortfall of the retained loss, see the works of Cai et al. [40], Cheung [46], and Chi and Tan [47]. Note that some other ceded loss functions appear in their work, there are however very close to the stop-loss one.

Unfortunately, one is seriously constrained when calculating these quantities analytically, as there are only a few cases where either the pdf or the survival function is available in a simple tractable form. To estimate the VaR or the stop-loss premium we must find fast and accurate approximations for these functions.

We discuss the use of an approximation of the pdf in terms of the gamma density and its orthonormal polynomials. This method has been studied in the recent works of Goffard et al. [83] and Jin et al. [102]. We emphasise here the computational aspect of this numerical method and detail some practical improvements. An exponential change of measure can be used to recover the pdf of $S_N$ when the claim sizes are governed by a heavy-tailed distribution. This refinement has been successfully applied in Chapter 3 to recover the density of the sum of lognormally distributed random variables.

This method is compared to a numerical inversion of the Laplace transform which is known to be efficient to recover the survival function of a compound distribution. The critical step in Laplace inversion is to select which numerical integration technique to apply. We implement a method inspired by the work of Abate and Whitt [2] which is

very similar to the method of Rolski et al. [147, Chapter 5, Section 5]. An approximation of the stop-loss premium is then proposed relying on the connection with the survival function of the equilibrium distribution of $S_N$. Note that Dufresne et al. [66] successfully applied a Laplace inversion based technique to the evaluation of stop-loss premiums.

To close, we want to emphasise the fact that the numerical methods also apply in a risk theory framework. The infinite-time ruin probability in the compound Poisson ruin model is equal to the survival function of a compound geometric distribution. The polynomial approximation and the Laplace inversion methods have been employed, and compared to solve this particular problem in the work of Goffard et al. [84]. We add a more original application by noting that the finite-time non-ruin probability with no initial reserves, again under the classical risk model assumptions, may be rewritten as the stop-loss premium associated with a compound Poisson distribution where the priority is expressed in terms of the premium rate and the time horizon.

The rest of the chapter is organised as follows. Section 4.2 introduces compound distributions, and details their role in risk theory. Section 4.3 presents the approximation method based on orthogonal polynomials. Section 4.4 presents the approximation through the numerical inversion of the Laplace transform. Section 4.5 is devoted to numerical illustrations where the performances of the two methods are compared; the MATHEMATICA code used is available online [82].

## 4.2 Compound distributions and risk theory

We introduce compound distributions along with a brief account of their importance in risk modeling.

### 4.2.1 Compound distributions

Let $S_N = \sum_{k=1}^{N} U_k$ be the aggregated claim amounts associated with a non-life insurance portfolio over a fixed time period. The number of claims, also called the claim frequency, is modeled by a counting random variable $N$ having a probability mass function $f_N$. The

claim sizes form a sequence $\{U_k\}_{k\in\mathbb{N}_0}$ of iid non-negative random variables with common pdf $f_U$. We further assume that the claim sizes are independent from the claim frequency.

As $S_N = 0$ whenever $N = 0$ (assuming this occurs with positive probability), the distribution of $S_N$ is the sum of a singular part (the probability mass $\mathbb{P}(S_N = 0) = f_N(0) > 0$) and a continuous part (describing $S_N$ where $N > 0$) with a defective pdf $f^+_{S_N}$ and cdf $F^+_{S_N}$. From the law of total probability, we have

$$f^+_{S_N}(x) = \sum_{n=1}^{\infty} f_N(n) f_U^{*n}(x),\ x \geq 0. \tag{4.3}$$

This density is intractable because of the infinite series. Furthermore, the summands are defined by repeated convolution of $f_U$ with itself which are rarely straightforward to evaluate. The methods presented in this work rely on the knowledge of the Laplace transform of $S_N$, given by

$$\mathscr{L}_{S_N}(t) = \mathcal{G}_N[\mathscr{L}_U(t)]\,,$$

where $\mathcal{G}_N(t) := \mathbb{E}[t^N]$ is the *probability generating function* of $N$. The simple expression of the Laplace transform has made possible the use of numerical methods involving the moments or transform inversion to recover the distribution of $S_N$. The distribution is typically recovered using Panjer's algorithm or a Fast Fourier Transform algorithm based on the inversion of the discrete Fourier transform; these two methods are compared in the work of Embrechts and Frei [69]. Our orthogonal polynomial method involves the standard integer moment sequence for $S_N$, in contrast to more exotic types of moments used by some recent methods. Gzyl and Tagliani [87] uses the fractional moments within a max-entropic based method, while Mnatsakanov and Sarkisian [127] performs an inversion of the scaled Laplace transform via the exponential moments. In addition to proposing an approximation for the survival function of $S_N$, we provide an efficient way to compute the usual stop-loss premium (4.1) for reinsurance applications.

### 4.2.2 Risk theory

In the classical risk model, the financial reserves of a non-life insurance company are modeled by the risk reserve process $\{R(t), t \geq 0\}$, defined as

$$R(t) = u + ct - \sum_{k=1}^{N(t)} U_k.$$

The insurance company holds an initial capital of amount $R(0) = u \geq 0$, and collects premiums at a constant rate of $c > 0$ per unit of time. The number of claims up to time $t \geq 0$ is governed by a homogeneous Poisson process $\{N(t), t \geq 0\}$ with intensity $\lambda$. The claim sizes are iid non-negative random variables independent from $N(t)$.

One of the goals of risk theory to evaluate an insurer's ruin probability, that is, the probability that the financial reserves eventually fall below zero. Of interest are both the finite-time ruin probability $\psi(u, T)$ and the infinite-time ruin probability, also called the *probability of ultimate ruin*, $\psi(u)$, which are defined as

$$\psi(u, T) = \mathbb{P}\Big(\inf_{0 \leq t \leq T} R(t) \leq 0\Big),$$

and

$$\psi(u) = \mathbb{P}\Big(\inf_{t \geq 0} R(t) \leq 0\Big).$$

These probabilities are often reformulated (for mathematical convenience) in terms of the associated claims surplus process $\{S(t), t \geq 0\}$,

$$S(t) = \sum_{k=1}^{N(t)} U_k - ct, \ t \geq 0,$$

specifically,

$$\psi(u, T) = \mathbb{P}\Big(\sup_{0 \leq t \leq T} S(t) \geq u\Big) \quad \text{and} \quad \psi(u) = \mathbb{P}\Big(\sup_{t \geq 0} S(t) \geq u\Big).$$

For a general background on risk theory and the evaluation of ruin probabilities, we refer the reader to the monograph of Asmussen and Albrecher [14].

The first connection between compound distributions and ruin probabilities is the following. If the net benefit condition is satisfied, i.e. if the premium rate exceeds the average cost of aggregated claims per unit of time, then the infinite-time ruin probability is given by the survival function of a geometric compound distribution. More precisely,

$$\psi(u) = \mathbb{P}\Big(S_N := \sum_{k=1}^{N} U_k^* > u\Big) = (1-\rho)\sum_{n=1}^{\infty} \rho^n \overline{F}_{U^*}^{*n}(u),$$

with $N \sim \mathsf{Geom}_0(\rho)$, $\rho = \lambda\,\mathbb{E}[U]/c < 1$, and with iid $U_k^*$ with pdf $f_{U^*}(x) = \overline{F}_U(x)/\,\mathbb{E}[U]$. This result is known as the Pollaczeck–Khinchine formula, see for instance Asmussen and Albrecher [14, Chapter IV, (2.2)]. Thus it is possible to evaluate the infinite-time ruin probability via Panjer's algorithm. If we are able to determine the Laplace transform of $V$ then we can also apply the polynomial method of Goffard et al. [83], the fractional moment based method of Gzyl et al. [86], and the exponential moments based method of Mnatsakanov et al. [128].

The second connection links the finite-time ruin probability with no initial reserves to the stop-loss premium associated with a compound distribution. If $N(t) \sim \mathsf{Poisson}(\lambda t)$ (i.e. claims arrive as a homogeneous Poisson process) then the finite-time ruin probability is given by

$$\psi(0,T) = 1 - \frac{1}{cT}\int_0^{cT} \mathbb{P}\Big(\sum_{i=1}^{N(T)} U_i \le x\Big)\,\mathrm{d}x\,. \tag{4.4}$$

This implies $\psi(0,T) = \mathbb{E}[\min(S_{N(T)}, cT)]/cT$ where $S_{N(T)} := \sum_{i=1}^{N(T)} U_i$, and hence

$$\begin{aligned}\psi(0,T) &= \frac{1}{cT}\,\mathbb{E}[\min(S_{N(T)}, cT)] \\ &= (cT)^{-1}\{\mathbb{E}[N(T)]\;\mathbb{E}[U_1] - \Pi_{cT}(S_{N(T)})\}\,.\end{aligned} \tag{4.5}$$

Lefèvre and Picard [117, Corollary 4.3] show that equations (4.4) and (4.5) hold in the more general case where the claim arrival process forms a *mixed Poisson process*. This connection has been exploited recently in Lefèvre et al. [118] where the influence of the claim size distribution on the ruin probabilities is studied via stochastic ordering considerations.

## 4.3 Orthogonal polynomial approximations

### 4.3.1 Approximating general density functions

Let $X$ be an arbitrary random variable with pdf $f_X$.[1] We assume that the density is unknown and we propose an approximation of the form

$$\widehat{f}_X(x) = \sum_{k=0}^{K} a_k p_k(x) w(x). \tag{4.6}$$

where $w$ is the *reference density*. The $\{p_k\}_{k\in\mathbb{N}_0}$ are the polynomials which are orthonormal with respect to $w$, just as in Chapters 1 and 3.

Just as earlier, we can generate the polynomials by the Gram–Schmidt procedure if $w$ admits moments of all orders (i.e. $\forall n \in \mathbb{N}_0 : \mathbb{E}[X^n] < \infty$ where $X \sim w$), and the resulting polynomials are complete in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ if $\int_{\mathbb{R}} \mathrm{e}^{s|x|} w(x)\,\mathrm{d}x < \infty$ for some $s > 0$.

Therefore, if $f_X/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ we have

$$f_X(x)/w(x) = \sum_{k=0}^{\infty} \langle f_X/w, p_k\rangle_w p_k(x)\,. \tag{4.7}$$

We label the coefficients as $a_k = \langle f_X/w, p_k\rangle_w = \mathbb{E}[p_k(X)]$ and rearrange (4.7) to be

$$f_X(x) = \sum_{k=0}^{\infty} a_k p_k(x) w(x). \tag{4.8}$$

The approximation (4.6) follows by simply truncating the series to $K+1$ terms.

Typical choices of reference distributions are ones that belong to a Natural Exponential Family with Quadratic Variance Function (NEF-QVF) which includes the normal, gamma, hyperbolic, Poisson, binomial, and Pascal distributions. This family of distributions is convenient as the associated orthogonal polynomials are classical, see the characterisation by Morris [129]. The polynomials are known explicitly, so the time-consuming Gram–Schmidt orthogonalisation procedure is unneccessary. Furthermore, it has been

[1] This section is written from the perspective of approximating a pdf, however the main results also hold if applied to a defective density.

shown in a paper by Provost [141] that the recovery of unknown densities from the knowledge of the moments of the distribution naturally leads to an approximation in terms of the gamma density and Laguerre polynomials when $X$ admits $\mathbb{R}_+$ as support, and in terms of the normal density and Hermite polynomials when $X$ has $\mathbb{R}$ as support.

### 4.3.2 Approximating densities of positive random variables

To approximate the pdf for positive $X$, a natural candidate for the reference density is the gamma density. It has been proven to be efficient in practice, see the work of Goffard et al. [83, 84], and Jin et al. [102]. The work of Papush et al. [138] showed that among the gamma, normal and lognormal distributions, the gamma distribution seems to be better suited to model certain aggregate losses. The lognormal distribution is a problematic choice. Even though the orthogonal polynomials are available in a closed form (see Section 3.A), Proposition 3.5 shows that they are incomplete in $\mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$. The case of the inverse Gaussian as basis received a treatment in the work of Nishii [133], where it is shown that the only way to get a complete system of polynomials is by using the Gram–Schmidt orthogonalisation procedure. Differentiating the density (as it is done in the case of NEF-QVF) does not lead to an orthogonal polynomial system, and starting from the Laguerre polynomials leads to a system of orthogonal functions which is not complete. A solution might be to exploit the bi-orthogonality property pointed out in the work of Hassairi and Zarai [92]. To close this review of reference densities, we mention the work of Nadarajah et al. [131] where Weibull and exponentiated exponential distributions are considered as reference density.

Consider $w$ to be the pdf of the $\mathsf{Gamma}(r, m)$ distribution and the associated orthonormal polynomials are given by

$$p_n(x) = (-1)^n \binom{n+r-1}{n}^{-\frac{1}{2}} L_n^{r-1}\Big(\frac{x}{m}\Big) = (-1)^n \left(\frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)}\right)^{-\frac{1}{2}} L_n^{r-1}\Big(\frac{x}{m}\Big),$$

where $\{L_n^{r-1}\}_{n\in\mathbb{N}_0}$ are the generalised Laguerre polynomials,

$$L_n^{r-1}(x) = \sum_{i=0}^{n} \binom{n+r-1}{n-i} \frac{(-x)^i}{i!} = \sum_{i=0}^{n} \frac{\Gamma(n+r)}{\Gamma(n-i+1)\Gamma(r+i)} \frac{(-x)^i}{i!},\ n \geq 0,$$

cf. the classical book by Szegö [158].

**Lemma 4.1.** *If $w$ is the pdf of the* $\mathsf{Gamma}(r, m)$ *distribution, the polynomial expansion* (4.8) *may be rewritten as*

$$f_X(x) = \sum_{i=0}^{\infty} c_i \gamma(r+i, m, x)\,, \tag{4.9}$$

*where*

$$c_i = \sum_{k=i}^{\infty} a_k \frac{(-1)^{i+k}}{i!\,(k-i)!} \sqrt{\frac{k!\Gamma(k+r)}{\Gamma(r)}}, \tag{4.10}$$

*and the function $\gamma(r, m, x)$ is the pdf of the* $\mathsf{Gamma}(r, m)$ *distribution.*

*Proof.* If we change the sum in (4.8) from iterating over Laguerre polynomials to iterating over monomials we get

$$f_X(x) = \sum_{k=0}^{\infty} a_k p_k(x)\gamma(r, m, x) = \sum_{i=0}^{\infty} b_i x^i \gamma(r, m, x)\,,$$

where

$$b_i = \sum_{k=0}^{\infty} \text{Coefficient}(x^i, a_k p_k(x)) = \frac{(-1)^i}{m^i i!} \sum_{k=i}^{\infty} a_k (-1)^k \binom{k+r-1}{k}^{-\frac{1}{2}} \binom{k+r-1}{k-i}.$$

We also note that

$$x^i \gamma(r, m, x) = m^i \frac{\Gamma(r+i)}{\Gamma(r)} \gamma(r+i, m, x),$$

so

$$f_X(x) = \sum_{i=0}^{\infty} b_i m^i \frac{\Gamma(r+i)}{\Gamma(r)} \gamma(r+i, m, x) = \sum_{i=0}^{\infty} c_i \gamma(r+i, m, x),$$

where we have set $c_i = b_i m^i \Gamma(r+i)/\Gamma(r)$. □

**Remark 4.2.** When $r = 1$ (that is, when $w(x)$ is the pdf of an exponential distribution) the formula for $c_i$, (4.10), simplifies to

$$c_i = \sum_{k=i}^{\infty} a_k (-1)^{i+k} \binom{k}{i}.$$

◇

The expression of the pdf in (4.9) resembles the one of an Erlang mixture, which are extensively used for risk modeling purposes, cf. Willmot and Woo [164], Lee and Lin [116], and Willmot and Lin [163]. However, the $c_i$ defined in (4.10) do not form a proper probability mass function as they are not always positive. Hence our approximation cannot be considered as an approximation through an Erlang mixture although it enjoys the same features when it comes to approximating the survival function and the stop-loss premium as shown in the following result.

**Proposition 4.3.** *Letting $\Gamma_u(r, m, x)$ be the survival function of the* $\mathsf{Gamma}(r, m)$ *distribution, we have:*

*(i) the survival function of $X$ is given by*

$$\overline{F}_X(x) = \sum_{i=0}^{\infty} c_i \Gamma_u(r+i, m, x) \quad \text{for } x \geq 0\,, \tag{4.11}$$

*(ii) the usual stop-loss premium of $X$ with priority $a \geq 0$ is given by*

$$\mathbb{E}\left[(X-a)_+\right] = \sum_{i=0}^{\infty} c_i \left[m(r+i)\Gamma_u(r+i+1, m, a) - a\Gamma_u(r+i, m, a)\right]. \tag{4.12}$$

*Proof.* If $f_X/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ then Lemma 4.1 allows us to write $f_X$ as in (4.9), and integrating this over $[x, \infty)$ yields the formula (4.11). Now consider the usual stop-loss premium of $X$, and note that

$$\begin{aligned}\mathbb{E}\left[(X-a)_+\right] &= \int_a^{\infty} (x-a) f_X(x)\,\mathrm{d}x \\ &= \int_a^{\infty} x f_X(x)\,\mathrm{d}x - a\overline{F}_X(a).\end{aligned} \tag{4.13}$$

Then notice that for every $i \in \mathbb{N}_+$, we have that

$$\begin{aligned}\int_a^{\infty} x\,\gamma(r+i, m, x)\,\mathrm{d}x &= \int_a^{\infty} x \frac{x^{r+i-1}\mathrm{e}^{-x/m}}{\Gamma(r+i)m^{r+i}}\,\mathrm{d}x \\ &= m\frac{\Gamma(r+i+1)}{\Gamma(r+i)} \int_a^{\infty} \frac{x^{r+i}\mathrm{e}^{-x/m}}{\Gamma(r+i+1)m^{r+i+1}}\,\mathrm{d}x \\ &= m(r+i)\Gamma_u(r+i+1, m, a).\end{aligned} \tag{4.14}$$

Therefore substituting (4.9) and (4.11) into (4.13) and simplifying with (4.14) yields (4.12). □

Let us make the connection between our approach and Erlang mixture more precise. Assuming that $f_X/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ then taking the Laplace transform on both side of (4.9) yields

$$\mathscr{L}_X(s) = \sum_{i=0}^{\infty} c_i \left(\frac{1}{1+sm}\right)^{r+i} = \left(\frac{1}{1+sm}\right)^{r} \mathcal{P}\left(\frac{1}{1+sm}\right),$$

where $\mathcal{P}(z) = \sum_{i=1}^{\infty} c_i z^i$ denotes the generating function of the $\{c_i\}_{i\in\mathbb{N}_0}$ coefficients. Now setting $z = \frac{1}{1+sm}$ allows to express the generating function $\mathcal{P}(z)$ in terms of the Laplace transform of $X$ as

$$\mathcal{P}(z) = z^{-r}\mathscr{L}_X\left(\frac{1-z}{zm}\right).$$

**Remark 4.4.** The approximation through an Erlang mixture consists in approximating the pdf of a nonnegative random variable $X$ as

$$f_X(x) = \sum_{i=1}^{\infty} c_i \gamma(i, m, x), \text{ for } x \geq 0.$$

The function $\mathcal{P}(z)$ becomes then the probability generating function (pgf) of a counting random variable $M$, where $c_i = \mathbb{P}(M = i), \text{ for } i \geq 1.$ ◇

The next example is designed to shed light on the link between our polynomial expansion and an Erlang mixture.

**Example 4.5.** Suppose that we are interested in approximating the pdf of an exponential random variable $\mathsf{Gamma}(1, \beta)$. The generating function of the coefficients is then

$$\mathcal{P}(z) = z^{1-r}\frac{m}{\beta + z(m-\beta)}.$$

If one takes $r = 1$ and $m = \beta$ then $\mathcal{P}(z) = 1$ and the polynomial representation reduces to the exponential pdf. Choosing $0 < m < \beta$ leads to $\mathcal{P}(z) = \frac{m/\beta}{1-z(1-m/\beta)}$, which is the pgf of a geometric random variable; this recovers the fact that an exponential random variable can be represented by a zero-truncated geometric sum of exponential random variables. For $m > \beta$, we have $\mathcal{P}(z) = \frac{m/\beta}{1+z(1-\beta/m)}$ which is an alternating sequence that decreases

geometrically fast. Recall that our polynomial expansion is valid only if $m > \beta/2$, which means that when $\beta/2 < m \leq \beta$ our approach coincides with the Erlang mixture technique. It does not when $m > \beta$. When $m \leq \beta/2$, the Erlang mixture representation holds even though the integrability condition, which is a sufficient one, does not hold.

The coefficients of the polynomials could be derived by differentiating the generating function $\mathcal{P}(z)$ as

$$c_i = \frac{1}{i!}\frac{\mathrm{d}^i}{\mathrm{d}z^i}\mathcal{P}(z)\Big|_{z=0} = \text{Coefficient}(i, \text{MaclaurinSeries}(\mathcal{P}(z))) \quad i \in \mathbb{N}_0\,.$$

In practice, the singularities of the function $\mathcal{P}(z)$ at zero mean this procedure is not viable. Instead, the $c_i$ are approximated by computing the $a_k$ and truncating their expression (4.10) up to a given order. The practical evaluation of the $a_k$ is discussed in Section 4.3.3.

A sufficient condition for $f_X/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ is

$$f_X(x) = \begin{cases} \mathcal{O}(\mathrm{e}^{-x/\delta}) & \text{as } x \to \infty \text{ with } m > \delta/2, \\ \mathcal{O}(x^\beta) & \text{as } x \to 0 \text{ with } r < 2(\beta+1). \end{cases} \tag{4.15}$$

When $X$ has a well-defined moment generating function one can typically choose $r$ and $m$ so this integrability condition is satisfied. When we consider heavy-tailed distributions, which is a desirable model characteristic in the applications, the integrability condition cannot be satisfied. The work-around is to use the expansion

$$\mathrm{e}^{-\theta x} f_X(x) = \sum_{k=0}^{\infty} a_k p_k(x) w(x),$$

for some $\theta > 0$. Thus, we can use

$$f_X(x) = \mathrm{e}^{\theta x} \sum_{k=0}^{\infty} a_k p_k(x) w(x) = \mathrm{e}^{\theta x} \sum_{i=0}^{\infty} c_i \gamma(r+i, m, x)$$

and since, when $1 - m\theta > 0$,

$$\mathrm{e}^{\theta x}\gamma(r+i, m, x) = (1-m\theta)^{-(r+i)}\gamma\Big(r+i, \frac{m}{1-m\theta}, x\Big)$$

we have

$$f_X(x) = \sum_{i=0}^{\infty} c_i (1-m\theta)^{-(r+i)} \gamma\Big(r+i, \frac{m}{1-m\theta}, x\Big) = \sum_{i=0}^{\infty} \widetilde{p}_i \gamma\Big(r+i, \widetilde{m}, x\Big),$$

where

$$\widetilde{p}_i = \frac{c_i}{(1-m\theta)^{r+i}} \quad \text{and} \quad \widetilde{m} = \frac{m}{1-m\theta}\,.$$

Calculating the $a_i$ and $c_i$, topic covered in Section 4.3.3, requires a Laplace transform of $\mathrm{e}^{-\theta x} f_X(x)$ which is given by

$$\mathscr{L}\{\mathrm{e}^{-\theta x} f_X(x)\}(t) = \mathscr{L}\{f_X(x)\}(t+\theta).$$

The method described above is the same (up to some constants) as approximating the exponentially tilted distribution. This idea has been used in Chapter 3. It is easily seen that taking $m > \theta^{-1}/2$ implies that $(\mathrm{e}^{-\theta x} f_X(x))/w(x) \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$.

### 4.3.3 Approximating densities of positive compound distributions

We now focus on variables $S_N$ which admit a compound distribution. Since these distributions have an atom at 0, we put aside this singularity and focus on the defective pdf $f^+_{S_N}$. The discussion in Sections 4.3.1 and 4.3.2 also applies to defective densities. Namely, if $f^+_{S_N}/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$ then the expansion in Lemma 4.1 is valid, we have

$$f^+_{S_N}(x) = \sum_{k=0}^{\infty} a_k p_k(x)\, \gamma(r, m, x) = \sum_{i=0}^{\infty} c_i \gamma(r+i, m, x), \text{ for } x > 0,$$

where $a_k = \int_0^\infty p_k(x) f^+_{S_N}(x)\,\mathrm{d}x$ and $c_i$ is given by (4.10). Truncating the first summation yields

$$f^+_{S_N}(x) \approx \sum_{k=0}^{K} a_k p_k(x)\, \gamma(r, m, x) = \sum_{i=0}^{K} \widehat{p}_i \gamma(r+i, m, x),$$

where $\widehat{p}_i = \sum_{k=i}^K a_k (-1)^{i+k}/[i!\,(k-i)!]\sqrt{k!\Gamma(k+r)/\Gamma(r)}$ for $i \le K$. The survival function $\overline{F}_{S_N}$ and the stop-loss premium $\mathbb{E}\left[(S_N - a)_+\right]$ follows from Proposition 4.3. If the inte-

grability condition is not satisfied then the exponentially tilted version of the defective pdf is expanded.

### Choice of $r$ and $m$

Firstly, we need to ensure that the choice of $r$ and $m$ satisfy the integrability condition (4.15). It is not simple to ensure this condition is met in general, so we look at the case of random sums of gamma distributed random variables. Define the radius of convergence of a random variable $X$ as

$$\rho_X := \sup\{s > 0,\ \mathscr{L}\{f_X\}(-s) < \infty\},$$

and consider the following result.

**Proposition 4.6.** *Let $S_N$ have summands where $U_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(r^*, m^*)$, then*

$$f^+_{S_N}(x) = \mathcal{O}(\exp\{-x\rho_{S_N}\}) \quad \textit{as } x \to \infty\,.$$

*Proof.* We have $f^+_{S_N}(x) = \sum_{n=1}^{\infty} f_N(n) f_{S_n}(x)$ and we also know that $S_n \sim \mathsf{Gamma}(nr^*, m^*)$. As $S_n$ is gamma distributed we have $f_{S_n}(x) \sim \frac{1}{m^*}\overline{F}_{S_n}(x)$ as $x \to \infty$. So applying the Chernoff inequality for each $S_n$ we have $\overline{F}_{S_n}(x) \leq \mathrm{e}^{-sx}\mathscr{L}_{S_n}(-s) = \mathrm{e}^{-sx}\mathscr{L}_U(-s)^n$ for all $s \in [0, \rho_U)$. Combining this we have

$$\begin{aligned} f^+_{S_N}(x) &\sim \frac{1}{m^*}\sum_{n=1}^{\infty} f_N(n)\overline{F}_{S_n}(x) \leq \frac{\mathrm{e}^{-sx}}{m^*}\sum_{n=1}^{\infty} f_N(n)\mathscr{L}_U(-s)^n \\ &\leq \frac{\mathrm{e}^{-sx}}{m^*}\,\mathbb{E}[\mathscr{L}_U(-s)^N] = \frac{\mathrm{e}^{-sx}}{m^*}\mathcal{G}_N(\mathscr{L}_U(-s)) = \frac{\mathrm{e}^{-sx}}{m^*}\mathscr{L}_{S_N}(-s) \end{aligned}$$

for $s \in [0, \rho_{S_N})$. This proves that $f^+_{S_N}(x) = \mathcal{O}(\mathrm{e}^{-sx})$ for $s \in [0, \rho_{S_N})$, and taking $s \nearrow \rho_{S_N}$ gives the result. □

Proposition 4.6 implies that for sums of gamma variables, the integrability condition is satisfied if $m > 1/(2\rho_{S_N})$.

The parameters for the polynomial approximations are set differently for the light-tailed and heavy-tailed cases. In the light-tailed cases moment matching of order 2 is the

natural procedure to set the values of $r$ and $m$, while ensuring that $m > (2\rho_{S_N})$. The two distributions we use for modeling the claim frequency $N$ are the *Poisson* and the *Pascal* distributions. We denote the Poisson distribution as $\mathsf{Poisson}(\lambda)$ and define the Pascal random variable to be the number of failures counted before observing $\alpha \in \mathbb{N}_+$ successes, denoted $\mathsf{Pascal}(\alpha, p)$.

**Example 4.7.** Let $N$ be Poisson distributed, the moment generating function of $S_N$ is then given by

$$\mathscr{L}_{S_N}(-s) = \exp\{\lambda[\mathscr{L}_U(-s) - 1]\}\,.$$

The radius of convergence of $S_N$ coincides with the one of $U$, $\rho_{S_N} = \rho_U$. In that case, we can set $r = 1$ and $m = \lambda\,\mathbb{E}[U]$ which corresponds to a moment matching procedure of order 1 or set $r = \lambda\,\mathbb{E}[U]^2/\,\mathbb{E}[U^2]$ and $m = \mathbb{E}[U^2]/\,\mathbb{E}[U]$ which, in turns, matches the two first moments.

**Example 4.8.** Let $N$ be Pascal distributed, the moment generating function of $S_N$ is then given by

$$\mathscr{L}_{S_N}(-s) = \left[\frac{p}{1 - q\mathscr{L}_U(-s)}\right]^{\alpha}.$$

The radius of convergence $\rho_{S_N}$ is the positive solution of the equation $\mathscr{L}_U(-s) = q^{-1}$. We set $r = 1$ and $m = \rho_{S_N}^{-1}$.

The parametrisation proposed in Example 4.8 is linked to the fact that it leads to the exact defective pdf in the case of a compound Pascal model with exponentially distributed claim sizes. The following lemma, adapted from [137], shows a useful correspondence between the Pascal and binomial distributions when used in compound sums with the exponential distribution.

**Lemma 4.9.** *Consider the random sums $X = \sum_{i=1}^{N_1} U_i$ and $Y = \sum_{i=1}^{N_2} V_i$, where*

$$N_1 \sim \mathsf{Pascal}(\alpha, p)\,, \quad U_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(1, \beta)\,, \quad N_2 \sim \mathsf{Binomial}(\alpha, q)\,, \quad V_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(1, p^{-1}\beta)\,,$$

*where $p \in (0, 1)$, $\alpha \in \mathbb{N}_+$, $p + q = 1$, and where $\beta > 0$. Then we have $X \overset{\mathcal{D}}{=} Y$.*

*Proof.* Both $X$ and $Y$ have the same Laplace transform, so $X \overset{\mathcal{D}}{=} Y$. □

**Corollary 4.10.** *Consider the compound sum* $S_N = \sum_{i=1}^{N} U_i$ *where* $N \sim \mathsf{Pascal}(\alpha, p)$ *and the* $U_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(1, \beta)$*. Then the survival function of* $S_N$ *is given by*

$$\overline{F}_{S_N}(x) = \sum_{i=1}^{\alpha} \binom{\alpha}{i} q^i p^{\alpha-i} \, \Gamma_u\left(i, p^{-1}\beta, x\right),$$

*and its stop-loss premium is given by*

$$\mathbb{E}\left[(S_N - a)_+\right] = \sum_{i=1}^{\alpha} \binom{\alpha}{i} q^i p^{\alpha-i} \left[\frac{i\beta}{p}\Gamma_u\Big(i+1, p^{-1}\beta, a\Big) - a\Gamma_u\left(i, p^{-1}\beta, a\right)\right].$$

*Proof.* By Lemma 4.9 we can instead consider the $S_N$ defined by $N \sim \mathsf{Binomial}(\alpha, q)$ and with $U_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(1, p^{-1}\beta)$. Noting that $S_n = U_1 + \cdots + U_n \sim \mathsf{Gamma}(n, p^{-1}\beta)$ gives the result. □

One conclusion of Corollary 4.10 is that the exact solution coincides with our approximation when $r = 1$ and $m = p^{-1}\beta$ (and with $K \geq \alpha - 1$). Note that $p\beta^{-1}$ is the solution of the equation $\mathscr{L}_U(-s) = q^{-1}$ which is consistent with the parametrisation proposed in Example 4.8.

In the heavy-tailed cases (i.e. when exponential tilting is required) we set $\theta = 1$ and $m = \theta/2 = 1/2$ (at the lower limit for $m$; this gives $\widetilde{m} = 1$), and choose $r = \mathbb{E}[U]$.

### Computation of the $a_k$

The inherent challenge of the implementation of the polynomial method remains the evaluation of the coefficients $\{a_k\}_{k\in\mathbb{N}_0}$. Recall that

$$a_k = \int_0^{\infty} p_k(x) f_{S_N}^{+}(x) \, \mathrm{d}x, \;\; k \geq 0.$$

We propose an evaluation based on the Laplace transform $\mathscr{L}\{f_{S_N}^{+}\}$. Define the generating function of the sequence $\{a_k d_k\}_{k\in\mathbb{N}_0}$ as $\mathcal{Q}(z) = \sum_{k=0}^{\infty} a_k d_k z^k$, where

$$d_k = \left(\frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}\right)^{1/2}, \quad \text{for } k \geq 0\,.$$

The following result establishes a link between the Laplace transform of $f^+_{S_N}$ and the generating function $\mathcal{Q}(z)$.

**Proposition 4.11.** *Assume that $f^+_{S_N}/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$, then we have*

$$\mathcal{Q}(z) = (1+z)^{-r}\mathscr{L}\{f^+_{S_N}\}\Big[\frac{-z}{m(1+z)}\Big]. \tag{4.16}$$

*Proof.* As $f^+_{S_N}/w \in \mathcal{L}^2(\mathbb{R}, w(x)\,\mathrm{d}x)$, the polynomial representation of $f^+_{S_N}$ follows from the application of Lemma 4.1 with

$$f^+_{S_N}(x) = \sum_{k=0}^{\infty}\sum_{i=0}^{k} a_k \frac{(-1)^{i+k}}{i!\,(k-i)!}\sqrt{\frac{k!\Gamma(k+r)}{\Gamma(r)}}\gamma(r+i, m, x). \tag{4.17}$$

Taking the Laplace transform in (4.17) yields

$$\begin{aligned}
\mathscr{L}\{f^+_{S_N}\}(s) &= \Big(\frac{1}{1+sm}\Big)^r \sum_{k=0}^{\infty} a_k \sum_{i=0}^{k}(-1)^{k+i}\left(\frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}\right)^{1/2}\binom{k}{i}\Big(\frac{1}{1+sm}\Big)^i \\
&= \Big(\frac{1}{1+sm}\Big)^r \sum_{k=0}^{\infty} a_k d_k (-1)^k \sum_{i=0}^{k}\binom{k}{i}\Big(\frac{-1}{1+sm}\Big)^i \\
&= \Big(\frac{1}{1+sm}\Big)^r \sum_{k=0}^{\infty} a_k d_k (-1)^k \Big(\frac{sm}{1+sm}\Big)^k \\
&= \Big(1-\frac{sm}{1+sm}\Big)^r \mathcal{Q}\Big(-\frac{sm}{1+sm}\Big).
\end{aligned}$$

Thus (4.16) follows from letting $z = -sm/(1+sm)$. □

The Laplace transform of the defective pdf $f^+_{S_N}$ is given by

$$\mathscr{L}\{f^+_{S_N}\}(s) = \mathscr{L}_{S_N}(s) - \mathbb{P}(N=0).$$

The coefficients of the polynomials can be derived after differentiation of the generating function $\mathcal{Q}(z)$ as

$$a_k = \frac{1}{d_k}\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}z^k}\mathcal{Q}(z)\Big|_{z=0} = \frac{1}{d_k}\mathrm{Coefficient}(k, \mathrm{MaclaurinSeries}(\mathcal{Q}(z))).$$

## 4.4 Laplace transform inversion approximations

The Laplace transform inversion approach described in Section 1.3.2 is applied here. It uses a method inspired by the work of Abate and Whitt [2], and the key equations are repeated here for clarity. For a function $f$ we define for $\ell = 1, 2, \ldots$

$$s_\ell(x) := \frac{\mathrm{e}^{a/2}}{2x}\mathscr{L}\{f\}\left(\frac{a}{2x}\right) + \frac{\mathrm{e}^{a/2}}{x}\sum_{k=1}^{\ell}(-1)^k \Re\left[\mathscr{L}\{f\}\left(\frac{a+2\pi \mathrm{i} k}{2x}\right)\right],$$

and with some positive integers $M_1$ and $M_2$, we have

$$f_{\text{approx}}(x) := \sum_{k=0}^{M_1}\binom{M_1}{k}2^{-M_1}s_{M_2+k}(x) \approx f(x)\,. \tag{4.18}$$

For a random sum $S_N$, we consider using the technique above to evaluate its survival function $\overline{F}_{S_N}$ and stop-loss premiums from its Laplace transform. We invert $\mathscr{L}\{\overline{F}_{S_N}\}$, but note that inverting $\mathscr{L}\{F_{S_N}\}$ produces almost identical results.

This inversion easily gives estimates of $\overline{F}_{S_N}$, though evaluating the stop-loss premiums requires extra thought. As noted in Dufresne et al. [66], we have that

$$\mathbb{E}\left[(S_N - d)_+\right] = \mathbb{E}[S_N]\overline{F_{S_N^*}}(d), \tag{4.19}$$

where $S_N^*$ is a random variable under the *equilibrium distribution* with density

$$f_{S_N^*}(x) = \begin{cases} \overline{F}_{S_N}(x)/\,\mathbb{E}[S_N], & \text{for } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

and Laplace transform

$$\mathscr{L}_{S_N^*}(s) = \frac{1 - \mathscr{L}_{S_N}(s)}{s\,\mathbb{E}[S_N]}.$$

The stop-loss premium is then obtained, replacing in (4.19) the survival function of $S_N^*$ by its approximation through (4.18).

## 4.5 Numerical illustrations

We illustrate the performance of the two proposed numerical procedures. Section 4.5.1 focuses on approximating the survival function and the stop-loss premium associated to aggregated claim sizes, while Section 4.5.2 considers the approximation of the finite-time ruin probability with no initial reserves using formula (4.5).

For each test case we compare the orthogonal polynomial approximation, the Laplace inversion approximation, and for the crude Monte Carlo approximation. For the cases when $U$ is gamma distributed, we use the fact that $S_n$ is Erlang distributed to produce an approximate distribution for $S_N$ by truncating $N$ to be less than some large level.

The parameters for the polynomial approximations has been discussed in Section 4.3.3, the calibration is depending on the assumptions over the claim frequency and claim sizes distribution. The parameters for the Laplace inversion technique are set to $M_1 = 11$, $M_2 = 15$ and $a = 18.5$ following the example of Rolski et al. [147, Chapter 5, Section 5]; note, this choice of $a$ implies that the discretisation error is less than $10^{-8}$, as derived from (1.9). Note, we do not use any built-in routines for the Laplace inversion, but simply implement (4.18).

In each plot, the first subplot shows the estimates each estimator produces, and the second shows the *approximate absolute error*. We define this, for estimator $i \in \{1, \ldots, I\}$, as

$$\begin{aligned}\text{ApproximateAbsoluteError}(\widehat{f}_i, x) &:= \widehat{f}_i(x) - \text{Median}\Big\{\widehat{f}_1(x), \ldots, \widehat{f}_I(x)\Big\} \\ &\approx \widehat{f}_i(x) - f(x) =: \text{AbsoluteError}(\widehat{f}_i, x)\,.\end{aligned}$$

When the different estimators cross each other, the median obtains an unrealistically jagged character. We therefore use as reference a slightly smoothed version of the median, achieved in MATHEMATICA using `GaussianFilter[Medians, 2]`. As noted earlier, all of the code used is available online [82].

### 4.5.1 Survival function and stop-loss premium computations

To ensure both estimators were implemented correctly, we applied the estimators to the case where $N \sim \mathsf{Pascal}(\alpha = 10, p = 3/4)$ and $U \sim \mathsf{Gamma}(r = 1, m = 1/6)$. Corollary 4.10 tells us the orthogonal approximation (with $r = 1$, $m = \lambda/p = 2/9$ and $K = \alpha - 1 = 9$) is equivalent to the true function, which we verified, and the Laplace inversion errors in Tables 4.1 and 4.2 are acceptably small.

Table 4.1: Relative errors for the Laplace inversion survival function estimator

| $x$ | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|
| Error | 7.27e-7 | 1.92e-6 | 5.86e-6 | 1.78e-5 | 4.01e-5 |

Table 4.2: Relative errors for the Laplace inversion stop-loss premium estimator

| $a$ | 0.5 | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|---|
| Error | 8.68e-7 | 2.27e-6 | 5.92e-6 | 1.12e-5 | -2.12e-5 |

**Test 4.12.** $N \sim \mathsf{Poisson}(\lambda = 2)$, *and* $U \sim \mathsf{Gamma}(r = 3/2, m = 1/3)$

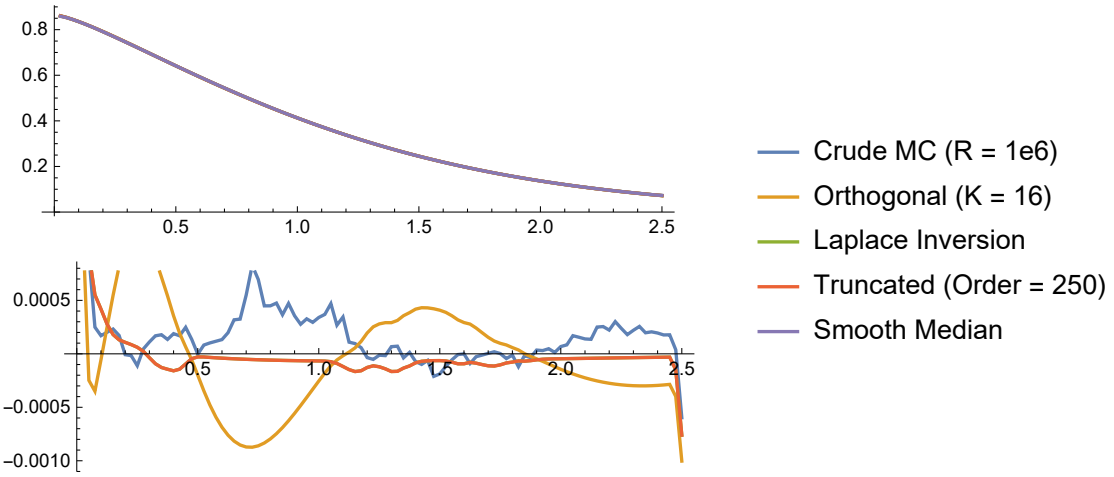



Figure 4.1: Survival function estimates and approximate absolute error.

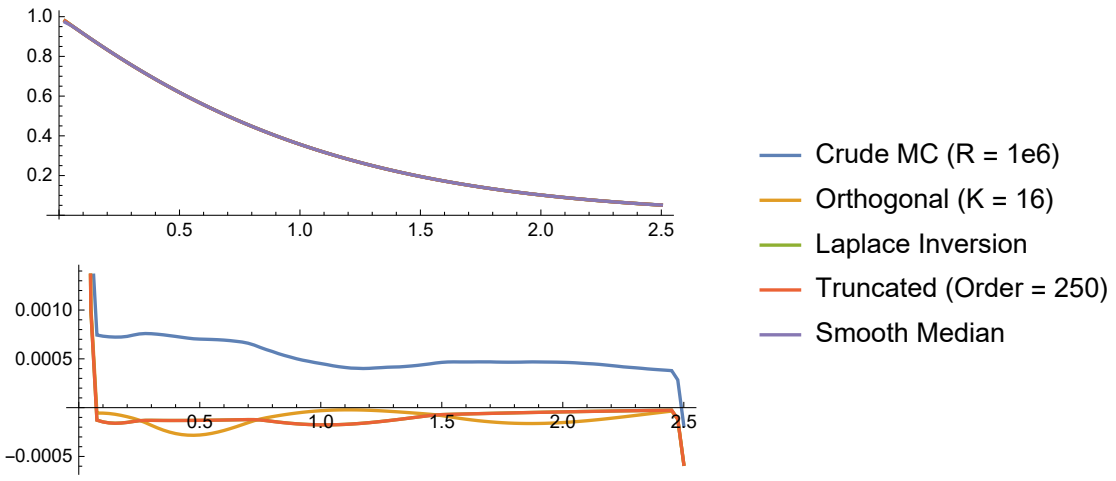



Figure 4.2: Stop-loss premium estimates and approximate absolute error.

**Test 4.13.** $N \sim \mathsf{Pascal}(\alpha = 10, p = 1/6)$, *and* $U \sim \mathsf{Gamma}(r = 3/2, m = 1/75)$

This test case (up to the scaling constant) has been considered by Jin et al. [102, Example 3]. In the plots for this test case, the orthogonal estimator, the Laplace inversion estimator, and the truncated estimator all give the same values and hence are hidden underneath the red line for the truncated estimator.

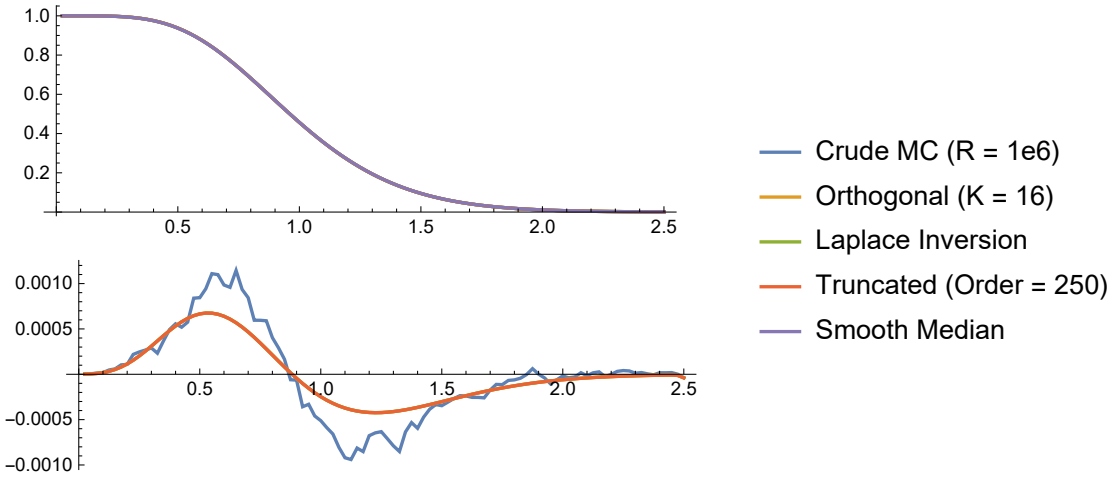



Figure 4.3: Survival function estimates and approximate absolute error.

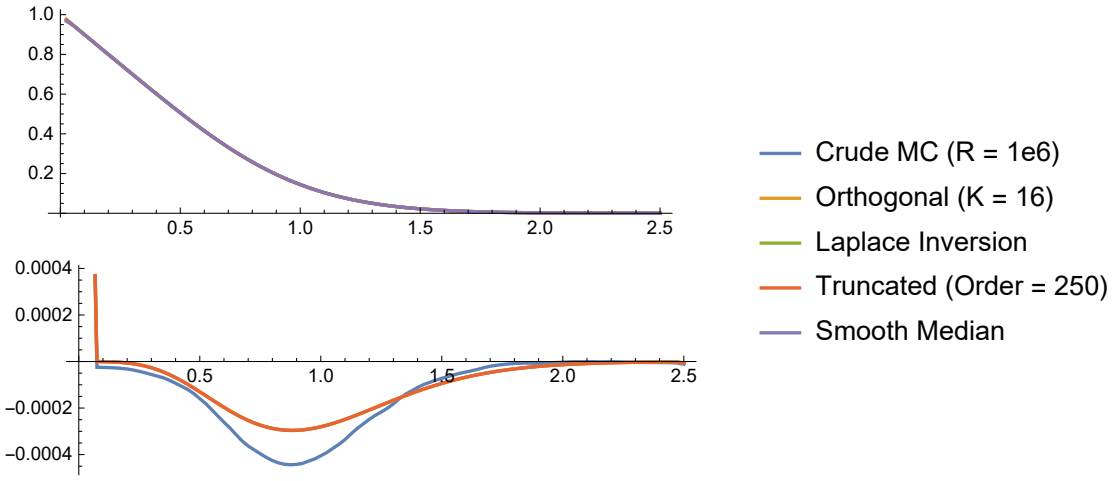



Figure 4.4: Stop-loss premium estimates and approximate absolute error.

**Test 4.14.** $N \sim \mathsf{Poisson}(\lambda = 4)$, *and* $U \sim \mathsf{Pareto}(a = 5, b = 11, \theta = 0)$

The Laplace inversion estimator breaks down for small values of $x$ or $a$ in this test case. The specific error given is an "out of memory" exception when MATHEMATICA is attempting to do some algebra with extremely large numbers. It is unclear whether a different implementation or selection of parameters would fix this behaviour.

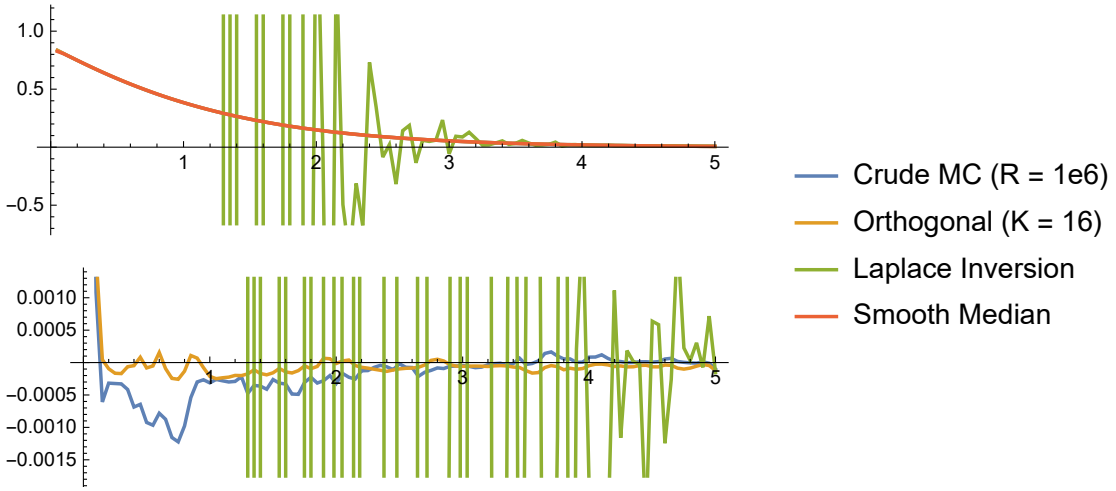



Figure 4.5: Survival function estimates and approximate absolute error.

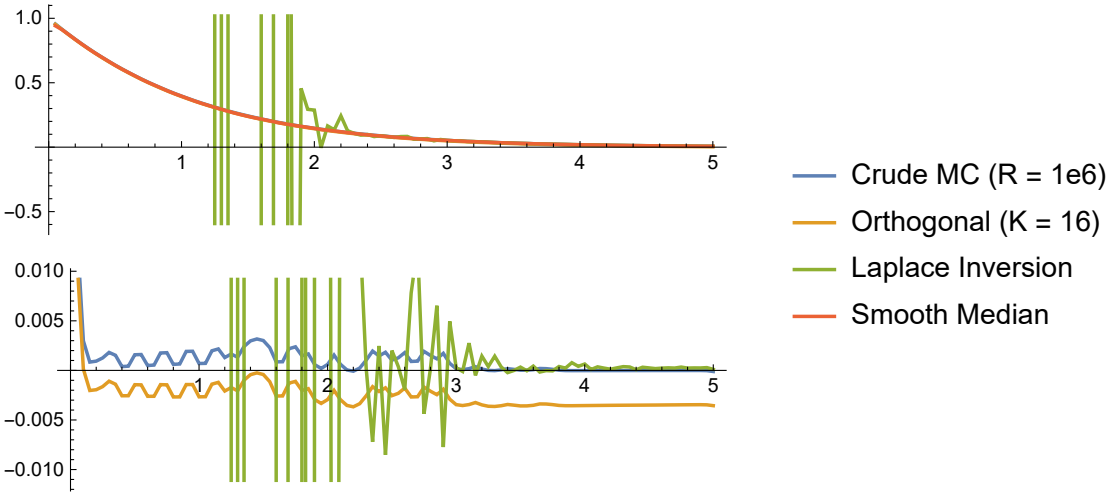



Figure 4.6: Stop-loss premium estimates and approximate absolute error.

**Test 4.15.** $N \sim \text{Pascal}(\alpha = 2, p = 1/4)$, *and* $U \sim \text{Weibull}(\beta = 1/2, \lambda = 1/2)$

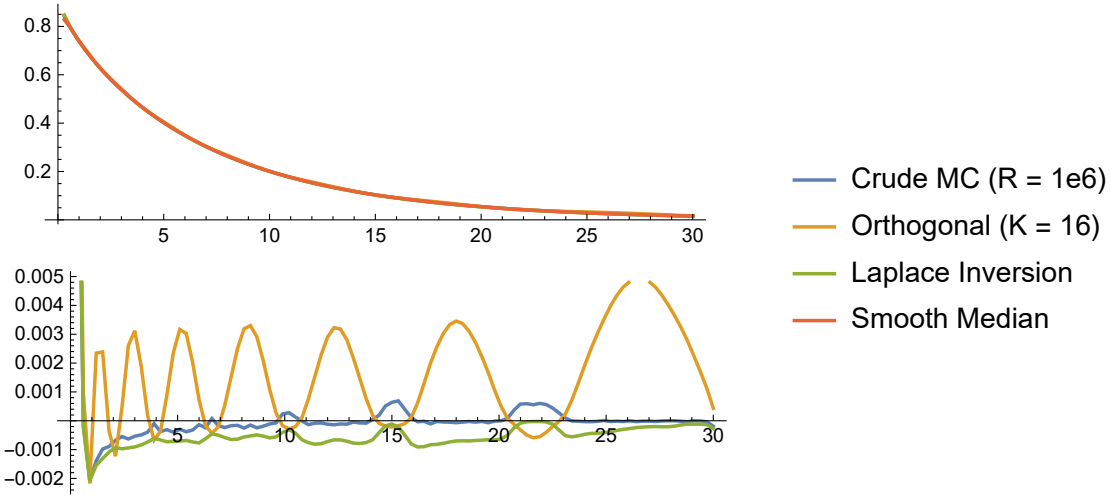



Figure 4.7: Survival function estimates and approximate absolute error.

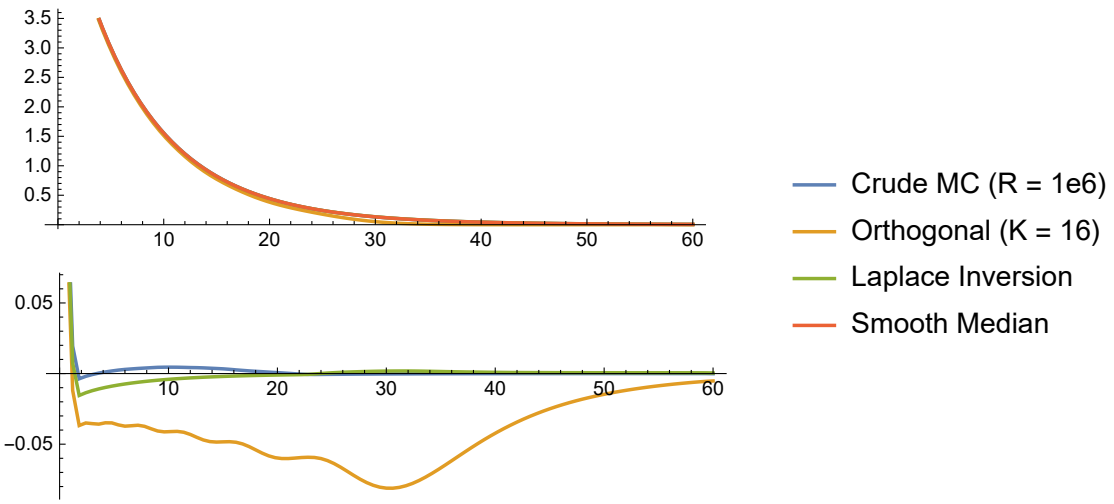



Figure 4.8: Stop-loss premium estimates and approximate absolute error.

### 4.5.2 Finite-time ruin probability with no initial reserve

The plots above have used common random numbers to smooth the estimators, however this isn't possible in the following plots so they will appear less smooth.

**Test 4.16.** $\lambda = 4$ *and* $U \sim \mathsf{Gamma}(r = 2, m = 2)$ *and* $c = 1$

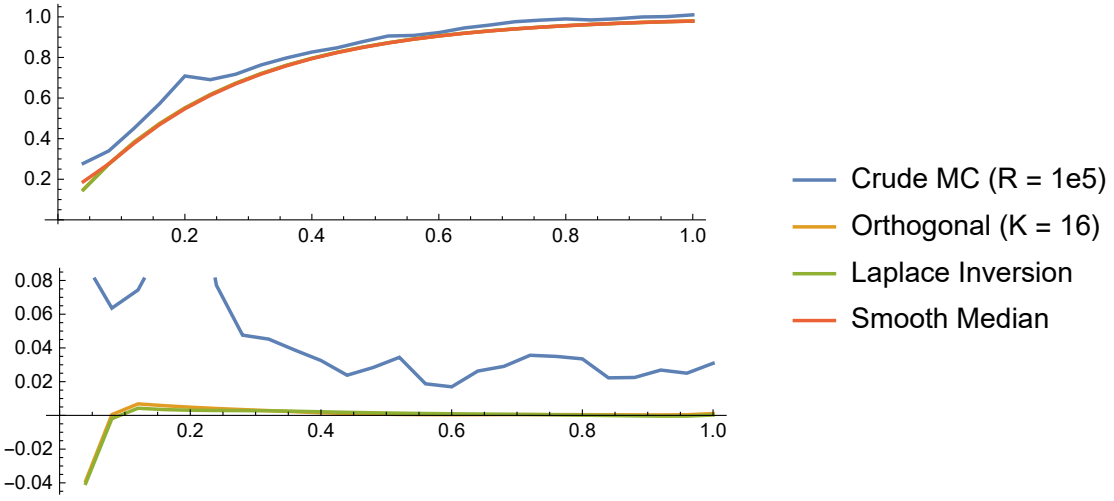



Figure 4.9: Ruin probability $\psi(0, t)$ estimates and approximate absolute error.

**Test 4.17.** $\lambda = 2$ *and* $U \sim \mathsf{Pareto}(a = 5, b = 11, \theta = 0)$ *and* $c = 1$

See the discussion of Test 4.14 for a description of the Laplace inversion estimator's poor behaviour when Pareto variables are involved.

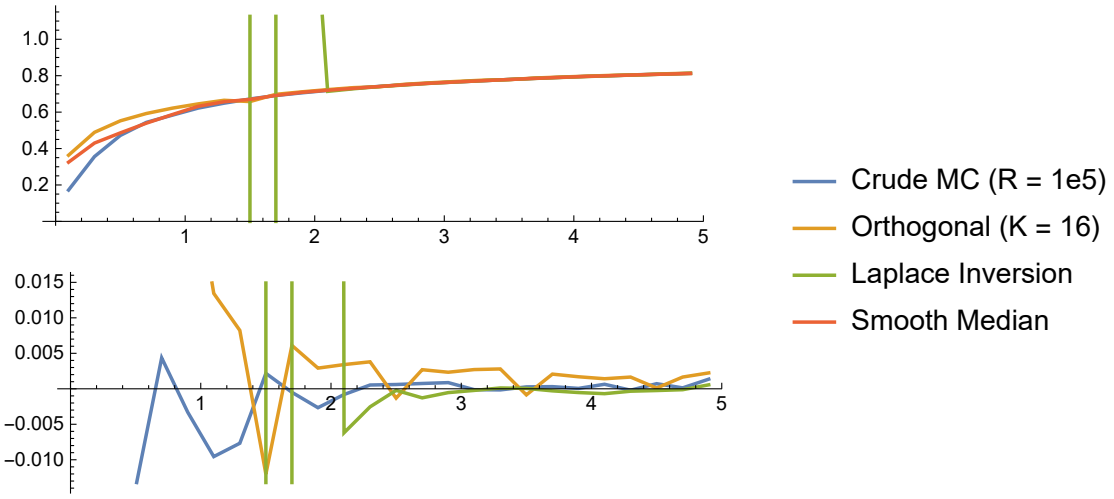



Figure 4.10: Ruin probability $\psi(0, t)$ estimates and approximate absolute error.

### 4.5.3 Concluding remarks

The orthogonal polynomial method has performed well across all the test cases studied. The accuracy is acceptable even with a rather small order of truncation $K = 16$. It produces an approximation having an analytical expression, which is desirable, and in a timely manner. The precision may be improved by adding more terms in the expansions. The main drawback is probably the need for a parametrisation tailored to the case studied.

The Laplace transform inversion method yields outstanding result in terms of accuracy. It failed to provide a stable approximation for Pareto distributed claim sizes. The parametrisation is automatic and seems to fit the different case studied (except the Pareto one).

The main conclusion is that both methods are easy to implement and are superior to a simple truncation or a crude Monte Carlo approach.

## Chapter 5 Authorship Statement

**Citation:** Thomas Taimre, Patrick J. Laub (2018), *Rare tail approximation using asymptotics and L1 polar coordinates*, Statistics and Computing (submitted)

The authors of this paper equally contributed to the following tasks:

1. conception and design of the project;
2. mathematical arguments, and interpretation of the results;
3. writing the publication.

In addition to this, I completed the majority of the computational work.

# Chapter 5

# Rare tail approximation using asymptotics and polar coordinates

## 5.1 Introduction

This chapter focuses on evaluating

$$\ell(\gamma) := \mathbb{P}(S > \gamma) \tag{5.1}$$

where $S := X_1 + \cdots + X_d$ for a fixed $d \in \mathbb{N}_+$ and where the $\gamma \in \mathbb{R}$ is large or increasing. As detailed above, this is often a difficult problem which does not have a simple closed-form solution.

When analytical solutions are unavailable, the next best option is numerical integration, and after that Monte Carlo integration (or quasi-Monte Carlo). Numerical integration algorithms applied to

$$\ell(\gamma) = \int_{\mathbb{R}^d} \mathbb{I}\{x_1 + \cdots + x_d > \gamma\} f_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

are typically slow, inaccurate, and misleading. This is because the indicator is rarely 1, floating-point errors accumulate, and the curse of dimensionality applies for $d$ larger than about 2 or 3. Some of these algorithms attempt to estimate the error in their result, but

there are few (if any) theoretical guarantees that these estimates are reliable.

Rare-event problems also cause difficulties for the crude Monte Carlo (CMC) estimator. This is obvious as the CMC estimator's relative error explodes for large $\gamma$ — that is, the CMC estimator $\widehat{\ell}_{\mathrm{CMC}}(\gamma) := \mathbb{I}\{S > \gamma\}$ has

$$\lim_{\gamma\to\infty} \mathrm{RelativeError}\{\widehat{\ell}_{\mathrm{CMC}}(\gamma)\} = \lim_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}[\widehat{\ell}_{\mathrm{CMC}}(\gamma)]}{\ell(\gamma)^2} = \lim_{\gamma\to\infty} \frac{\ell(\gamma)[1-\ell(\gamma)]}{\ell(\gamma)^2} = \infty\,.$$

Intuitively, the problem is because the indicator $\mathbb{I}\{S > \gamma\}$ is eventually always 0 when $\gamma$ gets very large. In response, various variance reduction techniques have been applied so that there are now a large collection of estimators with better performance in this setting, c.f. 'rare-event estimation' in [110, 15, 79].

There is, of course, no silver bullet for the problem. Some estimators only apply to specific distributions (e.g. [37] for sums of lognormals, [166] for sums of phase-type mixtures) or to certain classes of distributions (exponential tilting for light-tailed summands [110, 15], hazard-rate twisting or the Asmussen–Kroese method [21] for heavy-tailed summands). Other estimators are general but require specifying either some extra information (e.g. some conditional distributions for conditional Monte Carlo [13], or a sampling distribution for importance sampling). The most general estimators — such as the generalised splitting method, cross-entropy method, or Markov Chain Monte Carlo (MCMC) methods like [43] — are usually computationally demanding, they often depend upon an intelligent selection of input parameters to perform efficiently, and are somewhat complicated.

While we almost never have an analytic solution for $\ell(\gamma)$, it is somewhat common to know the *asymptotic approximation* to it, and this forms the basis for our proposed estimator. For example, if $\boldsymbol{X} \sim \mathsf{Lognormal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is positive definite, then it has been shown that [23]

$$\ell(\gamma) = \mathbb{P}(S > \gamma) \sim \sum_{i=1}^{d} \mathbb{P}(X_i > \gamma) \qquad \text{as } \gamma \to \infty \tag{5.2}$$

where $f(x) \sim g(x)$ denotes $\lim_{x\to\infty} f(x)/g(x) = 1$. Thus, one is tempted to label the RHS of (5.2) as $\widehat{\ell}_{\mathrm{Asym}}(\gamma)$ and use it as an approximation for $\ell(\gamma)$. For certain values of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ this asymptotic approximation can be accurate, in others it can be wildly inaccurate,

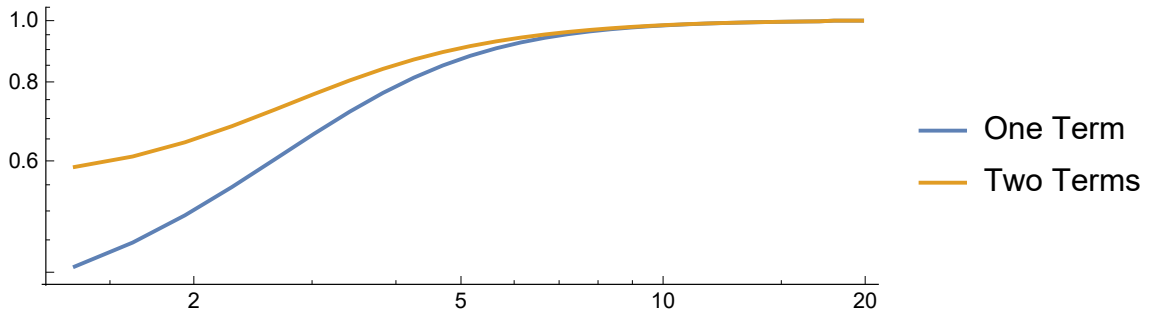



Figure 5.1: A comparison of $\ell(\gamma)$ and $\widehat{\ell}_{\mathrm{Asym}}(\gamma)$ for $X_1 + X_2$ where $X_1 \sim \mathsf{Lognormal}(0,1)$ is independent to $X_2 \sim \mathsf{Lognormal}(0, \frac{3}{4})$. The $y$ axis plots $\widehat{\ell}_{\mathrm{Asym}}(\gamma)/\ell(\gamma)$, and the $x$ axis shows $-\log_{10}\ell(\gamma)$. The two lines describe two possible asymptotics, the yellow "Two terms" describes $\widehat{\ell}_{\mathrm{Asym}}(\gamma)$ as given in (5.2) whereas the blue "One Term" uses just the first term of this sum.

depending on how fast the asymptotic approximation converges to the true value; see Figure 5.1 for an illustration where it is only when $\ell(\gamma) \lesssim 10^{-10}$ that the asymptotic form begins to give accurate estimates (i.e., $\widehat{\ell}_{\mathrm{Asym}}(\gamma)/\ell(\gamma) > 0.99$). A discussion of this phenomenon is in [37].

We propose an *importance sampling* estimator which incorporates the asymptotic approximation and uses Monte Carlo sampling to estimate the difference between $\ell(\gamma)$ and $\widehat{\ell}_{\mathrm{Asym}}(\gamma)$. In the estimator, the asymptotic approximation acts like a prior belief (though it is just a point-estimate) of the value of $\ell(\gamma)$, which gets corrected/updated with more samples.

The main drawback to importance sampling is *likelihood degeneration*, where one can face numerical errors if $\gamma$ or $d$ is extremely large. The degeneration caused by a large $d$ is only partially compensated by our method, so we take $d \leq 100$. To mitigate degeneration for large $\gamma$, we focus our attention of values of $\gamma$ which are moderately large but not unrealistically so. Our goal is to provide an estimator which is practically useful when $\ell(\gamma)$ is between roughly $10^{-3}$ and $10^{-7}$.

The range of probabilities that we consider are unusual as they are less rare than much of the standard rare-event literature. The orthodox approach is to construct an estimator $\widehat{\ell}(\gamma)$ and analyse the limit $\lim_{\gamma\to\infty} \mathbb{V}\mathrm{ar}(\widehat{\ell}(\gamma))/\ell(\gamma)^2$; if the limit is small (i.e., zero, bounded, or at least grows only at a polynomial rate) then the estimator is branded as a success

(it has 'vanishing relative error', 'bounded relative error', or is 'logarithmically efficient' respectively) regardless of its behaviour in the finite $\gamma$ situation. It can happen that these desirable limiting properties are only discernible in cases when the probabilities are truly minuscule (e.g. of order $10^{-10}$ or smaller); in a situation like this, the model error would surely dominate any estimation error.

The estimator is introduced in Section 5.2, the results from numerical comparisons are in Section 5.3, and Section 5.4 concludes the discussion.

## 5.2 The polar estimator

### 5.2.1 The general form

We construct an estimator of the quantity $\ell(\gamma) := \mathbb{P}(S > \gamma)$, where $S = X_1 + \cdots + X_d$ for large $\gamma$ by applying IS. Standard IS theory says to construct an estimator which samples from a distribution close to $f_{\boldsymbol{X}\,|\,S>\gamma}$ (that is, the distribution of $\boldsymbol{X}$ conditioned on $\{S > \gamma\}$), rather than the original $f_{\boldsymbol{X}}$. To do this, perform a change of variables so

$$\boldsymbol{X} \longrightarrow (S, \boldsymbol{\Theta}) := (X_1 + \cdots + X_d, \boldsymbol{X}/[X_1 + \cdots + X_d]) \; .$$

The new density $f_{(S,\boldsymbol{\Theta})}$ is available (if $f_{\boldsymbol{X}}$ is known), and is

$$f_{(S,\boldsymbol{\Theta})}(s, \boldsymbol{\theta}) = f_{\boldsymbol{X}}(s\boldsymbol{\theta}) \times |s|^{d-1} \, .$$

Consider IS in this new form. Imagine that we have a density $g_{(S,\boldsymbol{\Theta})}$ which is in some way similar to $f_{(S,\boldsymbol{\Theta})}$, and we also know the marginal density $g_S(s) := \int g_{(S,\boldsymbol{\Theta})}(s, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$ and the conditional density $g_{\boldsymbol{\Theta}|S} := g_{(S,\boldsymbol{\Theta})}/g_S$. If we truncate $g_{(S,\boldsymbol{\Theta})}$ so that $S > \gamma$ a.s., and use this as the IS measure, we get

$$\widehat{l}_{\mathrm{IS}}(\gamma) := \frac{\overline{G}_S(\gamma)}{R} \sum_{r=1}^{R} \frac{f_{(S,\boldsymbol{\Theta})}(S^{[r]}, \boldsymbol{\Theta}^{[r]})}{g_S(S^{[r]}) g_{\boldsymbol{\Theta}|S}(\boldsymbol{\Theta}^{[r]} \,|\, S^{[r]})} \quad \text{for} \quad \begin{array}{c} S^{[r]} \overset{\text{iid}}{\sim} g_{S|S>\gamma}, \\ \boldsymbol{\Theta}^{[r]} \overset{\text{ind}}{\sim} g_{\boldsymbol{\Theta}|S}(\,\cdot\,|S^{[r]}), \end{array} \tag{5.3}$$

where $\overline{G}_S(\gamma) := \int_\gamma^\infty g_S(s) \, \mathrm{d}s$, and $g_{S|S>\gamma} := g_S \mathbb{I}\{S > \gamma\}/\overline{G}_S(\gamma)$.

We investigate estimators of the general form of (5.3) which we call *polar estimators*. These are accurate when $g_{(S,\boldsymbol{\Theta})} = g_S \times g_{\boldsymbol{\Theta}|S}$ closely resembles $f_{(S,\boldsymbol{\Theta})} = f_S \times f_{\boldsymbol{\Theta}|S}$. This is done in two steps, by finding a *radial approximation* $g_S$ which approximates $f_S$, and an *angular approximation* $g_{\boldsymbol{\Theta}|S}$ similar to $f_{\boldsymbol{\Theta}|S}$, problems we discuss in the following sections.

### 5.2.2 The radial approximation

As mentioned in the introduction, we consider utilising the asymptotic form of the sum in our estimator — they form our radial approximation. To clarify the notation, we again define the relevant asymptotic forms:

**Definition 5.1** (Asymptotic form)**.** *If for some function* $\mathfrak{f}_S \in \mathcal{L}^1(\mathbb{R})$*, with tail* $\overline{\mathfrak{F}}_S(s) = \int_s^\infty \mathfrak{f}_S(x)\,\mathrm{d}x$*, and constant* $c_S \in \mathbb{R}_+$*, we have that*

$$f_S(s) \;\sim\; c_S \mathfrak{f}_S(s)\,, \qquad \text{for } s \to \infty \tag{5.4}$$

*then we say* $\mathfrak{f}_S$ *is an* asymptotic form *of* $f_S$*.* ◇

Thus, in the general form (5.3) we will use $g_S = \mathfrak{f}_S$ when it is available and is a proper pdf. There are some technicalities for the cases when $\mathfrak{f}_S$ does not form a proper pdf which we defer for now. The estimator resulting from this radial approximation is

$$\widehat{l}_{\mathrm{IS2}}(\gamma) := \frac{c_S \overline{\mathfrak{F}}_S(\gamma)}{R} \sum_{r=1}^{R} \frac{f_{(S,\boldsymbol{\Theta})}(S^{[r]}, \boldsymbol{\Theta}^{[r]})}{c_S \mathfrak{f}_S(S^{[r]}) g_{\boldsymbol{\Theta}|S}(\boldsymbol{\Theta}^{[r]} \mid S^{[r]})} \quad \text{for} \quad \begin{array}{c} S^{[r]} \overset{\text{iid}}{\sim} \mathfrak{f}_{S|S>\gamma}, \\ \boldsymbol{\Theta}^{[r]} \overset{\text{ind}}{\sim} g_{\boldsymbol{\Theta}|S}(\,\cdot\,|S^{[r]}). \end{array} \tag{5.5}$$

**Remark 5.2.** Define $\mathcal{R}(\gamma)$ by $\ell(\gamma) = \widehat{\ell}_{\mathrm{Asym}}(\gamma)\mathcal{R}(\gamma)$; n.b. $\widehat{\ell}_{\mathrm{Asym}}(\gamma) := c_S \mathfrak{F}_S(\gamma)$. We can see that $\widehat{\ell}_{\mathrm{IS2}}(\gamma)$ has a nice interpretation, because

$$\widehat{\ell}_{\mathrm{IS2}}(\gamma) = \widehat{\ell}_{\mathrm{Asym}}(\gamma) \times \widehat{\mathcal{R}}(\gamma)\,,$$

where $\widehat{\mathcal{R}}(\gamma)$ is a Monte Carlo estimate of $\mathcal{R}(\gamma)$. ◇

The recent applied probability literature has found the $\mathfrak{f}_S$ for a staggering array of distributions of $\boldsymbol{X}$. Perhaps the simplest case is when the $X_i$ are iid subexponential random

variables. By definition (cf. [73]), they satisfy

$$f_S(s) \;\sim\; d\, f_1(s)\,, \qquad \text{for } s \to \infty\,. \tag{5.6}$$

For sums of independent non-identically distributed subexponential variables (or for sums containing some subexponential and some lighter-tailed variables) we have

$$f_S(s) \;\sim\; \sum_{i=1}^{d} f_i(s) \;\sim\; \sum_{i\in I} f_i(s)\,, \qquad \text{for } s \to \infty \tag{5.7}$$

where $I$ is the set of indices of slowest tail decay. The asymptotics in (5.7) also hold in many regimes where dependence has been introduced, cf. [74, 165, 8, 9].

A distribution can satisfy a stronger property called *regular variation* which implies subexponentiality and hence the asymptotics above. Examples of regularly varying distributions are Cauchy, Fréchet, and Pareto distributions [33]. The lognormal and heavy-tailed Weibull distributions are subexponential but not regularly varying.

The Weibull distribution is interesting as it is a family which can be heavy-tailed, light-tailed (the Raleigh distribution is a special case), or on the boundary between these (i.e. the exponential distribution). The asymptotic form for the heavy-tailed Weibull sum is covered by (5.6) and (5.7) as the summands are subexponential. The difficulty in finding the asymptotics for the light-tailed case led the authors to investigate it in detail, leading to the paper [17] which uses results originally from [25].

**Proposition 5.3.** *Assume that $X_1, \ldots, X_d$ are iid light-tailed* $\mathsf{Weibull}(\beta, \lambda)$ *where $\beta > 1$, $\lambda \in \mathbb{R}_+$, $d \geq 2$. Then*

$$F_S(s) \;\sim\; \left[\frac{2\beta\pi}{\beta-1}\right]^{(d-1)/2} d^{-1/2} \left(\frac{s}{\lambda d}\right)^{\beta(d-1)/2} \overline{F}\left(\frac{s}{d}\right)^{d}, \qquad \textit{for } s \to \infty\,.$$

The exposition in [17] details this and more general asymptotics (i.e. the independent but non-identically distributed case, and when the variables are not exactly Weibull but are 'Weibull-like').

### 5.2.3 The angular approximation

The choice of angular approximation is not as obvious as was the choice of radial approximation. Finding a conditional density $g_{\boldsymbol{\Theta}|S}$ which is similar to $f_{\boldsymbol{\Theta}|S}$ has little to no precedent in the literature.

We can make a simplification by looking at a different conditional distribution. Instead of taking an $S$ which is larger than $\gamma$ and asking 'what is the distribution of $\boldsymbol{\Theta}$ given this $S$?', we can instead ask 'what is the distribution of $\boldsymbol{\Theta}$ given $S > \gamma$?'. This second conditional will resemble the first when $\gamma$ becomes large, since it is typically the case that $\mathbb{E}[S - \gamma \mid S > \gamma]$ converges quickly to zero. Also we have a computation benefit to finding a $g_{\boldsymbol{\Theta}|S>\gamma}$ which is similar to $f_{\boldsymbol{\Theta}|S>\gamma}$ as this distribution will be constant across all Monte Carlo iterates, in contrast to $g_{\boldsymbol{\Theta}|S^{[r]}}$ and $f_{\boldsymbol{\Theta}|S^{[r]}}$.

When it is possible, we follow the same approach as the radial approximation and utilise some asymptotic information. However, if one re-uses the previous asymptotic form, that is

$$f_{\boldsymbol{\Theta}|S}(\boldsymbol{\theta}|s) = \frac{f_{S,\boldsymbol{\Theta}}(s,\boldsymbol{\theta})}{f_S(s)} \sim \frac{f_{\boldsymbol{X}}(s\boldsymbol{\theta})|s|^{d-1}}{\mathfrak{f}_S(s)} =: g_{\boldsymbol{\Theta}|S}(\boldsymbol{\theta}|s)\,,$$

which is natural, then the estimator (5.5) degenerates to the deterministic

$$\widehat{l}_{\mathrm{IS2}}(\gamma) := \frac{\overline{\mathfrak{F}}_S(\gamma)}{R}\sum_{r=1}^{R} 1 = \overline{\mathfrak{F}}_S(\gamma)\,.$$

Unfortunately, the conditional $f_{\boldsymbol{\Theta}|S}$ and $f_{\boldsymbol{\Theta}|S}$ distributions are rarely examined in the literature, so we do not have a ready supply of their asymptotics.

When the summands are subexponential, then the distribution of $(\boldsymbol{\Theta} \mid S = s)$ as $s \to \infty$ degenerates to a discrete distribution over the unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$. This is just a re-casting of the principle of the single big jump (cf. Section 1.3.6 or [73]). One density, which we call the *optimistic density* (see the algorithm below), that is asymptotically equivalent to $(\boldsymbol{\Theta} \mid S = s)$ is

$$g_{\boldsymbol{\Theta}|S}(\boldsymbol{\theta} \mid s) = |s|^{d-1}\sum_{i=1}^{d} p_i(s) f_{\boldsymbol{X}_{-i}}(s\boldsymbol{\theta}_{-i})\mathbb{I}\{\theta_i = 1 - \mathbf{1}\cdot\boldsymbol{\theta}_{-i}\} \tag{5.8}$$

where the $p_i$ functions are defined by

$$p_i(s) = \frac{\overline{F}_i(s)}{\sum_{j=1}^{d} \overline{F}_j(s)} \,. \tag{5.9}$$

Algorithm 4 shows a method for sampling from this $g_{\boldsymbol{\Theta}|S}(\boldsymbol{\theta} \mid s)$, and Proposition 5.4 shows has the expected limiting distribution as $s \to \infty$.

**Algorithm 4** Sampling from the optimistic angular density

**procedure** OPTIMISTIC($s$, $F_1$, …, $F_d$)
  Simulate index $I$ in $\{1, \ldots, d\}$ by $\mathbb{P}(I = i) = p_i(s)$ from (5.9).
  **for** $i = 1$ **to** $d$ **except** $I$ **do**
    $X_i^* \leftarrow$ Random sample from $F_i$
  **end for**
  $X_I^* \leftarrow s - \sum_{i \neq I} X_i$ ▷ This can be negative, but we are optimistic
  **return** $\boldsymbol{\Theta} \leftarrow \boldsymbol{X}^*/s$
**end procedure**

**Proposition 5.4.** *The optimistic density* (5.8) *converges as* $s \to \infty$ *to the singular density*

$$g_\infty(\boldsymbol{\theta}) := \sum_{i=1}^{d} p_i \, \mathbb{I}\{\boldsymbol{\theta} = \boldsymbol{e}_i\} \,, \tag{5.10}$$

*where* $p_i = \lim_{s \to \infty} p_i(s)$ *for* $i = 1, \ldots, d$.

*Proof.* For some $\boldsymbol{t} = (t_1, \ldots, t_d)' \in \mathbb{R}^d$, the characteristic function of $g_{\boldsymbol{\Theta}|S}$ is

$$\phi_{g_{\boldsymbol{\Theta}|S}}(\boldsymbol{t} \mid s) = \mathbb{E} \exp\left(\mathrm{i}\, \boldsymbol{t}^\top \boldsymbol{\Theta}\right) = \mathbb{E}\left[\exp\left(\mathrm{i}\, \frac{\boldsymbol{t}}{s}^\top \boldsymbol{X}^*\right)\right]$$

where $\boldsymbol{X}^* = s\boldsymbol{\Theta}$ as in Algorithm 4.

So, with $I$ as the discrete variable defined in Algorithm 4, we have

$$\begin{aligned}
\phi_{g_{\boldsymbol{\Theta}|S}}(\boldsymbol{t} \mid s) &= \sum_{j=1}^{d} p_j(s)\, \mathbb{E}\left[\exp\left(\mathrm{i}\,\frac{\boldsymbol{t}}{s}^{\top} \boldsymbol{X}^*\right) \,\middle|\, I = j\right] \\
&= \sum_{j=1}^{d} p_j(s)\, \mathbb{E}\left[\exp\left\{\mathrm{i}\left[\frac{\boldsymbol{t}_{-j}}{s}^{\top} \boldsymbol{X}^*_{-j} + \frac{t_j}{s}(s - \mathbf{1}^{\top}\boldsymbol{X}^*_{-j})\right]\right\} \,\middle|\, I = j\right] \\
&= \sum_{j=1}^{d} p_j(s) \mathrm{e}^{\mathrm{i}t_j}\, \mathbb{E}\left[\exp\left(\mathrm{i}\,\frac{(\boldsymbol{t}_{-j} - t_j\mathbf{1})}{s}^{\top} \boldsymbol{X}_{-j}\right)\right] \\
&= \sum_{j=1}^{d} p_j(s) \mathrm{e}^{\mathrm{i}t_j} \phi_{\boldsymbol{X}_{-j}}\left(\frac{(\boldsymbol{t}_{-j} - t_j\mathbf{1})}{s}\right).
\end{aligned}$$

Therefore,

$$\lim_{s\to\infty} \phi_{g_{\boldsymbol{\Theta}|S}}(\boldsymbol{t}; s) = \sum_{j=1}^{d} p_j \mathrm{e}^{\mathrm{i}t_j} = \sum_{j=1}^{d} p_j \mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\top}\boldsymbol{e}_j}$$

which has the singular density as in (5.10). □

**Remark 5.5.** The polar estimator with the optimistic angular approximation (5.8) simplifies to

$$\widehat{l}_{\mathrm{IS2}}(\gamma) = \frac{c_S \overline{\mathfrak{F}}_S(\gamma)}{R} \sum_{r=1}^{R} \frac{\mathrm{HarmonicMean}(f_{X_1}(S^{[r]}\Theta_1^{[r]}), \ldots, f_{X_d}(S^{[r]}\Theta_d^{[r]}))}{c_S \mathfrak{f}_S(S^{[r]})}$$

where $S^{[r]} \overset{\mathrm{iid}}{\sim} \mathfrak{f}_{S|S>\gamma}$, $\boldsymbol{\Theta}^{[r]} \overset{\mathrm{ind}}{\sim} g_{\boldsymbol{\Theta}|S}(\,\cdot\,|S^{[r]})$. ◇

The conditional angular asymptotic distribution is harder to obtain in the case of light-tailed summands. The following example shows these distributions differ qualitatively when different copulas are considered.

**Example 5.6.** Consider $X_1$ and $X_2$ are $\mathsf{Exp}(1)$ variables which are: i) independent, ii) $\mathsf{Clayton}(1)$ dependent, or iii) $\mathsf{Ali\text{-}Mikhail\text{-}Haq}(\text{-}1)$ dependent. The sum densities are

$$f_S^{\mathrm{Ind}}(s) = s\mathrm{e}^{-s} \quad \text{for } s > 0\,,$$

$$f_S^{\mathrm{Cla}}(s) = \frac{2 - 2\cosh(s) + s\sinh(s)}{(\cosh(s) - 1)^2} \quad \text{for } s > 0$$

$$f_S^{\mathrm{AMH}}(s) = 8\,\mathrm{csch}(s)^3 \sinh(s/2)^4 \quad \text{for } s > 0$$

respectively, and hence for $s > 0$ and $\theta \in (0, 1)$

$$f^{\text{Ind}}_{\Theta_1|S}(\theta|s) = 1\,,$$

$$f^{\text{Cla}}_{\Theta_1|S}(\theta|s) = \frac{s\mathrm{e}^{-s\theta}(\mathrm{e}^s - \mathrm{e}^{s\theta})(\mathrm{e}^{s\theta} - 1)}{2 + s - 2\mathrm{e}^s + s\mathrm{e}^s}\,,$$

$$f^{\text{AMH}}_{\Theta_1|S}(\theta|s) = \frac{s\mathrm{e}^{-s\theta}(\mathrm{e}^s + \mathrm{e}^{2s\theta})}{2(\mathrm{e}^s - 1)}$$

respectively. It is interesting to note that the asymptotic independence of the Clayton copula would indicate that $f^{\text{Cla}}_{\Theta_1|S}(\theta|s)/f^{\text{Ind}}_{\Theta_1|S}(\theta|s) \to 1$ as $s \to \infty$ which is almost the case ($f^{\text{Cla}}/f^{\text{Ind}} \to 2$). In contrast, $f^{\text{AMH}}_{\Theta_1|S}(\theta|s)$ degenerates to a pair of atoms at 0 and 1 as $s \to \infty$.

One (light-tailed) case where we can find an asymptotic angular distribution is for light-tailed Weibull sums. The angular asymptotic can be extracted from the following result in [17].

**Proposition 5.7.** *Say $X_1, \ldots, X_d$ are iid light-tailed* $\mathsf{Weibull}(\beta, 1)$ *with survival function $\overline{F}(x) = \mathrm{e}^{-x^\beta}$ where $\beta > 1$, $d \geq 2$. Define the vector function $\boldsymbol{W}(x)$ component-wise by*

$$W_i(x) = \omega(x)(X_i - x/d)\,, \qquad \text{for } i = 1, \ldots, d\,,$$

*where $\omega(x) := \sqrt{2\beta(\beta-1)(x/d)^{\beta-2}}$. Then as $\gamma \to \infty$ we have*

$$(\boldsymbol{W}(\gamma) \mid S > \gamma) \xrightarrow{\mathcal{D}} \mathsf{Normal}\left(\mathbf{0}, (1-\rho)\boldsymbol{I} + \rho\right)\,,$$

*where $\rho = -1/(d-1)$.*

When the asymptotic angular approximation is unavailable, there are several backup options. One can select a $g_{\boldsymbol{\Theta}|S}$ from some family of distributions which has the appropriate support. If $\boldsymbol{X}$ has non-negative components, then the support of $g_{\boldsymbol{\Theta}|S}$ is the simplex $\mathbb{S}^{d-1} = \{\boldsymbol{\theta} \in \mathbb{R}^d_+ : \boldsymbol{\theta}^\top \mathbf{1} = 1\}$. To the authors' knowledge, the only commonly known distribution over $\mathbb{S}^{d-1}$ is the Dirichlet distribution.

In some experiments, we sampled $(\boldsymbol{\Theta} \mid S > \gamma)$ using MCMC, then used the maximum likelihood Dirichlet fit to the samples as an angular approximation in the polar estimator.

The results were disappointing — the Dirichlet distribution struggles to fit the multimodal angular distributions which are characteristic of subexponential sums conditioned on taking large values. We also attempted the MCMC flavour of the cross-entropy method as outlined by Chan and Kroese [43], though the multimodality led to extremely high variance estimates (relative to the much simpler Asmussen–Kroese method).

We also attempted to perform an approximation of the angular density using Bernstein polynomials. The angular density $f_{\boldsymbol{\Theta}|S}(\boldsymbol{\theta} \mid s) \propto f_{\boldsymbol{X}}(s\boldsymbol{\theta})$, so it is easy to calculate quantities which are proportional to the desired conditional density. Using Bernstein polynomials effectively constructed an approximation which was a mixture of Dirichlet distributions using these unnormalised angular density values. The results for these experiments are also omitted, since the number of mixture components required to create an accurate approximation easily becomes prohibitively large (then, the computation time for evaluating the pdf of the mixture becomes an issue).

## 5.3 Results

Below we show the estimates and the estimated relative errors for the polar estimator and the Asmussen–Kroese estimator for various distributions of $\boldsymbol{X}$. Each estimator is given $R = 10^5$ iid samples of $\boldsymbol{X}$.

The first test takes the sum of $d = 16$ independent lognormals random variables, where $X_i \sim \mathsf{Lognormal}(-i/d, i/d)$. Here, the sum is asymptotically like $X_d \sim \mathsf{Lognormal}(-1, 1)$, and the optimistic angular distribution is used.

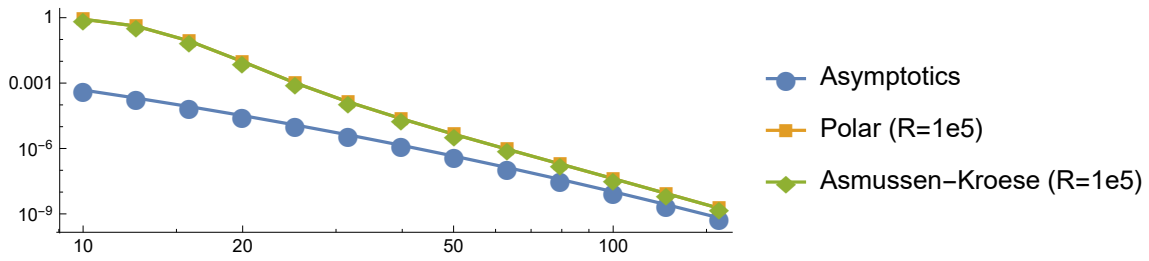



Figure 5.2: Estimates of $\mathbb{P}(S > \gamma)$ from each estimator.

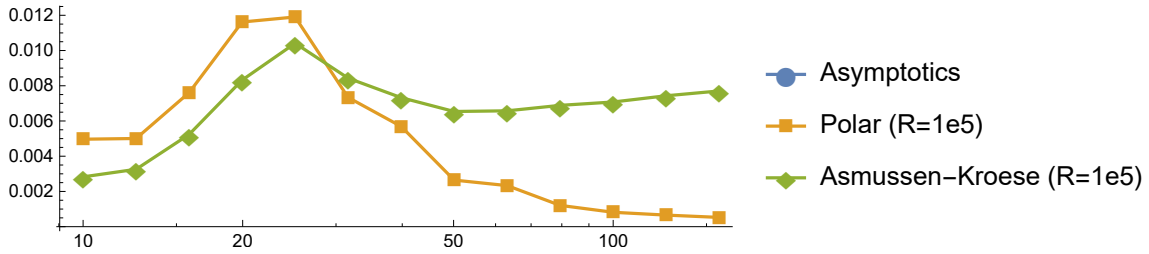



Figure 5.3: Estimated relative errors for each estimator.

The second test considers the sum of $d = 16$ independent Pareto random variables, where $X_i \sim \mathsf{Pareto}(i, 1, 0)$. The sum is asymptotically like $X_1 \sim \mathsf{Pareto}(1, 1, 0)$, and the optimistic angular distribution is used.

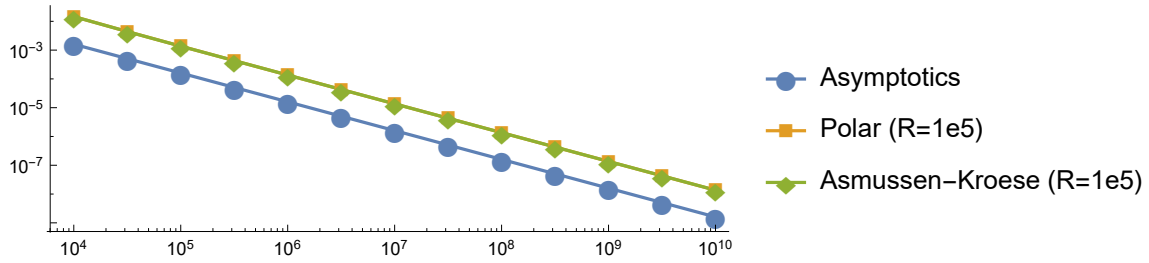



Figure 5.4: Estimates of $\mathbb{P}(S > \gamma)$ from each estimator.

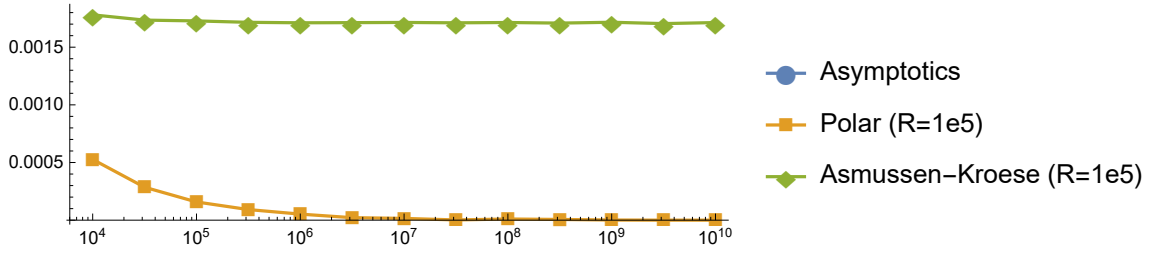



Figure 5.5: Estimated relative errors for each estimator.

The third test considers the sum of $d = 8$ independent heavy-tailed Weibull variables. The marginal distributions are $X_i \sim \mathsf{Weibull}(\frac{i}{d+1}, \frac{i}{10d})$. The sum is asymptotically like $X_1 \sim \mathsf{Weibull}(\frac{1}{9}, \frac{1}{80})$, and the optimistic angular distribution is used.

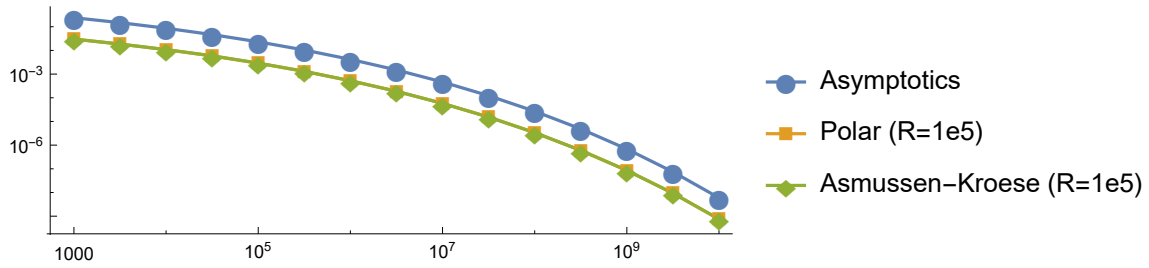



Figure 5.6: Estimates of $\mathbb{P}(S > \gamma)$ from each estimator.

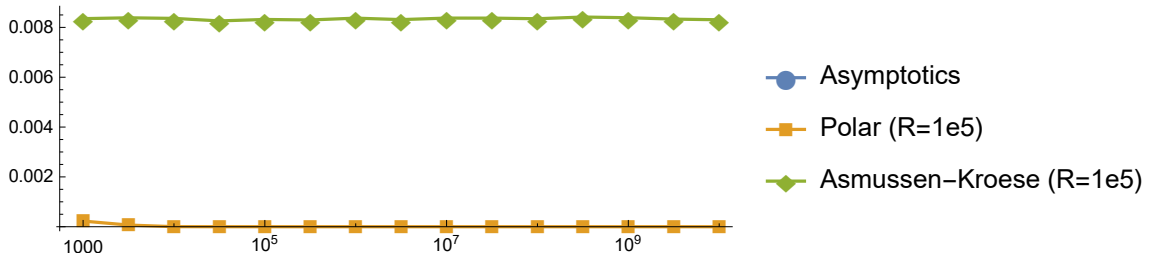



Figure 5.7: Estimated relative errors for each estimator.

The final test takes the sum of $d = 2$ iid light-tailed Weibulls, where $X_i \sim \mathsf{Weibull}(3, 1)$. An asymptotic survival function for the sum is given by Proposition 5.3, and the optimistic angular distribution used is from Proposition 5.7. Instead of the Asmussen–Kroese method, which is designed for subexponential summands, we have compared the polar estimator against *exponential tilting* (cf. (1.18) and the surrounding discussion). The exponential tilting method can be very easy to implement (in particular, when applied to distributions in the *natural exponential family*) but it takes some effort in this situation. There are no known ways to directly simulate from exponentially tilted Weibull distributions. We resort to the acceptance–rejection method with proposals coming from a gamma distribution. The specific gamma distribution is moment-matched with the asymptotic normal approximation for the exponentially tilted Weibull distribution, cf. Section 6 of [17].

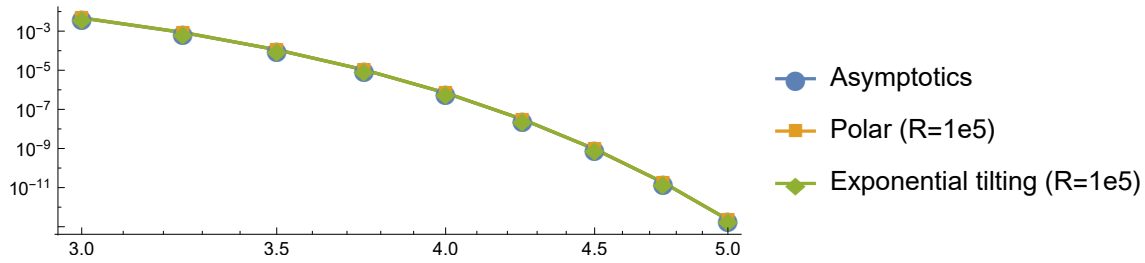



Figure 5.8: Estimates of $\mathbb{P}(S > \gamma)$ from each estimator.

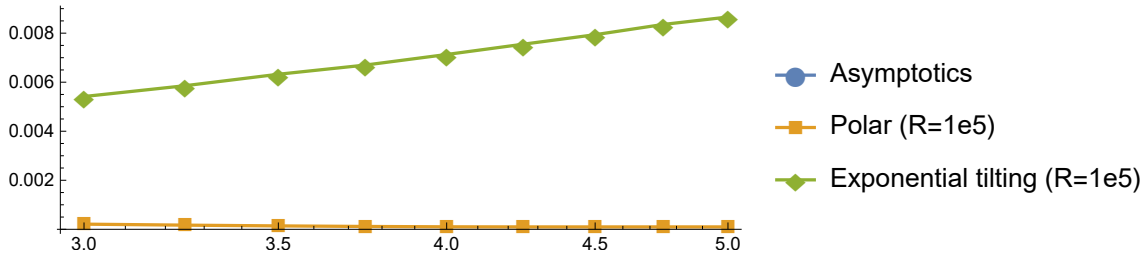



Figure 5.9: Estimated relative errors for each estimator.

## 5.4 Conclusion

On the tests performed so far, our estimator appears to perform about as well as the Asmussen–Kroese method, and outperforms all the other methods compared against (i.e. the improved cross-entropy method, fitting mixtures of Dirichlet variables, and Bernstein polynomial approximation). In particular, the polar estimator seems to consistently outperform the Asmussen–Kroese method whenever the summands are not identically distributed. In fairness, our implementation of the polar estimator is slower than the Asmussen–Kroese estimator since it is quite general, though with some optimisation the difference could be reduced. More research is needed to see whether the angular distribution can be more optimally chosen in the case of dependent summands.

## Chapter 6 Authorship Statement

**Citation:** Lars Nørvang Andersen, Patrick J. Laub, Leonardo Rojas-Nandayapa (2016), *Efficient simulation for dependent rare events with applications to extremes*, Methodology and Computing in Applied Probability

The authors of this paper equally contributed to the following tasks:

1. conception and design of the project;
2. mathematical arguments, and interpretation of the results;
3. writing the publication.

In addition to this, I completed the majority of the computational work and of the editing (e.g. checking grammar and typographical details).

# Chapter 6

# Rare maxima of random variables

## 6.1 Introduction

The distribution of the sum of random variables is, in some ways, similar to the maximum of those random variables, and many of the techniques used to analyse the former can be applied to the latter. For example, one well-known result connecting $S = \sum_{i=1} X_i$ to $M = \max_i X_i$ when the $X_i$ are subexponentially distributed is

$$\mathbb{P}(S > \gamma) \sim \mathbb{P}(M > \gamma) \qquad \text{as } \gamma \to \infty \,.$$

Another connection is that the distribution of $M$ is rarely known in a closed form, except when the $X_i$ are independent hence we ignore this simple case. In this chapter, I will give the results we achieved in approximating interesting quantities relating to these maxima distributions.

The estimators in this chapter apply to quite general problems, not just to maxima variables. We will first introduce them as relating to rare maxima of dependent random vectors, and then generalise. For a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$, the first problem we consider is estimating

$$\alpha(\gamma) = \mathbb{P}(M > \gamma) \,.$$

This problem has applications in many areas, for example in actuarial science (e.g. default

probabilities [14]), finance (e.g. probability of 'knock-out' in a barrier option [52]), survival analysis, reliability [142] and engineering (e.g. failure probability of a series circuit).

We construct estimators for this probability, which are in terms of

$$E(\gamma) = \sum_{i=1}^{d} \mathbb{I}\{X_i > \gamma\}\,,$$

the random variable which counts the number of $X_i$ which exceed $\gamma$.[1] Our two main estimators in this setting are

$$\widehat{\alpha}_1 = \sum_{i=1}^{d} \mathbb{P}(X_i > \gamma) + \frac{1}{R}\sum_{r=1}^{R}(1 - E_r(\gamma))\mathbb{I}\{E_r(\gamma) \geq 2\}\,, \text{ and} \tag{6.1}$$

$$\widehat{\alpha}_2 = \sum_{i=1}^{d} \mathbb{P}(X_i > \gamma) - \sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \mathbb{P}(X_i > \gamma, X_j > \gamma) + \frac{1}{R}\sum_{r=1}^{R}\Big[1 - E_r(\gamma) + \frac{E_r(\gamma)(E_r(\gamma)-1)}{2}\Big]\mathbb{I}\{E_r(\gamma) \geq 3\}\,. \tag{6.2}$$

where $R \in \mathbb{N}_+$ and the $E_r(\gamma)$ are derived from iid samples of $\boldsymbol{X}$. The fact that these are unbiased estimators of $\alpha(\gamma)$ follows from their construction — see equations (6.9) to (6.11) and the surrounding discussion below. Estimation of $\mathbb{P}(M > \gamma)$ is a difficult problem and treatments in the literature make distributional assumptions on $\boldsymbol{X}$. One such example is Adler et al. [6] where $\boldsymbol{X}$ is assumed to be multivariate normal. In this case, our estimator $\widehat{\alpha}_1$, with appropriate importance sampling, is the same as one of the estimators from [6].

The next problem we consider is estimating

$$\beta_n(\gamma) := \mathbb{E}[Y\mathbb{I}\{E(\gamma) \geq n\}]$$

for $n = 1, \ldots, d$ and some random variable $Y$. We do not make any assumptions of independence between the $\{X_i > \gamma\}$ events themselves or between the events and $Y$.

The subcase of $Y = 1$ a.s. has some interesting examples:

$$\beta_1(\gamma) = \mathbb{P}(M > \gamma) = \alpha(\gamma)\,, \quad \text{and} \quad \beta_n(\gamma) = \mathbb{P}(X_{(n)} > \gamma)$$

[1] We use $\mathbb{I}\{\cdot\}$ to denote the indicator function, and $\mathbb{I}\{\emptyset\} = 1$.

where $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(d)}$ are the order statistics of $\boldsymbol{X}$. The probability of a parallel circuit failing is a simple application for $\mathbb{P}(X_{(n)} > \gamma)$.

Our main $\beta_1$ estimator uses the fact that

$$\{M > \gamma\} := \bigcup_{i=1}^{d} \{X_i > \gamma\} = \{X_1 > \gamma\} \cup \left( \bigcup_{i=2}^{d} \{X_1 \leq \gamma, \ldots, X_{i-1} \leq \gamma, X_i > \gamma\} \right) \tag{6.3}$$

where the events in the union on the right are disjoint. This supplies a form of $\beta_1$ which is amenable to efficient Monte Carlo estimation:

$$\beta_1 = \sum_{i=1}^{d} \mathbb{E}[Y \, \mathbb{I}\{X_1 \leq \gamma, \ldots, X_{i-1} \leq \gamma\} \mid X_i > \gamma] \, \mathbb{P}\left(X_i > \gamma\right) . \tag{6.4}$$

As previously mentioned, while they are main example and motivation, the extremes considered so far are a very specific instance of estimators. We now turn our attention to the general set-up treated in the chapter.

Let $A(\gamma) = \cup_{i=1}^{d} A_i(\gamma)$ be the union of events $A_1(\gamma), \ldots, A_d(\gamma)$ for an index parameter $\gamma \in \mathbb{R}$. We consider the problem of estimating $\mathbb{P}(A(\gamma))$ when the events are rare, that is, $\mathbb{P}(A(\gamma)) \to 0$ as $\gamma \to \infty$. Define

$$\alpha(\gamma) := \mathbb{P}(A(\gamma)) \quad \text{and} \qquad E(\gamma) := \sum_{i=1}^{d} \mathbb{I}\{A_i(\gamma)\} .$$

Note that we recover our introductory example by having $A_i(\gamma) = \{X_i > \gamma\}$. Aside from this example, $A(\gamma)$ is quite general (a union of arbitrary events) and many interesting events arising in applied probability and statistics can be formulated as a union. The quantity $\beta_n(\gamma)$ is reminiscent of *expected shortfall* from risk management [123].

Traditional Monte Carlo methods are unreliable in the rare-event setting. We will use standard techniques from the *rare-event simulation methodology*, such as importance sampling for variance reduction and applicable measures of efficiency: *bounded relative error* and *logarithmic efficiency*, cf. [15, 79, 149]. The resulting estimators are among the most efficient possible under the most general assumptions.

The chapter is structured as follows. In Sections 6.2 and 6.3 we formally introduce our

estimators for $\alpha(\gamma)$ and $\beta_n(\gamma)$ respectively, we prove their validity, and show how to combine them with some existing variance reduction techniques; the efficiency properties for the general estimators are analysed in Section 6.4, in addition we further investigate the efficiency for certain important dependence structures. Finally, we evaluate the numerical performance of the estimators in Section 6.5.

## 6.2 Estimators of $\alpha$

In the following, we first explain the construction of our estimators of $\alpha$, then discuss possible variance reduction schemes. As the $\gamma$ notation can be cumbersome, we simply write $A = A(\gamma)$, $A_i = A_i(\gamma)$, $E = E(\gamma)$, $\alpha = \alpha(\gamma)$ and $\beta_n = \beta_n(\gamma)$. Also, we often write $\sum_i$, $\sum_{i<j}$, $\cup_i$, $\cap_i$ for $\sum_{i=1}^d$,$\sum_{1=i<j}^d$, $\cup_{i=1}^d$ and $\cap_{i=1}^d$. The $\cap$ notation is often dropped. We write $\mathbb{P}(A_i, A_j)$ instead of $\mathbb{P}(A_i \cap A_j)$, which is similar to the standard notation of $\mathbb{P}(X_i > \gamma, X_j > \gamma)$ to refer to $\mathbb{P}(\{X_i > \gamma\} \cap \{X_j > \gamma\})$. We also write $\mathbb{I}\{A_1 \dots A_j\}$ instead of $\mathbb{I}\{A_1 \cap \dots \cap A_j\}$. Lastly, we use the notation $\sum_{|I|=i}$ to refer to the summation over all subsets of indices $I \subset \{1, \dots, d\}$ such that $I$ contains $i$ indices ($|I| = i$).

### 6.2.1 Proposed estimators of $\alpha$

The inclusion–exclusion formula (IEF) provides a representation of $\alpha$ as a summation whose terms are decreasing in size. The formula is

$$\alpha = \mathbb{P}(A) = \sum_{i=1}^{d} (-1)^{i+1} \sum_{|I|=i} \mathbb{P}\Big( \bigcap_{i \in I} A_i \Big). \tag{6.5}$$

The IEF can rarely be used as its summands are increasingly difficult to calculate numerically. The $\mathbb{P}(A_i)$ terms are typically known, and the $\mathbb{P}(A_i, A_j)$ terms can frequently be calculated, however the remaining higher-dimensional terms are normally intractable for numerical integration algorithms (cf. the *curse of dimensionality* [15, Chapter IX]). Truncating the summation leads to bias, and indeed by the Bonferroni inequalities we

have:

$$\alpha \le \sum_{i=1}^{k} (-1)^{i-1} \sum_{|I|=i} \mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) \quad \text{if } 1 \le k < d \text{ and } k \text{ is odd,} \tag{6.6}$$

$$\alpha \ge \sum_{i=1}^{k} (-1)^{i-1} \sum_{|I|=i} \mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) \quad \text{if } 1 < k < d \text{ and } k \text{ is even.} \tag{6.7}$$

This higher-order intractability motivates our estimators which use the IEF rewritten in terms of $E = \sum_i \mathbb{I}\{A_i\}$.

**Proposition 6.1.** *For $i = 1, \ldots, d$,*

$$\sum_{|I|=i} \mathbb{I}\{\cap_{i\in I} A_i\} = \binom{E}{i} \mathbb{I}\{E \ge i\}\,. \tag{6.8}$$

*Proof.*

$$\sum_{|I|=i} \mathbb{I}\{\cap_{i\in I} A_i\} = \sum_{k=i}^{d} \sum_{|I|=i} \mathbb{I}\{\cap_{i\in I} A_i, E = k\} = \sum_{k=i}^{d} \binom{k}{i} \mathbb{I}\{E = k\} = \binom{E}{i} \mathbb{I}\{E \ge i\}\,.$$

□

Taking the expectation of (6.8) gives

$$\sum_{|I|=i} \mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) = \mathbb{E}\Big[\binom{E}{i} \mathbb{I}\{E \ge i\}\Big] \quad \text{for } i = 1, \ldots, d\,.$$

So the following has mean $\alpha$, and forms the nucleus of our $\widehat{\alpha}_i$ estimators:

$$\sum_{i=1}^{d} (-1)^{i-1} \binom{E}{i} \mathbb{I}\{E \ge i\}\,. \tag{6.9}$$

We present estimators which deterministically *calculate* the first larger terms of the IEF (6.5) and Monte Carlo (MC) *estimate* the remaining smaller terms using sample means of (6.8). We begin by constructing the single-replicate estimator $\widehat{\alpha}_1$ where the first summand

is calculated and the remaining terms are estimated:

$$\begin{aligned}\widehat{\alpha}_1 &:= \sum_i \mathbb{P}(A_i) + \sum_{i=2}^{d}\Big[(-1)^{i-1}\binom{E}{i}\mathbb{I}\{E \geq i\}\Big] \\ &= \sum_i \mathbb{P}(A_i) + (1-E)\mathbb{I}\{E \geq 2\}\,, \quad \text{using} \quad \sum_{k=0}^{n}(-1)^{k-1}\binom{n}{k} = 0\,. \end{aligned} \tag{6.10}$$

In identical fashion, the single-replicate estimator calculating the first two terms from the IEF is

$$\begin{aligned}\widehat{\alpha}_2 &:= \sum_i \mathbb{P}(A_i) - \sum_{i<j}\mathbb{P}(A_i, A_j) + \sum_{i=3}^{d}\Big[(-1)^{i-1}\binom{E}{i}\mathbb{I}\{E \geq i\}\Big] \\ &= \sum_i \mathbb{P}(A_i) - \sum_{i<j}\mathbb{P}(A_i, A_j) + \Big[1 - E + \frac{E(E-1)}{2}\Big]\mathbb{I}\{E \geq 3\}\,. \end{aligned} \tag{6.11}$$

Thus, for $n \in \{1, \ldots, d-1\}$,[1]

$$\widehat{\alpha}_n := \sum_{i=1}^{n}(-1)^{i-1}\sum_{|I|=i}\mathbb{P}\Big(\bigcap_{i\in I}A_i\Big) + \Big[\sum_{i=0}^{n}(-1)^i\binom{E}{i}\Big]\mathbb{I}\{E \geq n+1\}\,. \tag{6.12}$$

Thus, $\{\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{d-1}\}$ is a collection of estimators which allows the user to control the computational division of labour between numerical integration and Monte Carlo estimation. We will furthermore let $\widehat{\alpha}_0$ be the crude Monte Carlo estimator $\mathbb{I}\{E \geq 1\}$, and note that this falls under the definition in (6.12) if we interpret the empty sum as zero.

The $\widehat{\alpha}_n$ estimators are of decreasing variance in $n$, however each estimator carries the assumption that one can perform accurate numerical integration for 1 up to $n$ dimensions. As numerical integration can be slow and unreliable in high dimensions we focus on $\widehat{\alpha}_1$, and also show the numerical performance of $\widehat{\alpha}_2$.

In practice, theses estimators will exhibit very modest improvements when compared against their truncated IEF counterparts (i.e., the right side of (6.6) and (6.7)). When combined with importance sampling, as in Section 6.2.4, the improvement is marked. Furthermore, we will show that these estimators possess desirable efficiency properties which are preserved after combining with importance sampling.

[1]Note that by the IEF, we have $\widehat{\alpha}_d := \alpha$, so this possibility is ignored.

### 6.2.2 Discussion of $\widehat{\alpha}_1$ estimator

The estimator $\widehat{\alpha}_1$ has some nice interpretations. Recall the Boole–Fréchet inequalities

$$\max_i \mathbb{P}(A_i) \leq \alpha = \mathbb{P}(A) \leq \sum_i \mathbb{P}(A_i) =: \overline{\alpha}\,. \tag{6.13}$$

The stochastic part of $\widehat{\alpha}_1$ is an unbiased estimate of $\overline{\alpha} - \alpha \leq 0$. That is to say, $\widehat{\alpha}_1$ MC estimates the difference between the target quantity $\alpha$ and its upper bound given by the Boole–Fréchet inequalities, $\overline{\alpha}$. Similarly, we often have

$$\alpha(\gamma) \sim \sum_i \mathbb{P}(A_i(\gamma))\,,^{1}$$

for example when the $A_i$ exhibit a weak dependence structure. In this case, we can say that $\widehat{\alpha}_1$ MC estimates the difference between $\alpha$ and its (first-order) asymptotic expansion.

### 6.2.3 Relation of $\widehat{\alpha}_n$ estimators to control variates

An alternative construction of $\{\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{d-1}\}$ is to add *control variates* to the crude Monte Carlo estimator $\widehat{\alpha}_0$. We begin by adding the control variate $E$ to $\widehat{\alpha}_0$ with weight $\tau \in \mathbb{R}$:

$$\widehat{\alpha}_1^{\tau} := \mathbb{I}\{E \geq 1\} - \tau\Big[E - \sum_i \mathbb{P}(A_i)\Big]\,.$$

Setting $\tau = 1$ means this estimator simplifies to $\widehat{\alpha}_1$. Next, we add the control variates $E$ and $-\frac{1}{2}E(E-1)$ to $\widehat{\alpha}_0$, and setting the corresponding weights to 1 gives $\widehat{\alpha}_2$. This pattern goes on.

### 6.2.4 Combining $\widehat{\alpha}_1$ with importance sampling

The family of estimators $\widehat{\alpha}_n$ can be combined with the variance reduction technique called *importance sampling* (IS), cf. [15, 79]. Standard IS theory suggests that we should focus on IS distributions where the event of interest $A = \cup_i A_i = \{E \geq 1\}$ occurs almost surely.

[1] Using the standard notation that $f(x) \sim g(x)$ means $\lim_{x\to\infty} f(x)/g(x) = 1$.

A convenient way of constructing such a distribution is as a *mixture distribution*. Say that we condition on $A_i$ with probability

$$p_i := \frac{\mathbb{P}(A_i)}{\sum_j \mathbb{P}(A_j)} = \frac{\mathbb{P}(A_i)}{\overline{\alpha}}\,, \qquad \text{for } i = 1, \ldots, d\,.$$

A heuristic motivation for this selection comes from a rare-event setting where the asymptotic relationship $\mathbb{P}(A_i(\gamma), A_j(\gamma)) = \mathrm{o}(\mathbb{P}(A_i(\gamma)))$ often occurs for all $i \neq j$. In such a case

$$\mathbb{P}\left(A_i(\gamma) \mid A(\gamma)\right) = \frac{\mathbb{P}(A_i(\gamma))}{\sum_j \mathbb{P}(A_j(\gamma))(1 + \mathrm{o}(1))} \sim p_i(\gamma)\,, \qquad \text{as } \gamma \to \infty\,.$$

Now consider the measure

$$\mathbb{Q}^{[1]}(\mathscr{A}) = \sum_i p_i\, \mathbb{P}(\mathscr{A} \mid A_i) \qquad \forall \mathscr{A} \in \mathcal{F}\,,$$

which induces the likelihood ratio of $L^{[1]} := \mathrm{d}\,\mathbb{Q}^{[1]} \,/\, \mathrm{d}\,\mathbb{P} = \overline{\alpha}/E$. As

$$\overline{\alpha} + (1 - E)\mathbb{I}\{E \geq 2\}L^{[1]} = \overline{\alpha}\Big(1 + \frac{1 - E}{E}\Big) = \frac{\overline{\alpha}}{E} \quad \text{ under } \mathbb{Q}^{[1]}\,,$$

we can see that $\widehat{\alpha}_1$ under this change of measure, with $R \in \mathbb{N}_+$ replicates, is

$$\widehat{\alpha}_1^{[1]} := \frac{1}{R}\sum_{r=1}^{R} \frac{\overline{\alpha}}{E_r^{[1]}}\,, \tag{6.14}$$

where the superscript "[1]" indicates that the $E_r^{[1]}$ are (independently) sampled under $\mathbb{Q}^{[1]}$. This estimator corresponds to one from the paper of Adler et al. [5], though applied in a more general way (they consider rare maxima of normally distributed vectors).

Continuing in the same pattern, consider the *second-order* IS distributions where $\{E \geq 2\}$ occurs almost surely, to be applied to $\widehat{\alpha}_2$. Say that we choose to condition on $A_i \cap A_j$ with probability

$$p_{ij} := \frac{\mathbb{P}(A_i, A_j)}{\sum_{m<n} \mathbb{P}(A_m, A_n)} = \frac{\mathbb{P}(A_i, A_j)}{q}\,, \qquad \text{for } 1 \leq i < j \leq d\,,$$

defining $q := \sum_{i<j} \mathbb{P}(A_i, A_j)$. Now consider the measure

$$\mathbb{Q}^{[2]}(\mathscr{A}) = \sum_{i<j} p_{ij}\, \mathbb{P}(\mathscr{A} \mid A_i, A_j) \qquad \forall \mathscr{A} \in \mathcal{F}\,,$$

which induces a likelihood ratio of

$$L^{[2]} := \frac{\mathrm{d}\,\mathbb{Q}^{[2]}}{\mathrm{d}\,\mathbb{P}} = \frac{q}{\sum_{i<j} \mathbb{I}\{A_i A_j\}} = \frac{q}{\binom{E}{2}} = \frac{2q}{E(E-1)}\,.$$

Thus, after simplifying, the estimator $\widehat{\alpha}_2$ under $\mathbb{Q}^{[2]}$ is

$$\widehat{\alpha}_2^{[2]} := \overline{\alpha} - \frac{2q}{R} \sum_{r=1}^{R} \frac{1}{E_r^{[2]}}\,. \tag{6.15}$$

**Remark 6.2.** As the $\mathbb{Q}^{[2]}$-mean of $\frac{2}{E}$ is less than 1, this fraction can be seen as a correction term for the two-term truncation of (6.5). We know from (6.7) that $\alpha \geq \overline{\alpha} - q$. $\diamond$

Both of these IS algorithms have some extra requirements for their use. The first-order estimators require that we can simulate from $\mathbb{P}(\,\cdot \mid A_i)$ and can calculate the $\mathbb{P}(A_i)$. The second-order estimator requires that we can simulate from $\mathbb{P}(\,\cdot \mid A_i, A_j)$ and that we can calculate the $\mathbb{P}(A_i)$ and $\mathbb{P}(A_i, A_j)$. In the rare maxima case, integration routines in MATHEMATICA or MATLAB can usually calculate these probabilities; it is simulating from the conditional distributions which can be the prohibitive requirement, particularly for $\widehat{\alpha}_2^{[2]}$.

## 6.3 Estimators of $\beta_n$

Now, we turn our attention to the estimation of $\beta_n := \mathbb{E}[Y\mathbb{I}\{E \geq n\}]$. We start with $\beta_1$, and rewrite the partition (6.3) in terms of the general $A_i$:

$$A := \bigcup_{i=1}^{d} A_i = A_1 \cup (A_1^{\mathsf{c}} A_2) \cup \cdots \cup (A_1^{\mathsf{c}} \ldots A_{d-1}^{\mathsf{c}} A_d)\,. \tag{6.16}$$

This gives us (the generalised version of (6.4))

$$\begin{aligned}\beta_1 = \mathbb{E}[Y \mid A_1]\,\mathbb{P}(A_1) + \mathbb{E}[Y\mathbb{I}\{A_1\} \mid A_2]\,\mathbb{P}(A_2) \\ + \cdots + \mathbb{E}[Y\mathbb{I}\{A_1^{\mathsf{c}} \ldots A_{d-1}^{\mathsf{c}}\} \mid A_d]\,\mathbb{P}(A_d)\,.\end{aligned}$$

If we assume it is possible to sample from the $\mathbb{P}(\,\cdot \mid A_i)$ conditional distributions—the same assumption required to use the first-order IS estimator $\widehat{\alpha}_1^{[1]}$ from Section 6.2.4—then each of these conditional expectations can be estimated by sample means:

$$\widehat{\beta}_1 := \sum_{i=1}^{d} \frac{\mathbb{P}(A_i)}{\lceil R/d \rceil} \sum_{r=1}^{\lceil R/d \rceil} Y_{i,r}\mathbb{I}\{A_1^{\mathsf{c}} \ldots A_{i-1}^{\mathsf{c}}\}_{i,r}\,. \tag{6.17}$$

Here, the $Y_{i,r}$ and $\mathbb{I}\{\cdot\}_{i,r}$ are sampled independently and conditional on $A_i$. The following proposition gives the partition of the event $\{E \geq i\}$:

**Proposition 6.3.** *Consider a finite collection of events $\{A_1, \ldots, A_d\}$ and for each subset $I \subset \{1, 2, \ldots, d\}$ define* [1]

$$B_I := \bigcap_{j \in I} A_j, \qquad C_I := \bigcap_{\substack{k \notin I,\\ k < \max I}} A_k^c.$$

*Then*

$$\{E \geq m\} = \bigcup_{|I|=m} B_I = \bigcup_{|I|=m} B_I C_I. \tag{6.18}$$

*Moreover, the collection of sets $\{B_I C_I : |I| = m\}$ is disjoint.*

*Proof.* The first equality in (6.18) is straightforward from the definition of the random variable $E$. For the second equality, the relation $\supseteq$ follows trivially; to prove the opposite relation $\subseteq$ it remains to show that if $\omega$ is such that $\omega \in B_I$ and $\omega \notin C_I$, then there exists $I'$ such that $|I'| = m$ and $\omega \in B_{I'}C_{I'}$. Notice that if $\omega \notin C_I$, then there exists a nonempty set $J$ satisfying $\max J < \max I$, with $j \in J$ if and only if $\omega \notin A_j^c$. Select $I'$ as the set

[1] Using the convention that $\cap_\emptyset = \Omega$.

formed by the smaller $m$ elements of $I \cup J$. In consequence,

$$\omega \in \Big(\bigcap_{j\in I\cup J} A_j\Big)\Big(\bigcap_{\substack{k\notin I\cup J,\\ k\le \max I}} A_k^c\Big) \subseteq \Big(\bigcap_{j\in I'} A_j\Big)\Big(\bigcap_{\substack{k\notin I',\\ k\le \max I'}} A_k^c\Big) = B_{I'}C_{I'}.$$

This completes the proof of the second equivalence in (6.18).

Next we show that the collection of sets $\{B_I C_I : |I| = m\}$ is disjoint. Consider two sets of indexes $I_1$ and $I_2$ such that $|I_1| = |I_2| = m$ and $I_1 \neq I_2$. Take $i$ such that $i \in I_1$, $i \notin I_2$ and w.l.o.g. further assume that $i < \max I_2$. Then $B_{I_1} \subseteq A_i$ while $C_{I_2} \subseteq A_i^c$. □

This proposition implies that

$$\beta_n = \mathbb{E}\Big[Y\mathbb{I}\Big\{\bigcup_{|I|=n} B_I\Big\}\Big] = \mathbb{E}\Big[Y\mathbb{I}\Big\{\bigcup_{|I|=n} B_I C_I\Big\}\Big] = \sum_{|I|=n} \mathbb{E}[Y\mathbb{I}\{C_I\} \mid B_I]\,\mathbb{P}(B_I).$$

Therefore, if (i) reliable estimates of $\mathbb{P}(B_I)$ are available, and (ii) it is possible to simulate from the conditional measures $\mathbb{P}(\,\cdot \mid B_I)$, then the following is an unbiased estimator of $\mathbb{E}[Y\mathbb{I}\{E \ge n\}]$:

$$\widehat{\beta}_n := \sum_{|I|=n} \frac{\mathbb{P}(B_I)}{\lceil R/\binom{d}{n}\rceil} \sum_{r=1}^{\lceil R/\binom{d}{n}\rceil} Y_{I,r}\mathbb{I}\{C_I\}_{I,r}\,. \tag{6.19}$$

Here, similar to before, $Y_{I,r}$ and $\mathbb{I}\{\cdot\}_{I,r}$ denote independent sampling conditioned on $B_I$.

Notice that a permutation of the sets $A_1, \ldots, A_d$ will result in a different collection of events $C_I$, and also a slightly different estimator.

### 6.3.1 Applying $\widehat{\beta}_i$ to estimate $\alpha$

The $\widehat{\beta}_i$ estimators can be used in various ways to estimate the probability $\alpha = \mathbb{P}(A)$. The simplest way is to set $Y = 1$ a.s. in $\widehat{\beta}_1$ (6.19), leading to the estimator

$$\widehat{(\beta_1 \ddagger \alpha)} := \mathbb{P}(A_1) + \sum_{i=2}^{d} \frac{\mathbb{P}(A_i)}{\lceil R/(d-1)\rceil} \sum_{r=1}^{\lceil R/(d-1)\rceil} \mathbb{I}\{A_1^{\mathsf{c}} \ldots A_{i-1}^{\mathsf{c}}\}_{i,r}\,, \tag{6.20}$$

using the notation from (6.17). Note, we achieve minor improvement in (6.20) over (6.19) when $Y = 1$ a.s. as $\mathbb{E}[1 \mid A_1] = 1$ does not require estimation.

More effective estimators can be constructed if we use $\widehat{\beta}_n$ to estimate terms from $\widehat{\alpha}_{n-1}$ (6.12). We label the random terms in $\widehat{\alpha}_n$ as

$$R_n := \Big[\sum_{i=0}^{n}(-1)^i\binom{E}{i}\Big]\mathbb{I}\{E \geq n+1\}, \quad \text{and say} \quad \mathcal{R}_n := \mathbb{E}[R_n]\,. \tag{6.21}$$

Now, if we choose $Y := \sum_{i=0}^{n-1}(-1)^i\binom{E}{i}$ then it is obvious that

$$\beta_n := \mathbb{E}\Big\{\Big[\sum_{i=0}^{n-1}(-1)^i\binom{E}{i}\Big]\mathbb{I}\{E \geq n\}\Big\} = \mathcal{R}_{n-1}.$$

This leads to the set of estimators

$$\begin{aligned}\widehat{(\beta_n \ddagger \alpha)} := &\sum_{i=1}^{n-1}(-1)^{i-1}\sum_{|I|=i}\mathbb{P}\Big(\bigcap_{i\in I}A_i\Big)\\ &+\sum_{|I|=n}\frac{\mathbb{P}(B_I)}{\lceil R/\binom{d}{n}\rceil}\sum_{r=1}^{\lceil R/\binom{d}{n}\rceil}\Big[\sum_{i=0}^{n-1}(-1)^i\binom{E_{I,r}}{i}\Big]\mathbb{I}\{E \geq n\}_{I,r}\,,\end{aligned}$$

for $n = 2, \ldots d-1$. In particular, for $n = 2$

$$\widehat{(\beta_2 \ddagger \alpha)} := \sum_i \mathbb{P}(A_i) + \sum_{i<j}\frac{\mathbb{P}(A_i, A_j)}{\lceil R/\binom{d}{2}\rceil}\sum_{r=1}^{\lceil R/\binom{d}{2}\rceil}(1 - E_{ij,r})\mathbb{I}\{E \geq 2\}_{ij,r}\,, \tag{6.22}$$

where the $ij$ subscript indicates sampling conditional on $A_iA_j$, similar to before.

## 6.4 Efficiency results

We analyse the performance of the estimators in a rare-event setting. Recall that in such a setting, $\{A_1(\gamma), \ldots, A_d(\gamma)\}$ denotes an indexed collection of not necessarily independent rare events and our objective is to calculate $\alpha(\gamma) = \mathbb{P}(\bigcup_i^d A_i(\gamma))$ as $\gamma \to \infty$. For such a *rare-event* estimation problem there are specialised concepts of efficiency. In Section 6.4.2

these definitions of efficiency are introduced. In addition, we provide efficiency criteria for the proposed estimators under very general assumptions.

In Sections 6.4.3 and 6.4.4 we specialise in rare events associated with extremes. In such a framework, we show when the estimator $\widehat{\alpha}_1$ is efficient for: i) a vast array of multivariate distributions with identical marginals in Section 6.4.3, and ii) the specific cases of normal and elliptical distributions in Section 6.4.4. For this section we take the number of replicates $R$ to be 1.

### 6.4.1 Variance Reduction

First we compare the efficiency of our proposed estimator $\widehat{\alpha}_1$ against that of the crude Monte Carlo (CMC) estimator $\widehat{\alpha}_0(\gamma)$ of $\alpha(\gamma) := \mathbb{P}(A(\gamma))$. An upper bound for $\mathbb{V}\mathrm{ar}\,\widehat{\alpha}_0(\gamma)$ is

$$\mathbb{V}\mathrm{ar}\,\widehat{\alpha}_0(\gamma) = \mathbb{P}(A(\gamma))[1 - \mathbb{P}(A(\gamma)] < \mathbb{P}(A(\gamma)) \leq \sum_i \mathbb{P}(A_i(\gamma))\,.$$

This implies that the variance of the CMC estimator is of order $\mathcal{O}(\max_i \mathbb{P}(A_i(\gamma)))$, which is the best possible without making any further assumptions. In contrast an upper bound of $\mathbb{V}\mathrm{ar}\,\widehat{\alpha}_1(\gamma) = \mathbb{V}\mathrm{ar}\,R_1$, where $R_1 = (1-E)\mathbb{I}\{E \geq 2\}$ from (6.21), is

$$\mathbb{V}\mathrm{ar}\,\widehat{\alpha}_1(\gamma) \leq \mathbb{E}[R_1^2] < 2\,\mathbb{E}\Big[\binom{E}{2}\mathbb{I}\{E \geq 2\}\Big] \underset{(6.8)}{=} 2\sum_{i<j}\mathbb{P}(A_i(\gamma), A_j(\gamma))\,. \tag{6.23}$$

Thus the variance of our estimator $\widehat{\alpha}_1(\gamma)$ is of order $\mathcal{O}(\max_{i<j}\mathbb{P}(A_i(\gamma), A_j(\gamma)))$, so we can conclude that $\widehat{\alpha}_1(\gamma)$ is asymptotically superior to CMC.

Next we turn our attention to the estimator $\widehat{\beta}_n$. The following proposition shows that the reduction of variance of the estimator $\widehat{\beta}_n$ is of at least of a factor $\max_{|I|=n}\mathbb{P}(B_I)$ with respect to the non-conditional (crude) version estimator $\widehat{\beta}_n^{[0]}$ defined as

$$\widehat{\beta}_n^{[0]} := \sum_{|I|=n} \frac{1}{\lceil R/\binom{d}{n}\rceil} \sum_{r=1}^{\lceil R/\binom{d}{n}\rceil} Y_{Ir}\mathbb{I}\{B_I C_I\}\,. \tag{6.24}$$

**Proposition 6.4.**

$$\mathbb{V}\mathrm{ar}(\widehat{\beta}_n) \leq \max_{|I|=n} \mathbb{P}(B_I)\, \mathbb{V}\mathrm{ar}(\widehat{\beta}_n^{[0]})\,.$$

*Proof.* Let $W_I := Y\mathbb{I}\{C_I\}$. By independence of the $W_I$ we can write the variance of $\widehat{\beta}_n$ as

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\widehat{\beta}_n) = \mathbb{V}\mathrm{ar}\Big(\sum_{|I|=n} W_I\, \mathbb{P}(B_I) \,\Big|\, B_I\Big) &= \sum_{|I|=n} \mathbb{P}(B_I)^2\, \mathbb{V}\mathrm{ar}(W_I \mid B_I) \\ &\leq \max_{|I|=n} \mathbb{P}(B_I) \sum_{|I|=n} \mathbb{P}(B_I)\, \mathbb{V}\mathrm{ar}(W_I \mid B_I)\,.\end{aligned}$$

Now, observe that

$$\begin{aligned}\mathbb{P}(B_I)\, \mathbb{V}\mathrm{ar}(W_I \mid B_I) &\leq \mathbb{P}(B_I)\, \mathbb{E}[W_I^2 \mid B_I] - \mathbb{P}(B_I)^2\, \mathbb{E}[W_I \mid B_I]^2 \\ &= \mathbb{E}[W_I^2 \mathbb{I}\{B_I\}] - \mathbb{E}[W_I \mathbb{I}\{B_I\}]^2 = \mathbb{V}\mathrm{ar}\left[W_I \mathbb{I}\{B_I\}\right].\end{aligned}$$

Thus we have proven that

$$\mathbb{V}\mathrm{ar}(\widehat{\beta}_n) \leq \max_{|I|=n} \mathbb{P}(B_I) \sum_{|I|=n} \mathbb{V}\mathrm{ar}(W_I \mathbb{I}\{B_I\}) = \max_{|I|=n} \mathbb{P}(B_I) \sum_{|I|=n} \mathbb{V}\mathrm{ar}(\widehat{\beta}_0)\,.$$

□

### 6.4.2 Efficiency criteria

We now ask if and when $\widehat{\alpha}_1$ and $\widehat{\beta}_n$ are efficient in the rare-event sense. We must first define efficiency, as there are several common benchmarks for the efficiency of a rare-event estimator.

**Definition 6.5.** *An estimator $\widehat{p}_\gamma$ of some rare probability $p_\gamma$ which satisfies $\forall \varepsilon > 0$*

$$\limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{p}_\gamma}{p_\gamma^{2-\varepsilon}} = 0 \tag{6.25a}$$

$$\limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{p}_\gamma}{p_\gamma^{2}} < \infty \tag{6.25b}$$

$$\limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{p}_\gamma}{p_\gamma^{2}} = 0 \tag{6.25c}$$

*has* logarithmic efficiency *(LE)* (6.25a), bounded relative error *(BRE)* (6.25b), *or* vanishing relative error *(VRE)* (6.25c) *respectively.*

The levels of efficiency in Definition 6.5 are given in increasing order of strength, that is, VRE $\Rightarrow$ BRE $\Rightarrow$ LE. As VRE is often too difficult a goal, we focus on BRE and LE. The following proposition gives an alternative form of the conditions in (6.25) for the specific case of our estimator $\widehat{\alpha}_1$.

**Proposition 6.6.** *The estimator $\widehat{\alpha}_1$ has LE iff it holds that $\forall \varepsilon > 0$*

$$\limsup_{\gamma\to\infty} \frac{\max_{i<j} \mathbb{P}(A_i(\gamma), A_j(\gamma))}{\max_k \mathbb{P}(A_k(\gamma))^{2-\varepsilon}} = 0\,, \tag{6.26}$$

*and has BRE iff*

$$\limsup_{\gamma\to\infty} \frac{\max_{i<j} \mathbb{P}(A_i(\gamma), A_j(\gamma))}{\max_k \mathbb{P}(A_k(\gamma))^{2}} < \infty\,. \tag{6.27}$$

*Proof.* We prove the LE claim (6.26). Proof of the BRE claim follows the same arguments. ($\Rightarrow$) We can see that

$$\mathbb{V}\mathrm{ar}\,\widehat{\alpha}_1(\gamma) \geq \mathbb{V}\mathrm{ar}\,\mathbb{I}\{E \geq 2\} = \mathbb{P}(E \geq 2)\ \mathbb{P}(E \leq 1)\,, \quad \mathbb{P}(E \leq 1) \to 1\,, \tag{6.28}$$

and

$$\mathbb{P}(E \geq 2) \geq \binom{d}{2}^{-1} \sum_{n=2}^{d} \binom{n}{2} \mathbb{P}(E = n) \underset{(6.8)}{=} \binom{d}{2}^{-1} \sum_{i<j} \mathbb{P}(A_i(\gamma), A_j(\gamma))\,. \tag{6.29}$$

So, $\forall \varepsilon > 0$,

$$\begin{aligned} 0 &\underset{(6.25a)}{=} \limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{\alpha}_1(\gamma)}{\mathbb{P}(A)^{2-\varepsilon}} \underset{(6.13)\,\&\,(6.28)}{>} \limsup_{\gamma\to\infty} \frac{\mathbb{P}(E \geq 2)}{\left(\sum_k \mathbb{P}(A_k(\gamma))\right)^{2-\varepsilon}} \\ &\underset{(6.29)}{\geq} \left[d^{2-\varepsilon}\binom{d}{2}\right]^{-1} \limsup_{\gamma\to\infty} \frac{\max_{i<j} \mathbb{P}(A_i(\gamma), A_j(\gamma))}{\max_k \mathbb{P}(A_k(\gamma))^{2-\varepsilon}} \end{aligned}$$

which implies (6.26).

($\Leftarrow$) We can see that, $\forall \varepsilon > 0$,

$$\limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}\,\widehat{\alpha}_1(\gamma)}{\mathbb{P}(A)^{2-\varepsilon}} \underset{(6.13)\,\&\,(6.23)}{<} \limsup_{\gamma\to\infty} \frac{2\sum_{i<j}\mathbb{P}(A_i(\gamma),A_j(\gamma))}{(\max_k \mathbb{P}(A_k(\gamma)))^{2-\varepsilon}}$$
$$\leq 2\binom{d}{2} \limsup_{\gamma\to\infty} \frac{\max_{i<j}\mathbb{P}(A_i(\gamma),A_j(\gamma))}{\max_k \mathbb{P}(A_k(\gamma))^{2-\varepsilon}} \underset{(6.26)}{=} 0\,,$$

which implies (6.25a). □

**Example 6.7.** If the $A_i$ events are independent then the estimator $\widehat{\alpha}_1$ has BRE.

For the efficiency of our $\widehat{\beta}_n$ estimators, the following proposition provides a very simple yet non-trivial condition for BRE.

**Proposition 6.8.** *The estimator $\widehat{\beta}_n(\gamma)$ has BRE if*

$$\limsup_{\gamma\to\infty} \frac{\max_{|I|=n}\mathbb{P}(B_I)}{\beta_n(\gamma)} < \infty.$$

*Proof.* By Proposition 6.4 and the hypothesis we have

$$\limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}(\widehat{\beta}_n(\gamma))}{\beta_n^2(\gamma)} \leq \limsup_{\gamma\to\infty} \frac{\max_{|I|=n}\mathbb{P}(B_I)\,\mathbb{V}\mathrm{ar}(\widehat{\beta}_n^{[0]}(\gamma))}{\beta_n^2(\gamma)}$$
$$\leq c\,\limsup_{\gamma\to\infty} \frac{\mathbb{V}\mathrm{ar}(\widehat{\beta}_n^{[0]}(\gamma))}{\beta_n(\gamma)}\,.$$

Since $\widehat{\beta}_n^{[0]}$ is an estimator in crude form then $\mathbb{V}\mathrm{ar}(\widehat{\beta}_n^{[0]}(\gamma)) = \mathcal{O}(\beta_n(\gamma))$ as $\gamma \to \infty$, so the proof is complete. □

**Corollary 6.9.** *The estimator $\widehat{(\beta_1 \ddagger \alpha)}$ from* (6.20) *has BRE.*

### 6.4.3 Efficiency for identical marginals and dependence

In this and the following subsections, we concentrate on rare events associated to extremes. More precisely, we let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an arbitrary random vector and define $M = \max_i X_i$. Therefore, we define $A_i(\gamma) = \{X_i > \gamma\}$ implying that the event of interest $A$ is equivalent to $\{M > \gamma\}$.

For now, we assume that the $X_i$ have identical marginal distributions. This simplifies the condition for BRE of $\widehat{\alpha}_1$, (6.27), so that it is now solely determined by the *copula* of $\boldsymbol{X}$. We investigate some common tail dependence measures of copulas (tail dependence parameter and residual tail index) and also some common families of copulas (Archimedean copulas) to see when the estimator $\widehat{\alpha}_1$ exhibits efficiency.

### Asymptotic dependence

The most basic measurement of tail dependence between a pair $(X_i, X_j)$ with common marginal distribution $F$ and copula $C_{ij}$ (cf. [103, 132]) is

$$\lambda_{ij} = \lim_{\upsilon \to 1} \mathbb{P}(X_i > \upsilon \mid X_j > \upsilon) = \lim_{\upsilon \to 1} \frac{1 - 2\upsilon + C_{ij}(\upsilon, \upsilon)}{1 - \upsilon}$$

where $\lambda_{ij} \in [0, 1]$ is called the *(upper) tail dependence parameter (or coefficient)* [103, 123]. We say the $(X_i, X_j)$ pair exhibit *asymptotic independence* (AI) when $\lambda_{ij} = 0$, or *asymptotic dependence* (AD) when $\lambda_{ij} > 0$. The canonical examples given for each case are the (non-degenerate) bivariate normal distribution for AI, and the bivariate Student $t$ distribution for AD [155].

For $\widehat{\alpha}_1$ to have BRE, all pairs in $\boldsymbol{X}$ must exhibit AI. This is a necessary but not sufficient condition, therefore we will employ a more refined tail dependence measurement.

### Residual tail index

We must first define two classes of functions:

- $L(x)$ is *slowly-varying* (at $\infty$) if $L(cx)/L(x) \to 1$ as $x \to \infty$ for all $c > 0$,
- $f(x)$ is *regularly-varying* (at $\infty$) with index $\tau > 0$ if it takes the form $f(x) = L(x)x^{-\tau}$ for some $L(x)$ which is slowly-varying (cf. [33, 144]).

We will assume, w.l.o.g., the marginals of $\boldsymbol{X}$ to be unit Fréchet distributed (i.e., $F_1(x) = \exp(-x^{-1}) \sim 1 - x^{-1}$). Ledford and Tawn [113, 114, 115] first noted that the joint survival

functions for a wide array of bivariate distributions satisfy

$$\mathbb{P}(X_i > \gamma, X_j > \gamma) \sim L(\gamma)\gamma^{-1/\eta} \qquad \text{as } \gamma \to \infty \tag{6.30}$$

for a slowly-varying $L(\gamma)$ and an $\eta \in (0, 1]$. In other words, (6.30) says that $\mathbb{P}(X_i > \gamma, X_j > \gamma)$ is regularly-varying with index $1/\eta$.

The index is called the *residual tail index* [58, 134].[1] When $(X_i, X_j)$ exhibit AD (AI) then we typically have $\eta = 1$ ($\eta < 1$).[2] For independent components we have $\eta = 1/2$, so Ledford and Tawn [113] describe bivariate distributions with $\eta = 1/2$ as having *near independence*. When $\eta < 1/2$ the random pair take large values together less frequently than they would if independent.

Returning to our original problem of estimating $\alpha(\gamma)$, let us label the residual tail index for every $(X_i, X_j)$ pair of $\boldsymbol{X}$ as $\eta_{ij}$. Also, let $\eta = \max_{ij} \eta_{ij}$ and $L$ be the associated slowly varying function. The following proposition outlines how these values relate to efficiency of $\widehat{\alpha}_1$:

**Proposition 6.10.** *If* (6.30) *is satisfied for the maximal pair of* $\boldsymbol{X}$*, that is,*

$$\max_{i<j} \mathbb{P}(X_i > \gamma, X_j > \gamma) \sim L(\gamma)\gamma^{-1/\eta} \qquad \textit{as } \gamma \to \infty\,,$$

*then the estimator* $\widehat{\alpha}_1$ *has: i) BRE if* $\eta < 1/2$ *or if* $\eta = 1/2$ *and* $L(\gamma) \not\to \infty$ *as* $\gamma \to \infty$*, ii) LE if* $\eta = 1/2$*.*

*Proof.* Label the components of $\boldsymbol{X}$ such that

$$\max_{i<j} \mathbb{P}(X_i > \gamma, X_j > \gamma) = \mathbb{P}(X_1 > \gamma, X_2 > \gamma)$$

then the condition for LE becomes, $\forall \varepsilon > 0$

$$\limsup_{\gamma\to\infty} \frac{\max_{i<j} \mathbb{P}(X_i \geq \gamma, X_j \geq \gamma)}{\max_k \mathbb{P}(X_k \geq \gamma)^{2-\varepsilon}} = \limsup_{\gamma\to\infty} \frac{L(\gamma)\gamma^{-1/\eta}}{(\gamma^{-1})^{2-\varepsilon}} = \limsup_{\gamma\to\infty} L(\gamma)\gamma^{2-\frac{1}{\eta}-\varepsilon} = 0$$

which is equivalent to $\eta \in (0, 1/2]$; the $\eta = 1/2$ case has LE as $\gamma^{-\varepsilon}L(\gamma) \to 0$ for all $\varepsilon > 0$

[1] The older (and less insightful) name for $\eta$ is the *coefficient of tail dependence* [113, 143].

[2] Hashorva [91] has found a case where an elliptically distributed $(X_i, X_j)$ has $\eta = 1$ and AI.

Table 6.1: Residual tail dependence index $\eta$ and $L(x)$ for various copulas. This is a subset of Table 1 of [96] (their row numbers are preserved).

(a) Copulas with BRE.

| # | **Name** | $\eta$ | $L(x)$ |
|---|---|---|---|
| 1 | Ali-Mikhail-Haq | 0.5 | $1+\tau$ |
| 2 | BB10 in Joe | 0.5 | $1+\theta/\tau$ |
| 3 | Frank | 0.5 | $\delta/(1-\mathrm{e}^{-\delta})$ |
| 4 | Morgenstern | 0.5 | $1+\tau$ |
| 5 | Plackett | 0.5 | $\delta$ |
| 6 | Crowder | 0.5 | $1+(\theta-1)/\tau$ |
| 7 | BB2 in Joe | 0.5 | $\theta(\delta+1)+1$ |
| 8 | Pareto | 0.5 | $1+\delta$ |
| 9 | Raftery | 0.5 | $\delta/(1-\delta)$ |

(b) Copulas without BRE.

| # | **Name** | $\eta$ | $L(x)$ |
|---|---|---|---|
| 11 | Joe | 1 | $2-2^{1/\delta}$ |
| 12 | BB8 in Joe | 1 | $2-2(1-\delta)^{\theta-1}$ |
| 13 | BB6 in Joe | 1 | $2-2^{1/(\delta\theta)}$ |
| 14 | Extreme value | 1 | $2-V(1,1)$ |
| 15 | B11 in Joe | 1 | $\delta$ |
| 16 | BB1 in Joe | 1 | $2-2^{1/\delta}$ |
| 17 | BB3 in Joe | 1 | $2-2^{1/\theta}$ |
| 18 | BB4 in Joe | 1 | $2^{-1/\delta}$ |
| 19 | BB7 in Joe | 1 | $2-2^{1/\theta}$ |

(see Proposition 1.3.6 part (v) of [33]). Similarly we have BRE for $\eta \in (0, 1/2)$, but for the $\eta = 1/2$ case we also require that $L(\gamma) \not\to \infty$. □

Heffernan [96] has conveniently compiled a directory of $\eta$ and $L(x)$ for many copulas which satisfy (6.30). A summary of these results is given in Table 6.1. In reading Heffernan's directory, one can spot two trends: normally $\eta \in \{1/2, 1\}$ and $L$ is a constant. The oft-cited Gaussian copula is the only exception for both of these trends in Heffernan's directory, having $\eta = (1+\rho)/2$ and $L(x) \propto (\log x)^{-\rho/(1+\rho)}$; Section 6.4.4 deals with the Gaussian case in detail.

### Archimedean Copulas

Some of the most frequently used copulas are in the family of *Archimedean copulas*. These are very general models and are widely used in applications due to their flexibility. A copula is Archimedean if there exists a function $\psi$ such that the copula $C$ can be written as

$$C(u_1, \ldots, u_n) = \psi^{\leftarrow}(\psi(u_1) + \cdots + \psi(u_n)),$$

where the function $\psi$ is called the *generator* of the copula. If $\psi^{\leftarrow}$ is the Laplace transform of a non-negative random variable, then $\psi$ defines a generator for a valid Archmidean copula, however other constructions of Archimedean copulas are possible [132, p. 74]. For Archimedean copulas we can restate the BRE condition (6.27) in terms of the generator $\psi$.

**Theorem 6.11** (Thm. 3.4 of [44]). *Let $(U_1, \ldots, U_n) \sim C$ where $C$ is an Archimedean copula with generator $\psi$. If $\psi^{\leftarrow}$ is twice continuously differentiable and its second derivative is bounded at 0 then $\forall\, i \neq j$*

$$\lim_{u \to 0} \frac{\mathbb{P}(U_i \geq 1 - ux_1, U_j \geq 1 - ux_2)}{u^2} < \infty$$

*for any $0 < x_1,\, x_2 < \infty$.*

**Corollary 6.12.** *Consider using $\widehat{\alpha}_1$ for a distribution with common marginal distributions and a copula $C$. If $C$ satisfies the conditions of Theorem 6.11 then $\widehat{\alpha}_1$ has BRE.*

Charpentier and Segers [44] have helpfully created a directory of Archimedean copulas from which we can see if the BRE conditions from Corollary 6.12 are satisfied. Using this information, we provide a summary of the efficiency status of many Archimedean copulas in Table 6.2.

The efficiency of $\widehat{\alpha}_1$ can be proved without the assumption of identical marginal distributions, but the efficiency must be shown case-by-case for each family of distributions. The next section does this for the multivariate normal distribution and for some elliptical distributions.

### 6.4.4 Efficiency for the case of normal and elliptical distributions

The efficiency characteristics of normally and elliptically distributed random vectors are very similar. This section defines these distributions, outlines their asymptotic properties, then shows the conditions in which $\widehat{\alpha}_1$ exhibits levels of asymptotic efficiency.

Table 6.2: Examples of Archimedean copula families. Names (if they are named) and generator functions are listed, as are the ranges for which $\theta$ is valid and the subset of $\theta$ which ensures that $\widehat{\alpha}_1$ has BRE. A $\Theta$ in the final column means that all valid $\theta$ ensure BRE. The families listed appear in Table 4.1 of [132] and Table 1 of [44].

| # | **Name** | **Generator** $\psi(t)$ | **Valid** $\theta$ | **Efficient** $\theta$ |
|---|---|---|---|---|
| 1 | Clayton | $\frac{1}{\theta}(t^{-\theta}-1)$ | $[-1,\infty)$ | $\Theta$ |
| 2 | | $(1-t)^{\theta}$ | $[1,\infty)$ | $\{1\}$ |
| 3 | Ali–Mikhail–Haq | $\log\frac{1-\theta(1-t)}{t}$ | $[-1,1)$ | $\Theta$ |
| 4 | Gumbel–Hougaard | $(-\log t)^{\theta}$ | $[1,\infty)$ | $\{1\}$ |
| 5 | Frank | $-\log\frac{\mathrm{e}^{-\theta t}-1}{\mathrm{e}^{-\theta}-1}$ | $\mathbb{R}$ | $\Theta\setminus\{0\}$ |
| 6 | | $-\log[1-(1-t)^{\theta}]$ | $[1,\infty)$ | $\{1\}$ |
| 7 | | $-\log[\theta t+(1-\theta)]$ | $(0,1]$ | $\Theta$ |
| 8 | | $\frac{1-t}{1+(\theta-1)t}$ | $[1,\infty)$ | $\Theta$ |
| 9 | | $\log(1-\theta\log t)$ | $(0,1]$ | $\Theta$ |
| 10 | | $\log(2t^{-\theta}-1)$ | $(0,1]$ | $\Theta$ |
| 11 | | $\log(2-t^{\theta})$ | $(0,1/2]$ | $\Theta$ |
| 12 | | $(\frac{1}{t}-1)^{\theta}$ | $[1,\infty)$ | $\{1\}$ |
| 13 | | $(1-\log t)^{\theta}-1$ | $(0,\infty)$ | $\Theta$ |
| 14 | | $(t^{-1/\theta}-1)^{\theta}$ | $[1,\infty)$ | $\{1\}$ |
| 15 | | $(1-t^{1/\theta})^{\theta}$ | $[1,\infty)$ | $\{1\}$ |
| 16 | | $(\frac{\theta}{t}+1)(1-t)$ | $[0,\infty)$ | $\Theta$ |
| 17 | | $-\log\frac{(1+t)^{-\theta}-1}{2^{-\theta}-1}$ | $\mathbb{R}$ | $\Theta\setminus\{0\}$ |
| 18 | | $\mathrm{e}^{\theta/(t-1)}$ | $[2,\infty)$ | $\emptyset$ |
| 19 | | $\mathrm{e}^{\theta/t}-\mathrm{e}^{\theta}$ | $(0,\infty)$ | $\Theta$ |
| 20 | | $\mathrm{e}^{t^{-\theta}}-\mathrm{e}$ | $(0,\infty)$ | $\Theta$ |
| 21 | | $1-[1-(1-t)^{\theta}]^{1/\theta}$ | $[1,\infty)$ | $\{1\}$ |
| 22 | | $\arcsin(1-t^{\theta})$ | $(0,1]$ | $\Theta$ |

### Definitions and categories of elliptical distributions

Let $\mathsf{Normal}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. Denote the corresponding density $\phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\cdot, \cdot)$, and write $\sigma_i^2 := \boldsymbol{\Sigma}_{ii}$, $\rho_{ij} := \boldsymbol{\Sigma}_{ij}/(\sigma_i \sigma_j)$. The normal distribution belong to the class of *elliptical distributions*, which we denote $\mathsf{Elliptical}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F)$, where $F$ is the cdf of a positive random variable. We define $\boldsymbol{X} \sim \mathsf{Elliptical}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F)$ as

$$\boldsymbol{X} \stackrel{\mathcal{D}}{=} \boldsymbol{\mu} + R\,\boldsymbol{C}\,U \tag{6.31}$$

where $R \sim F$ is called the *radial component*, $U$ is (independent of $R$ and) distributed uniformly on the $d$-dimensional unit hypersphere, and $\boldsymbol{C} \in \mathbb{R}^{d \times d}$ satisfies $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Sigma}$. For background on elliptical distributions, see [11]. The efficiency of $\widehat{\alpha}_1$ turns out to be related with max-domain of attraction (MDA) of the radial component. The MDA is known from standard extreme value theory, see [58].

We consider some subclasses of elliptical distributions depending on the MDA of the radial distribution:

- $F \in \mathsf{MDA}(\mathsf{Fréchet})$, then Theorem 4.3 of [100] implies that $\boldsymbol{X}$ has asymptotic dependence and $\widehat{\alpha}_1$ is never efficient (see Section 6.4.3).

- $F \in \mathsf{MDA}(\mathsf{Weibull})$, then components of $\boldsymbol{X}$ are light-tailed and uninteresting (in a rare-event context).

- $F \in \mathsf{MDA}(\mathsf{Gumbel})$, this is the interesting case which includes the normal distribution. Hashorva [90] label these the *type I elliptical random vectors*.

### Efficiency for type I elliptical distributions

Take $\boldsymbol{X} \sim \mathsf{Elliptical}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F)$ where the radial distribution $F \in \mathsf{MDA}(\mathsf{Gumbel})$ has support $(0, x_F)$, for some $x_F \in \overline{\mathbb{R}}$, and where $\{\sigma_1, \ldots, \sigma_d\}$ are in decreasing order. By definition

of the Gumbel MDA, one can find a scaling function $w(x)$ satisfying

$$\lim_{x\to x_F} \frac{\overline{F}(x + t/w(x))}{\overline{F}(x)} = \mathrm{e}^{-t}.$$

One frequently takes $w(x) := \overline{F}(x)/\int_x^{x_F} \overline{F}(s)\,\mathrm{d}s$. Also, define $\upsilon_i(\gamma) := (\gamma - \mu_i)/\sigma_i$ and $a_{ij} := \sigma_j/\sigma_i$. If $\rho_{ij} \geq a_{ij}$ then set

$$\mu_{ij} := \mu_j \quad \text{and} \quad \kappa_{ij} := \sigma_j$$

otherwise for $\rho_{ij} < a_{ij}$

$$\mu_{ij} := \frac{\mu_i - a_{ij}\rho_{ij}(\mu_1 + \mu_2) + a^2\mu_j}{\alpha_{ij}(1 - \rho_{ij}^2)} \quad \text{and} \quad \kappa_{ij} := \frac{\sigma_i^2\sigma_j^2(1 - \rho_{ij}^2)}{\sigma_i^2 - 2\rho_{ij}\sigma_i\sigma_j + \sigma_j^2}\,.$$

We now apply the asymptotic properties outlined in the Appendix to assess the efficiency of $\widehat{\alpha}_1$ for type I elliptical distributions.

**Theorem 6.13.** *Consider* $\boldsymbol{X} \sim \mathsf{Elliptical}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F)$ *where* $F \in \mathsf{MDA}(\mathsf{Gumbel})$*, and let*

$$\kappa := \max_{i<j} \kappa_{ij}\,, \quad \mu := \max_{i<j\,:\,\kappa=\kappa_{ij}} \mu_{ij}\,, \quad \textit{and} \quad \upsilon(\gamma) := (\gamma - \mu)/\kappa + \mathrm{o}(1)\,.$$

*If* $\kappa > \sigma_1$,[1] *then* $\widehat{\alpha}_1$ *has LE if*

$$\forall \varepsilon > 0 \quad \limsup_{\gamma\to x_F} \frac{w(\upsilon(\gamma))\overline{F}(\upsilon(\gamma))}{w(\upsilon_1(\gamma))\overline{F}(\upsilon_1(\gamma))^{2-\varepsilon}} < \infty\,. \tag{6.32}$$

*Moreover, if* (6.32) *holds for* $\varepsilon = 0$ *then* $\widehat{\alpha}_1$ *has BRE.*

*Proof.* It follows from (6.35) and Theorem 6.17 in the Appendix. □

**Example 6.14** (Kotz Type III)**.** One family of type I elliptical distributions, is the *Kotz Type III* distributions, defined by

$$\overline{F}(\gamma) = (K + \mathrm{o}(1))\gamma^N \exp(-ru^\delta), \quad w(\gamma) = r\delta\gamma^{\delta-1}\,, \quad \text{for } \gamma > 0,$$

[1] This implies that the Savage condition (see Appendix) is fulfilled at least for one pair.

with $K, \delta, N > 0$. In this case it is clear that

$$\lim_{\gamma\to\infty} \frac{w(\upsilon(\gamma))}{w(\upsilon_1(\gamma))} = \left(\frac{\sigma_1}{\kappa}\right)^{\delta-1} < \infty,$$

while

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{\overline{F}(\upsilon(\gamma))}{\overline{F}(\upsilon_1(\gamma))^2} \\
&= \limsup_{\gamma\to\infty} \left(\frac{\sigma_1^2}{\kappa\gamma}\right)^N \exp\left\{-r\left(\left(\frac{\gamma-\mu}{\kappa}\right)^\delta - 2\left(\frac{\gamma-\mu_1}{\sigma_1}\right)^\delta\right)\right\}, \\
&= \limsup_{\gamma\to\infty} \left(\frac{\sigma_1^2}{\kappa\gamma}\right)^N \exp\left\{-r\left(\frac{\gamma^\delta - \delta\mu\gamma^{\delta-1} + \mathrm{o}(\gamma^{\delta-1})}{\kappa^\delta} - \frac{\gamma^\delta - \delta\mu_1\gamma^{\delta-1} + \mathrm{o}(\gamma^{\delta-1})}{\sigma_1^\delta/2}\right)\right\}.
\end{aligned}$$

Hence, $\widehat{\alpha}_1$ has BRE in the following cases

- $\sigma_1^\delta > 2\kappa^\delta$, or
- $\sigma_1^\delta = 2\kappa^\delta$, $\delta > 1$ and $\mu_1 > \mu$.

The estimator $\widehat{\alpha}_1$ has LE if $\sigma_1^\delta = 2\kappa^\delta$, and is inefficient when $\sigma_1^\delta < 2\kappa^\delta$.

**Example 6.15** (Normal distributions)**.** The normal distribution is a Kotz III type distribution with $\delta = 2$. Hence, $\widehat{\alpha}_1$ has BRE if $\sigma_1^2 > 2\kappa^2$, or $\sigma_1^2 = 2\kappa^2$ and $\mu_1 > \mu$. The estimator $\widehat{\alpha}_1$ has LE if $\sigma_1^2 = 2\kappa^2$, and is inefficient when $\sigma_1^2 < 2\kappa^2$.

Frequently, a set of random variables represents as a stochastic process $\{X_n\}_{n\geq 0}$. The value of $\mathbb{P}(M > \gamma)$, with $M := \max_{1\leq n\leq d} X_n$, in such cases usually valuable. The simplest case to take is when all $X_n$ have identical marginals such as in stationary processes; one such example is the autoregressive (AR) process.

**Example 6.16** (AR(1) processes)**.** Say $X_t = \varphi X_{t-1} + \varepsilon_t$, where $|\varphi| < 1$ and $\varepsilon_t$ are iid $\mathsf{Normal}_1(0, \sigma_\varepsilon^2)$, and we start the process in stationarity. We have that each $X_i$ has the

same marginal distribution, $X_i \sim \mathsf{Normal}_1(0, \sigma_\varepsilon^2/(1-\varphi^2))$, and

$$\max_{i<j} \mathbb{P}(X_i > \gamma, X_j > \gamma) = \begin{cases} \mathbb{P}(X_\bullet > \gamma, X_{\bullet+1} > \gamma) & \text{if } \varphi > 0 \\ \mathbb{P}(X_\bullet > \gamma, X_{\bullet+2} > \gamma) & \text{if } \varphi < 0 \\ \mathbb{P}(X_\bullet > \gamma)^2 & \text{if } \varphi = 0 \end{cases} .$$

For $\varphi \neq 0$ we know that

$$(X_{\bullet+1} \mid X_\bullet = \gamma) \sim \mathsf{Normal}_1(\varphi\gamma, \sigma_\varepsilon^2), \text{ and } (X_{\bullet+2} \mid X_\bullet = \gamma) \sim \mathsf{Normal}_1(\varphi^2\gamma, \sigma_\varepsilon^2(1-\varphi^4)/(1-\varphi^2)) .$$

When $\varphi = 0$ the $X_i$ are independent and $\widehat{\alpha}_1$ is trivially efficient, and when $\varphi \in (-1,1)\backslash\{0\}$ we have (noting that $\{X_\bullet > \gamma\} \to \{X_\bullet = \gamma\}$) that

$$\begin{aligned} \lim_{\gamma\to\infty} \frac{\max_{i<j} \mathbb{P}(X_i > \gamma, X_j > \gamma)}{\max_i \mathbb{P}(X_i > \gamma)^2} &= \lim_{\gamma\to\infty} \frac{\mathbb{P}(X_\bullet > \gamma, X_{\bullet+(1\text{ or }2)} > \gamma)}{\mathbb{P}(X_\bullet > \gamma)^2} \\ &= \lim_{\gamma\to\infty} \frac{\mathbb{P}(X_{\bullet+(1\text{ or }2)} > \gamma \mid X_\bullet = \gamma)}{\mathbb{P}(X_\bullet > \gamma)} \\ &= 0 \end{aligned}$$

as $\sigma_\varepsilon^2 < \sigma_\varepsilon^2(1-\varphi^4)/(1-\varphi^2)) < \sigma_\varepsilon^2/(1-\varphi^2)$. Therefore, we have BRE of $\widehat{\alpha}_1$ for all stationary AR(1) processes.

## 6.5 Numerical experiments

We explore the performance of the estimators for the problem of $\mathbb{P}(M > \gamma)$ for $M = \max_i X_i$, where $\boldsymbol{X}$ is multivariate normal and multivariate Laplace distributed. The following notation is used: $\boldsymbol{X}_{-i}$ ($\boldsymbol{X}_{-i,-j}$) is the random vector $\boldsymbol{X}$ with $X_i$ ($X_i$ and $X_j$) removed, $\boldsymbol{0}$ is the vector of zeros, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{x}^\top$ is the transpose of $\boldsymbol{x}$, and $X \perp Y$ means $X$ and $Y$ are independent. We use some standard distributions: $\mathsf{Exponential}(\lambda)$ for exponential ($f(x) \propto \mathrm{e}^{-\lambda x}$), $\mathsf{InverseGaussian}(\mu, \lambda)$ for inverse Gaussian ($f(x) \propto x^{-3/2}\mathrm{e}^{-\lambda(x-\mu)^2/(2\mu^2 x)}$), $\mathsf{Laplace}()$ for Laplace (defined in Case 2 below). The MATLAB and MATHEMATICA code used to generate them are available online [10].

## Case 1: Multivariate Normal distributions

Let $\boldsymbol{X} \sim \mathsf{Normal}_d(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I} + \rho$; that is, each $X_i \sim \mathsf{Normal}_1(0,1)$ and $\mathbb{C}\mathrm{orr}(X_i, X_j) = \rho$. We implement the first- and second-order IS regimes. The necessary conditional distributions are well-known and simple; both $\boldsymbol{X}_{-i} \mid X_i$ and $\boldsymbol{X}_{-i,-j} \mid (X_i, X_j)$ are normally distributed [11]. Sampling from $X_i \mid X_i > \gamma$ can be easily done by acceptance–rejection with shifted exponential proposals [145] (or by inverse transform sampling [15, Remark 2.4], though this can be problematic using only double precision arithmetic). To simulate $(X_i, X_j) \mid \min(X_i, X_j) > \gamma$ we use Botev's MATLAB library [35], but also remark that a Gibb's sampler is a commonly used alternative [38, 145].

## Case 2: Multivariate Laplace distributions

Let $\boldsymbol{X} \sim \mathsf{Laplace}()$. We can define this distribution by

$$\boldsymbol{X} \overset{\mathcal{D}}{=} \sqrt{R}\boldsymbol{Y}\,, \quad \text{where } \boldsymbol{Y} \sim \mathsf{Normal}_d(\mathbf{0}, \boldsymbol{I}), R \sim \mathsf{Exponential}(1), \boldsymbol{Y} \perp R\,.$$

The distribution has been applied in a financial context [99], and is examined in [68, 109]. From the former we have that the density of $\mathsf{Laplace}()$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = 2(2\pi)^{-d/2} K_{(d/2)-1}\Big(\sqrt{2\boldsymbol{x}^\top \boldsymbol{x}}\Big)\left(\sqrt{\tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{x}}\right)^{1-(d/2)}$$

where $K_n$ denotes the modified Bessel function of the second kind of order $n$.

To implement the first-order IS algorithm we need the conditional distributions $X_i \mid X_i > \gamma$ and $\boldsymbol{X}_{-i} \mid \boldsymbol{X}_i$. Assuming $\gamma > 0$ we can derive that $(X_i \mid X_i > \gamma) \sim \mathsf{Exponential}(\sqrt{2})$. Further calculation gives

$$\boldsymbol{X}_{-1} \mid X_1 \;\overset{\mathcal{D}}{=}\; \frac{X_1}{Y_1}\boldsymbol{Y}_{-1} \mid (\sqrt{R}Y_1 = X_1) \;\overset{\mathcal{D}}{=}\; \frac{X_1}{Y_{1,X_1}}\boldsymbol{Y}_{-1}\,,$$

where $Y_{1,X_1} \sim (Y_1 \mid \sqrt{R}Y_1 = X_1)$, noting that $Y_{1,X_1} \perp \boldsymbol{Y}_{-1}$ because of the independence between the entries of $\boldsymbol{Y}$. Direct calculation gives

$$f_{Y_i \mid \sqrt{R}Y_i}(y_i \mid x_i) = 2|y_i| \exp\Big(-x_i^2/y_i^2 - x_i^2/2 + \sqrt{2}|x_i|\Big)/(\sqrt{\pi}y_i^2)$$

which is the density of $\sqrt{X}$ where $X \sim \mathsf{InverseGaussian}(\sqrt{2}|x_i|, 2x_i^2)$. This is summarised in the following algorithm.

**Algorithm 5** Sampling $\boldsymbol{X}_{-i} \mid X_i > \gamma$ for the Laplace distribution

1: $X_i \leftarrow \mathsf{Exponential}(\sqrt{2})$
2: $Y_{i,X_i} \leftarrow \mathsf{InverseGaussian}(\sqrt{2}|X_i|, 2X_i^2)$.
3: $\boldsymbol{Y}_{-i} \leftarrow \mathsf{Normal}_{d-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$.
4: **return** $X_i \boldsymbol{Y}_{-i} / Y_{i,X_i}$.

### 6.5.1 Test setup

The estimators tested are $\widehat{\alpha}_0$ (crude Monte Carlo) and $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_1^{[1]}$, $\widehat{\alpha}_2^{[2]}$, $(\widehat{\beta_1 \ddagger \alpha})$, $(\widehat{\beta_2 \ddagger \alpha})$, defined in (6.1), (6.2), (6.14), (6.15), (6.20) and (6.22) respectively. As a reference, we show the true value $\alpha$ (calculated by numerical integration using MATHEMATICA), and the first two truncations of the IEF: $\overline{\alpha}(\gamma) := \sum_i \mathbb{P}(X_i > \gamma)$ and $\overline{\alpha}(\gamma) - q(\gamma) := \sum_i \mathbb{P}(X_i > \gamma) - \sum_{i<j} \mathbb{P}(X_i > \gamma, X_j > \gamma)$. Each estimator is given $R = 10^6$, and an asterisk is placed in table entries where the corresponding estimate had 0 variance (i.e., the estimator had degenerated).

### 6.5.2 Results

| Estimators | $\gamma$ | | | |
|---|---|---|---|---|
| | 2 | 4 | 6 | 8 |
| $\alpha$ | 5.633e-02 | 1.095e-04 | 3.838e-09 | 2.481e-15 |
| $\widehat{\alpha}_0$ | 5.651e-02 | 1.140e-04 | 0* | 0* |
| $\overline{\alpha}$ | 9.100e-02 | 1.267e-04 | 3.946e-09 | 2.488e-15 |
| $\overline{\alpha}{-}q$ | 4.000e-02 | 1.055e-04 | 3.827e-09 | 2.480e-15 |
| $\widehat{\alpha}_1$ | 5.650e-02 | 1.047e-04 | 3.946e-09* | 2.488e-15* |
| $\widehat{\alpha}_2$ | 5.605e-02 | 1.075e-04 | 3.827e-09* | 2.480e-15* |
| $\widehat{\alpha}_1^{[1]}$ | 5.637e-02 | 1.096e-04 | 3.837e-09 | 2.481e-15 |
| $\widehat{\alpha}_2^{[2]}$ | 5.633e-02 | 1.095e-04 | 3.838e-09 | 2.481e-15 |
| $\widehat{(\beta_1 \ddagger \alpha)}$ | 5.634e-02 | 1.095e-04 | 3.838e-09 | 2.480e-15 |
| $\widehat{(\beta_2 \ddagger \alpha)}$ | 5.631e-02 | 1.095e-04 | 3.838e-09 | 2.481e-15 |

Table 6.3: Estimates of $\mathbb{P}(M > \gamma)$ where $M = \max_i X_i$ and $\boldsymbol{X} \sim \mathsf{Normal}_4(\mathbf{0}_4, \boldsymbol{\Sigma})$, $\rho = 0.75$.

| Estimators | $\gamma$ | | | |
|---|---|---|---|---|
| | 2 | 4 | 6 | 8 |
| $\widehat{\alpha}_0$ | 3.109e-03 | 4.075e-02 | 1* | 1* |
| $\overline{\alpha}$ | 6.154e-01 | 1.566e-01 | 2.822e-02 | 3.142e-03 |
| $\overline{\alpha}{-}q$ | 2.899e-01 | 3.665e-02 | 2.827e-03 | 1.147e-04 |
| $\widehat{\alpha}_1$ | 2.977e-03 | 4.429e-02 | 2.822e-02* | 3.142e-03* |
| $\widehat{\alpha}_2$ | 5.077e-03 | 1.839e-02 | 2.827e-03* | 1.147e-04* |
| $\widehat{\alpha}_1^{[1]}$ | 6.918e-04 | 4.639e-04 | 1.747e-04 | 2.192e-05 |
| $\widehat{\alpha}_2^{[2]}$ | 7.838e-08 | 8.647e-05 | 1.237e-05 | 4.010e-08 |
| $\widehat{(\beta_1 \ddagger \alpha)}$ | 6.564e-05 | 7.046e-05 | 6.227e-05 | 4.362e-05 |
| $\widehat{(\beta_2 \ddagger \alpha)}$ | 3.493e-04 | 1.593e-05 | 6.883e-06 | 3.340e-07 |

Table 6.4: Absolute relative errors of the estimates in Table 6.3.

| Estimators | $\gamma$ | | | |
|---|---|---|---|---|
| | 2 | 4 | 6 | 8 |
| $\widehat{\alpha}_0$ | 2.309e-01 | 1.068e-02 | 0 | 0 |
| $\widehat{\alpha}_1$ | 2.557e-01 | 5.099e-03 | 0 | 0 |
| $\widehat{\alpha}_2$ | 1.885e-01 | 1.414e-03 | 0 | 0 |
| $\widehat{\alpha}_1^{[1]}$ | 2.817e-02 | 3.071e-05 | 4.650e-10 | 9.972e-17 |
| $\widehat{\alpha}_2^{[2]}$ | 9.901e-03 | 4.244e-06 | 1.908e-11 | 8.575e-19 |
| $\widehat{(\beta_1 \ddagger \alpha)}$ | 1.929e-02 | 2.089e-05 | 3.197e-10 | 6.994e-17 |
| $\widehat{(\beta_2 \ddagger \alpha)}$ | 1.306e-02 | 5.265e-06 | 2.310e-11 | 1.035e-18 |

Table 6.5: Standard deviations of the estimates in Table 6.3.

| Estimators | $\gamma$ | | | |
|---|---|---|---|---|
| | 6 | 8 | 10 | 12 |
| $\alpha$ | 4.093e-04 | 2.435e-05 | 1.442e-06 | 8.526e-08 |
| $\widehat{\alpha}_0$ | 3.910e-04 | 2.000e-05 | 2.000e-06 | 0* |
| $\overline{\alpha}$ | 4.130e-04 | 2.441e-05 | 1.443e-06 | 8.527e-08 |
| $\overline{\alpha}-q$ | 4.093e-04 | 2.435e-05 | 1.442e-06 | 8.526e-08 |
| $\widehat{\alpha}_1$ | 4.120e-04 | 2.441e-05* | 1.443e-06* | 8.527e-08* |
| $\widehat{\alpha}_2$ | 4.093e-04* | 2.435e-05* | 1.442e-06* | 8.526e-08* |
| $\widehat{\alpha}_1^{[1]}$ | 4.093e-04 | 2.435e-05 | 1.442e-06 | 8.526e-08 |
| $\widehat{(\beta_1 \ddagger \alpha)}$ | 4.093e-04 | 2.435e-05 | 1.442e-06 | 8.526e-08 |

Table 6.6: Estimates of $\mathbb{P}(M > \gamma)$ where $M = \max_i X_i$ and $\boldsymbol{X} \sim \mathsf{Laplace}()$, $d = 4$.

| Estimators | $\gamma$ | | | |
|---|---|---|---|---|
| | 6 | 8 | 10 | 12 |
| $\widehat{\alpha}_0$ | 4.472e-02 | 1.786e-01 | 3.873e-01 | 1* |
| $\overline{\alpha}$ | 8.959e-03 | 2.473e-03 | 6.987e-04 | 2.003e-04 |
| $\overline{\alpha}-q$ | 8.067e-05 | 8.266e-06 | 8.757e-07 | 9.506e-08 |
| $\widehat{\alpha}_1$ | 6.516e-03 | 2.473e-03* | 6.987e-04* | 2.003e-04* |
| $\widehat{\alpha}_2$ | 8.067e-05* | 8.266e-06* | 8.757e-07* | 9.506e-08* |
| $\widehat{\alpha}_1^{[1]}$ | 8.470e-06 | 1.023e-05 | 3.019e-05 | 1.577e-05 |
| $\widehat{(\beta_1 \ddagger \alpha)}$ | 4.515e-05 | 2.948e-05 | 2.151e-06 | 2.833e-06 |

Table 6.7: Absolute relative errors of the estimates in Table 6.6.

| Estimators | $\gamma$ | | | |
|---|---|---|---|---|
| | 6 | 8 | 10 | 12 |
| $\widehat{\alpha}_0$ | 1.977e-02 | 4.472e-03 | 1.414e-03 | 0 |
| $\widehat{\alpha}_1$ | 1.000e-03 | 0 | 0 | 0 |
| $\widehat{\alpha}_2$ | 0 | 0 | 0 | 0 |
| $\widehat{\alpha}_1^{[1]}$ | 2.735e-05 | 8.581e-07 | 2.752e-08 | 8.189e-10 |
| $\widehat{(\beta_1 \ddagger \alpha)}$ | 1.937e-05 | 6.086e-07 | 1.908e-08 | 5.990e-10 |

Table 6.8: Standard deviations of the estimates in Table 6.6.

### 6.5.3 Discussion

We begin with some trends which we expected to find in the results:

- all estimators outperform crude Monte Carlo $\widehat{\alpha}_0$,
- the estimators which calculate $\mathbb{P}(X_i > \gamma)$ outperform those which do not,
- the estimators which calculate $\mathbb{P}(X_i > \gamma, X_j > \gamma)$ outperform those which only use the univariate $\mathbb{P}(X_i > \gamma)$,
- the importance sampling estimators improve upon their original counterparts,
- the second-order IS improves upon the first-order IS.

Also noticed in the performance of the $\widehat{\alpha}$ estimators:

- the $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ estimators often degenerated (i.e. had zero variance) to $\overline{\alpha}$ and $\overline{\alpha}-q$ respectively,
- the degeneration begin for smaller $\gamma$ when the $\boldsymbol{X}$ had a weaker dependence structure.

Table 6.9 shows the degeneration of the estimators in various examples involving multivariate normal distributions.

The fact that the estimators degenerate is not wholly undesirable, as they degenerate to the deterministic functions $\overline{\alpha}$ and $\overline{\alpha}-q$ which are highly accurate when degeneration occurs. Obviously, for very large $\gamma$ one would not resort to Monte Carlo methods as the asymptote $\overline{\alpha}$ would be accurate enough for most purposes; one could use the $\widehat{\alpha}$ estimators until the sample variance is below some threshold, then switch to the faster deterministic estimators $\overline{\alpha}$ and $\overline{\alpha}-q$.

Regarding the $\widehat{(\beta_1 \ddagger \alpha)}$ and $\widehat{(\beta_2 \ddagger \alpha)}$ estimators:

- their performance is roughly the same as than their $\widehat{\alpha}_1^{[1]}$ and $\widehat{\alpha}_2^{[2]}$ counterparts,

| Test cases | | $\gamma$ | | | |
|---|---|---|---|---|---|
| $d$ | $\rho$ | 2 | 4 | 6 | 8 |
| 3 | -0.25 | 0.00957 | 1* | 1* | 1* |
| | 0 | 0.00255 | 1* | 1* | 1* |
| | 0.5 | 0.00166 | 1* | 1* | 1* |
| | 0.75 | 0.005 | 0.165 | 1* | 1* |
| 4 | -0.25 | 0.00955 | 1* | 1* | 1* |
| | 0 | 0.0185 | 1* | 1* | 1* |
| | 0.5 | 0.00139 | 1* | 1* | 1* |
| | 0.75 | 0.00484 | 0.283 | 1* | 1* |
| Average | | 0.00663 | 0.806 | 1 | 1 |

(a) $\widehat{\alpha}_1$ to $\overline{\alpha}$

| Test cases | | $\gamma$ | | | |
|---|---|---|---|---|---|
| $d$ | $\rho$ | 2 | 4 | 6 | 8 |
| 3 | -0.25 | 1* | 1* | 1* | 1* |
| | 0 | 0.151* | 1* | 1* | 1* |
| | 0.5 | 0.0764 | 1* | 1* | 1* |
| | 0.75 | 0.0172 | 0.754 | 1* | 1* |
| 4 | -0.25 | 1* | 1* | 1* | 1* |
| | 0 | 0.189 | 1* | 1* | 1* |
| | 0.5 | 0.0153 | 1* | 1* | 1* |
| | 0.75 | 0.0175 | 0.502 | 1* | 1* |
| Average | | 0.308 | 0.907 | 1 | 1 |

(b) $\widehat{\alpha}_2$ to $\overline{\alpha}-q$

Table 6.9: Ratios of absolute relative errors for pairs of estimators. Each row corresponds to a separate distribution for $\boldsymbol{X}$, each being $\mathsf{Normal}_d$ distributed with standard normal marginals and constant correlation $\rho$.

- they perform better when the dependence between the variables is weak.

One must remember that the $\widehat{\beta}_i$ estimators are valid for a much larger class of problems (estimating expectations, not just probabilities). Also, we would expect that the $\widehat{\beta}_i$-based estimators compare favorably to the $\widehat{\alpha}_i^{[i]}$ IS-based estimators when $d$ is large, as the method involves no likelihood term which can degenerate.

## 6.6 Conclusion

We presented new estimators for the tail probability of a union of dependent rare events. The key idea in both estimators is that the tail probability of the such a rare event can be well approximated by the Bonferroni approximations:

$$\alpha = \mathbb{P}(A) \approx \sum_{i=1}^{k} (-1)^{i-1} \sum_{|I|=i} \mathbb{P}\Big(\bigcap_{i \in I} A_i\Big) \text{ for } k = 1, 2\,.$$

We provided conditions which ensure $\widehat{\alpha}_1$ and $\widehat{\beta}_i$ have logarithmic efficiency and bounded relative error. The estimators were tested on the classical example of rare maxima of random vectors. Furthermore, the fact that our $\widehat{\beta}_i$ estimators can be applied to a more general setting makes them useful for a larger variety of estimation problems.

### 6.6.1 Future work

We did not discuss stratification strategies for $\widehat{\beta}_i$ that could result in further reductions in variance. Nor did we investigate which permutations of the $A_i$ minimise the variance of $\widehat{\beta}_i$. Further investigation into the use of $\widehat{\beta}_i$ to estimate tail probabilities of order statistics would be of value.

## 6.A Elliptical distribution asymptotics

### 6.A.1 Asymptotic properties of normal distributions

In general, for an $\boldsymbol{X} \sim \mathsf{Normal}_d(\mathbf{0}, \boldsymbol{\Sigma})$, Theorem 2.6.1 of Bryc [39] states that for all measurable $A \subset \mathbb{R}^d$ the

$$\lim_{n\to\infty} \frac{1}{n^2} \log \mathbb{P}(\boldsymbol{X} \geq nA) = -\inf_{\boldsymbol{x}\in A} \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}. \tag{6.33}$$

The asymptotic properties of elliptical distributions also relate to this quadratic programming problem, which Hashorva [89, 90] denotes as

$$\mathcal{P}(\boldsymbol{\Sigma}^{-1}, \boldsymbol{t}) := \text{minimise } \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \text{ under the linear constraint } \boldsymbol{x} \geq \boldsymbol{t}. \tag{6.34}$$

The program $\mathcal{P}(\boldsymbol{\Sigma}^{-1}, \boldsymbol{t})$ is usually minimised at the boundary $\boldsymbol{t}$, and hence the asymptotic form (6.33) is very simple. This occurs when $\boldsymbol{\Sigma}^{-1}\boldsymbol{t} > \mathbf{0}$ (componentwise), a condition often called the *Savage condition* after Richard Savage [152]. For the cases when the Savage condition fails, the asymptotics change as some components of $\boldsymbol{X}$ become irrelevant in the limit. Figure 6.1 graphically shows some contours of $\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$ for some $\boldsymbol{\Sigma}$ which do and do not satisfy the Savage condition.

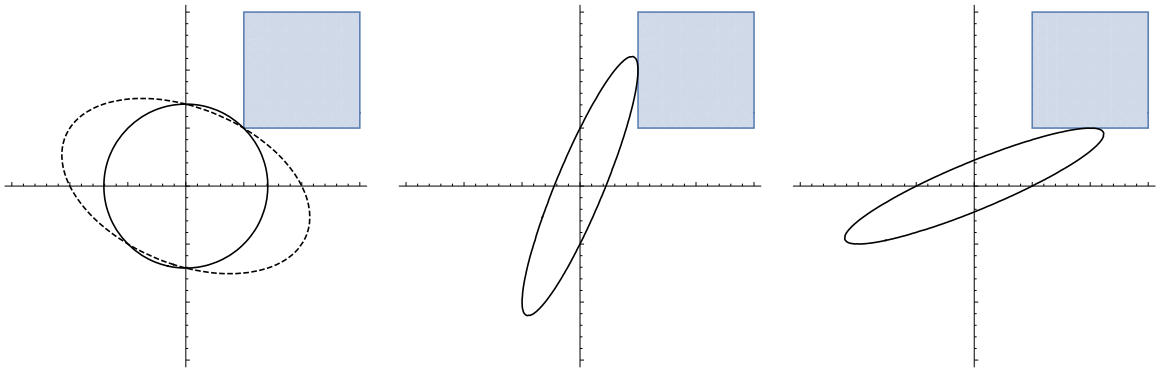


Figure 6.1: Contours of $\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$ for example $\boldsymbol{\Sigma}$ which: (a) satisfy the Savage condition (i.e., $\boldsymbol{\Sigma}^{-1}\mathbf{1} > \mathbf{0}$), and (b)–(c) do not satisfy the condition. The covariance matrices, in MATLAB notation, are: (a) $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $\boldsymbol{\Sigma} = [2, -1/2; -1/2, 1]$, (b) $\boldsymbol{\Sigma} = [1, 2; 2, 5]$, and (c) $\boldsymbol{\Sigma} = [5, 2; 2, 1]$.

### 6.A.2 Asymptotic properties of type I elliptical distributions

Take $\boldsymbol{X} \sim \mathsf{Elliptical}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F)$ where the radial distribution $F \in \mathsf{MDA}(\mathsf{Gumbel})$ has support $(0, x_F)$, for some $x_F \in \overline{\mathbb{R}}$, and where $\{\sigma_1, \ldots, \sigma_d\}$ are in decreasing order. The univariate and bivariate asymptotics, $\mathbb{P}(X_i > \gamma)$ and $\mathbb{P}(X_i > \gamma, X_j > \gamma)$, can be written in terms of the scaling function $w(\gamma)$ and of $\overline{F}((\gamma - \mu)/\kappa)$ for some particular $\mu$ and $\kappa$. Theorem 12.3.1 of Berman [32] gives the univariate case,

$$\mathbb{P}(X_i > \gamma) = (1 + \mathrm{o}(1)) \frac{\overline{F}(\upsilon_i(\gamma))}{\sqrt{2\pi \upsilon_i(\gamma) w(\upsilon_i(\gamma))}} \quad \text{as } \gamma \to x_F \tag{6.35}$$

where $\upsilon_i(\gamma) = (\gamma - \mu_i)/\sigma_i$. The bivariate case, i.e. $\mathbb{P}(X_i > \gamma, X_j > \gamma)$, relies on the following constants. Define $a_{ij} := \sigma_j/\sigma_i$. If $\rho_{ij} \geq a_{ij}$ then define

$$\mu_{ij} := \mu_j \quad \text{and} \quad \kappa_{ij} := \sigma_j$$

otherwise for $\rho_{ij} < a_{ij}$

$$\mu_{ij} := \frac{\mu_i - a_{ij}\rho_{ij}(\mu_1 + \mu_2) + a^2 \mu_j}{\alpha_{ij}(1 - \rho_{ij}^2)} \quad \text{and} \quad \kappa_{ij} := \frac{\sigma_i^2 \sigma_j^2 (1 - \rho_{ij}^2)}{\sigma_i^2 - 2\rho_{ij}\sigma_i\sigma_j + \sigma_j^2}.$$

**Theorem 6.17.** *Let $(X_i, X_j)$ be a pair from a type I elliptical random vector $\boldsymbol{X} \sim E(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F)$ and consider $\gamma \nearrow x_F$. Then with $\upsilon_{ij}(\gamma) = (\gamma - \mu_{ij})/\kappa_{ij} + c_{ij}(\gamma)$ for some $c_{ij}(\gamma) \in \mathrm{o}(1)$,*

$$\mathbb{P}(X_i > \gamma, X_j > \gamma) = \overline{F}(\upsilon_{ij}(\gamma)) \times \begin{cases} \Big(2\pi\upsilon_{ij}(\gamma)w(\upsilon_{ij}(\gamma))\Big)^{-1/2}(1 + \mathrm{o}(1)), & \text{if } \rho_{ij} > a_{ij}, \\ \Big(2\pi\upsilon_{ij}(\gamma)w(\upsilon_{ij}(\gamma))\Big)^{-1}(C_{a,\rho} + \mathrm{o}(1)), & \text{if } \rho_{ij} < a_{ij}, \end{cases}$$

*for a $C_{a,\rho} \in \mathbb{R}_+$. Furthermore, if either $\mu_i \geq \mu_j$ or $\lim_{\gamma \to x_F} w(\gamma)/\gamma < \infty$, then there exists a $C_\rho \in \mathbb{R}_+$ such that*

$$\mathbb{P}(X_i > \gamma, X_j > \gamma) = \overline{F}(\upsilon_{ij}(\gamma))\Big(2\pi\upsilon_{ij}(\gamma)w(\upsilon_{ij}(\gamma))\Big)^{-1/2}(C_\rho + \mathrm{o}(1)), \qquad \text{if } \rho_{ij} = a_{ij}.$$

*Proof.* Use Theorem 2 of Hashorva [90]. First we consider the case $a_{ij} < \rho_{ij}$. In such a case it holds that

$$\begin{aligned} \lim_{\gamma \to x_F} \sqrt{\frac{w(\upsilon_j(\gamma))}{\upsilon_j(\gamma)}}\,(\upsilon_i(\gamma) - \rho_{ij}\upsilon_j(\gamma)) &= \lim_{\gamma \to x_F} \sqrt{w(\upsilon_j(\gamma))\upsilon_j(\gamma)}\left(\frac{\upsilon_i(\gamma)}{\upsilon_j(\gamma)} - \rho_{ij}\right) \\ &= \lim_{\gamma \to x_F} \sqrt{w(\upsilon_j(\gamma))\upsilon_j(\gamma)}\,(a_{ij} - \rho_{ij}) = -\infty. \end{aligned}$$

Hence, the hypotheses of Case i) of Theorem 2 of Hashorva [90] hold and the first result follows. In the case where $a_{ij} = \rho_{ij}$ then

$$\lim_{\gamma \to x_F} \sqrt{\frac{w(\upsilon_j(\gamma))}{\upsilon_j(\gamma)}}\,(\upsilon_i(\gamma) - \rho_{ij}\upsilon_j(\gamma)) = \lim_{\gamma \to x_F} \sqrt{\frac{w(\upsilon_j(\gamma))}{\upsilon_j(\gamma)}}\frac{(\mu_j - \mu_i)}{\sigma_i}.$$

The last limit remains bounded from above if either $\mu_i > \mu_j$ or $\lim_{\gamma\to\infty} w(\gamma)/\gamma < \infty$. For the case $a_{ij} > \rho_{ij}$ we define $a_{ij}(\gamma) := \upsilon_i(\gamma)/\upsilon_j(\gamma)$ so $\lim_{\gamma\to\infty} a_{ij}(\gamma) = a_{ij}$.

We let

$$\tau_{ij}(\gamma) = \sqrt{\frac{1 - 2a_{ij}(\gamma)\rho_{ij} + a_{ij}^2(\gamma)}{1 - \rho_{ij}^2}}, \quad \tau_{ij} := \lim_{\gamma\to\infty} \tau_{ij}(\gamma) = \sqrt{\frac{1 - 2a_{ij}\rho_{ij} + a_{ij}^2}{1 - \rho_{ij}^2}}.$$

The results follows by noting that

$$\upsilon_j(\gamma)\tau_{ij}(\gamma) = \upsilon_{ij}(\gamma), \qquad \upsilon_{ij}(\gamma) = \frac{\gamma - \mu_{ij}}{\tau_{ij}} + \mathrm{o}(1).$$

□

# Bibliography


[1] Joseph Abate, Gagan L. Choudhury, and Ward Whitt. On the Laguerre method for numerically inverting Laplace transforms. *INFORMS Journal on Computing*, 8(4):413–427, 1995.

[2] Joseph Abate and Ward Whitt. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10(1):5–87, 1992.

[3] Joseph Abate and Ward Whitt. A unified framework for numerically inverting Laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421, 2006.

[4] Adnan A. Abu-Dayya and Norman C. Beaulieu. Outage probabilities in the presence of correlated lognormal interferers. *IEEE Transactions on Vehicular Technology*, 43(1):164–173, 1994.

[5] Robert J. Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, volume 12 of *Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 1990.

[6] Robert J. Adler, Jose Blanchet, and Jingchen Liu. Efficient Monte Carlo for high excursions of Gaussian random fields. *The Annals of Applied Probability*, 22(3):1167–1214, 2012.

[7] John Aitchison and James AC Brown. *The Lognormal Distribution with Special Reference to Its Uses in Economics*, volume 5 of *University of Cambridge Department of Applied Economics Monographs*. Cambridge University Press, 1957.

[8] Stan Alink, Matthias Löwe, and Mario V. Wüthrich. Diversification of aggregate dependent risks. *Insurance: Mathematics and Economics*, 35(1):77–95, 2004.

[9] Stan Alink, Matthias Löwe, and Mario V. Wüthrich. Diversification for general copula dependence. *Statistica Neerlandica*, 61(4):446–465, 2007.

[10] Lars Nørvang Andersen, Patrick J. Laub, and Leonardo Rojas-Nandayapa. *Online accompaniment for "Rare-event simulation for extrema of dependent random variables"*, 2016. Available at `https://github.com/Pat-Laub/RareMaxima`.

[11] Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Probability and Mathematical Statistics. John Wiley & Sons, 3rd edition, 2003.

[12] Søren Asmussen. *Applied Probability and Queues*, volume 51 of *Stochastic Modelling and Applied Probability Series*. Springer–Verlag, 2nd edition, 2003.

[13] Søren Asmussen. Conditional Monte Carlo for sums, with applications to insurance and finance. *Annals of Actuarial Science*, 12(2):455–478, 2018.

[14] Søren Asmussen and Hansjörg Albrecher. *Ruin Probabilities*, volume 14 of *Advanced Series on Statistical Science and Applied Probability*. World Scientific, 2nd edition, 2010.

[15] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57 of *Stochastic Modelling and Applied Probability Series*. Springer, 2007.

[16] Søren Asmussen, Pierre-Olivier Goffard, and Patrick J. Laub. *Online accompaniment for "Orthonormal polynomial expansions and lognormal sum densities"*, 2016. Available at `https://github.com/Pat-Laub/SLNOrthogonalPolynomials`.

[17] Søren Asmussen, Enkelejd Hashorva, Patrick J. Laub, and Thomas Taimre. Tail asymptotics for light-tailed Weibull-like sums. *Probability and Mathematical Statistics*, 37(2), 2017.

[18] Søren Asmussen, Jevgenijs Ivanovs, Patrick J. Laub, and Hailiang Yang. Phase-type models in life insurance: Fitting and valuation of equity-linked benefits. *Risks*, 2018. To be submitted.

[19] Søren Asmussen, Jens Ledet Jensen, and Leonardo Rojas-Nandayapa. Exponential family techniques for the lognormal left tail. *Scandinavian Journal of Statistics*, 43(3):774–787, 2016.

[20] Søren Asmussen, Jens Ledet Jensen, and Leonardo Rojas-Nandayapa. On the Laplace transform of the lognormal distribution. *Methodology and Computing in Applied Probability*, 18(2):441âĂŞ–458, 2016.

[21] Søren Asmussen and Dirk P. Kroese. Improved algorithms for rare event simulation with heavy tails. *Advances in Applied Probability*, 38(2):545–558, 2006.

[22] Søren Asmussen, Olle Nerman, and Marita Olsson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[23] Søren Asmussen and Leonardo Rojas-Nandayapa. Asymptotics of sums of lognormal random variables with Gaussian copula. *Statistics & Probability Letters*, 78(16):2709–2714, 2008.

[24] Efstathios Avdis and Ward Whitt. Power algorithms for inverting Laplace transforms. *INFORMS Journal on Computing*, 19(3):341–355, 2007.

[25] August A. Balkema, Claudia Klüppelberg, and Sidney I. Resnick. Densities with Gaussian tails. *Proceedings of the London Mathematical Society*, 3(3):568–588, 1993.

[26] Ole E. Barndorff-Nielsen and David Roxbee Cox. *Asymptotic Techniques for Use in Statistics*, volume 31 of *Monographs on Statistics and Applied Probability*. Springer Science & Business Media, 1989.

[27] Norman C. Beaulieu, Adnan A. Abu-Dayya, and Peter J. McLane. Estimating the distribution of a sum of independent lognormal random variables. *IEEE Transactions on Communications*, 43(12):2869–2873, 1995.

[28] Norman C. Beaulieu and Faruq Rajwani. Highly accurate simple closed-form approximations to lognormal sum distributions and densities. *IEEE Communications Letters*, 8(12):709–711, 2004.

[29] Norman C. Beaulieu and Qiong Xie. An optimal lognormal approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology*, 53(2):479–489, 2004.

[30] Mário N. Berberan-Santos. Expressing a probability density function in terms of another PDF: A generalized Gram–Charlier expansion. *Journal of Mathematical Chemistry*, 42(3):585–594, 2007.

[31] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*, volume 100 of *Graduate Texts in Mathematics*. Springer–Verlag, 1984.

[32] Simeon Berman. *Sojourns and Extremes of Stochastic Processes*. Wadsworth and Brooks/Cole Statistics/Probability Series. CRC Press, 1992.

[33] Nicholas H. Bingham, Charles M. Goldie, and Jozef L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1989.

[34] Mogens Bladt. A review on phase-type distributions and their use in risk theory. *ASTIN Bulletin: The Journal of the IAA*, 35(1):145–161, 2005.

[35] Zdravko I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B*, 79(1):125–148, 2017.

[36] Zdravko I. Botev, Joseph F. Grotowski, and Dirk P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010.

[37] Zdravko I. Botev, Robert Salomone, and Daniel MacKinlay. Fast and accurate computation of the distribution of sums of dependent log-normals. *arXiv Preprint 1705.03196*, 2017.

[38] Jon A. Breslaw. Random sampling from a truncated multivariate normal distribution. *Applied Mathematics Letters*, 7(1):1–6, 1994.

[39] Wlodzimierz Bryc. *The Normal Distribution: Characterizations with Applications*, volume 100 of *Lecture Notes in Statistics*. Springer Science & Business Media, 2012.

[40] Jun Cai, Ken Seng Tan, Chengguo Weng, and Yi Zhang. Optimal reinsurance under VaR and CTE risk measures. *Insurance: Mathematics and Economics*, 43(1):185–196, 2008.

[41] Mathieu Cambou, Marius Hofert, and Christiane Lemieux. Quasi-random numbers for copula models. *Statistics and Computing*, 27(5):1307–1329, 2017.

[42] Frédéric Cérou and Arnaud Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007.

[43] Joshua C. C. Chan and Dirk P. Kroese. Improved cross-entropy method for estimation. *Statistics and Computing*, 22(5):1031–1040, 2012.

[44] Arthur Charpentier and Johan Segers. Tails of multivariate Archimedean copulas. *Journal of Multivariate Analysis*, 100(7):1521–1537, 2009.

[45] Jean-Pierre Chateau and Daniel Dufresne. Gram–Charlier processes and applications to option pricing. *Journal of Probability and Statistics*, volume 2017, 2017.

[46] Ka Chun Cheung. Optimal reinsurance revisited: a geometric approach. *ASTIN Bulletin*, 40(1):221–239, 005 2010.

[47] Yichun Chi and Ken Seng Tan. Optimal reinsurance under VaR and CVaR risk measures: a simplified approach. *ASTIN Bulletin*, 41(2):487–509, 2011.

[48] Theodore Seio Chihara. On generalized Stieljes–Wigert and related orthogonal polynomials. *Journal of Computational and Applied Mathematics*, 5(4):291–297, 1979.

[49] Theodore Seio Chihara. *An Introduction to Orthogonal Polynomials*. Courier Corporation, 2011.

[50] Jacob Stordal Christiansen. The moment problem associated with the Stieljes–Wigert polynomials. *Journal of Mathematical Analysis and Applications*, 277(1):218–245, 2003.

[51] Leon Cohen. On the generalization of the Edgeworth/Gram–Charlier series. *Journal of Mathematical Chemistry*, 49(3):625–628, 2011.

[52] Rama Cont. *Encyclopedia of Quantitative Finance*. John Wiley & Sons, 2010.

[53] Robert M. Corless, Gaston H. Gonnet, David E. G. Hare, David J. Jeffrey, and Donald E. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.

[54] Harald Cramér. *On Some Classes of Series used in Mathematical Statistics*. Hoffenberg, 1926.

[55] Harald Cramér. *On the Mathematical Theory of Risk*. Skandia Jubilee Volume, Stockholm, 1930.

[56] Edwin L. Crow and Kunio Shimizu, editors. *Lognormal Distributions: Theory and Applications*, volume 88 of *Statistics Textbooks and Monographs*. Marcel Dekker, 1988.

[57] Pieter-Tjerk De Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[58] Laurens De Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Operations Research and Financial Engineering. Springer Science & Business Media, 2007.

[59] Michel Denuit, Jan Dhaene, Marc J. Goovaert, and Rob Kaas. *Actuarial Theory for Dependent Risk: Measures, Orders and Models*. John Wiley & Sons, 2006.

[60] Jean-Dominique Deuschel and Daniel W. Stroock. *Large Deviations*, volume 342. AMS Chelsea Publishing, 2001.

[61] Josef Dick and Friedrich Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.

[62] Klaus Duellmann. *Regulatory Capital*, volume IV, pages 1525–1538. John Wiley & Sons, 2010.

[63] Daniel Dufresne. Laguerre series for Asian and other options. *Mathematical Finance*, 10(4):407–428, 2000.

[64] Daniel Dufresne. The log-normal approximation in financial and other computations. *Advances in Applied Probability*, 36(3):747–773, 2004.

[65] Daniel Dufresne. Sums of lognormals. *Proceedings of the 43rd Actuarial Science Research Conference*, 2008.

[66] Daniel Dufresne, Jose Garrido, and Manuel Morales. Fourier inversion formulas in option pricing and insurance. *Methodology and Computing in Applied Probability*, 11(3):359–383, 2009.

[67] Daniel Dufresne and H. Li. Pricing Asian options: convergence of Gram–Charlier series. Technical report, Actuarial Research Clearing House, 2016.

[68] Torbjørn Eltoft, Taesu Kim, and Te-Won Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

[69] Paul Embrechts and Marco Frei. Panjer recursion versus FFT for compound distributions. *Mathematical Methods of Operations Research*, 69(3):497–508, 2009.

[70] Paul Embrechts, Giovanni Puccetti, Ludger Rüschendorf, Ruodu Wang, and Antonela Beleraj. An academic response to Basel 3.5. *Risks*, 2(1):25–48, 2014.

[71] Lawrence Fenton. The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, 8(1):57–67, 1960.

[72] Carlo Fischione, Fabio Graziosi, and Fortunato Santucci. Approximation for a sum of on-off lognormal processes with wireless applications. *IEEE Transactions on Communications*, 55(10):1984–1993, 2007.

[73] Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Operations Research and Financial Engineering. Springer Science & Business Media, 2013.

[74] Serguei Foss and Andrew Richards. On sums of conditionally independent subexponential random variables. *Mathematics of Operations Research*, 35(1):102–119, 2010.

[75] Antoine Frachot, Pierre Georges, and Thierry Roncalli. Loss distribution approach for operational risk. Technical report, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France, 2001.

[76] Antoine Frachot, Olivier Moudoulaud, and Thierry Roncalli. Loss distribution approach in practice. Technical report, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France, 2003.

[77] Xin Gao, Hong Xu, and Dong Ye. Asymptotic behavior of tail density for sum of correlated lognormal variables. *International Journal of Mathematics and Mathematical Sciences*, volume 2009, 2009. Article ID 630857.

[78] Walter Gautschi. *Numerical Analysis*. SpringerLink: Bücher. Birkhäuser Boston, 2nd edition, 2012.

[79] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53 of *Stochastic Modelling and Applied Probability Series*. Springer, 2003.

[80] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, and Tim Zajic. Splitting for rare event simulation: analysis of simple cases. In *Proceedings of the 28th Conference on Winter simulation*, pages 302–308. IEEE Computer Society, 1996.

[81] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, and Tim Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600, 1999.

[82] Pierre-Olivier Goffard and Patrick J. Laub. *Online accompaniment for "Two numerical methods to evaluate stop-loss premiums"*, 2017. Available at `https://github.com/Pat-Laub/ActuarialOrthogonalPolynomials`.

[83] Pierre-Olivier Goffard, Stéphane Loisel, and Denys Pommeret. Polynomial approximations for bivariate aggregate claims amount probability distributions. *Methodology and Computing in Applied Probability*, 19(1):151–174, 2015.

[84] Pierre-Olivier Goffard, Stéphane Loisel, and Denys Pommeret. A polynomial expansion to approximate the ultimate ruin probability in the compound Poisson ruin model. *Journal of Computational and Applied Mathematics*, 296:499–511, 2016.

[85] Archil Gulisashvili and Peter Tankov. Tail behavior of sums and differences of log-normal random variables. *Bernoulli*, 22(1):444–493, 2016.

[86] Henryk Gzyl, Pier Luigi Novi Inverardi, and Aldo Tagliani. Determination of the probability of ultimate ruin probability by maximum entropy applied to fractional moments. *Insurance: Mathematics and Economics*, 53(2):457–463, 2013.

[87] Henryk Gzyl and Aldo Tagliani. Determination of the distribution of total loss from the fractional moments of its exponential. *Applied Mathematics and Computation*, 219(4):2124–2133, 2012.

[88] Anders Hald. The early history of the cumulants and the Gram–Charlier series. *International Statistical Review*, 68(2):137–153, 2000.

[89] Enkelejd Hashorva. Asymptotics and bounds for multivariate Gaussian tails. *Journal of Theoretical Probability*, 18(1):79–97, 2005.

[90] Enkelejd Hashorva. Asymptotic properties of type I elliptical random vectors. *Extremes*, 10(4):175–206, 2007.

[91] Enkelejd Hashorva. On the residual dependence index of elliptical distributions. *Statistics & Probability Letters*, 80(13):1070–1078, 2010.

[92] Abdelhamid Hassairi and Mohammed Zarai. Characterization of the cubic exponential families by orthogonality of polynomials. *The Annals of Probability*, 32(3):2463–2476, 2004.

[93] Amin Hassan Zadeh. *Actuarial applications of multivariate phase-type distributions: model calibration and credibility*. PhD thesis, Université de Montréal, 2009.

[94] Amin Hassan Zadeh, Bruce L. Jones, and David A. Stanford. The use of phase-type models for disability insurance calculations. *Scandinavian Actuarial Journal*, 2014(8):714–728, 2014.

[95] Marwane Ben Hcine and Ridha Bouallegue. Highly accurate log skew normal approximation to the sum of correlated lognormals. In *Proceedings of the 2014 International Conference on Networks & Communications*, 2015.

[96] Janet E Heffernan. A directory of coefficients of tail dependence. *Extremes*, 3(3):279–290, 2000.

[97] Markus Hegland. Lecture notes for the AMSI Summer School course on Computational Mathematics, 2017.

[98] Chris C. Heyde. On a property of the lognormal distribution. *Journal of the Royal Statistical Society: Series B*, 5:392–393, 1963.

[99] Zhi Huang and Perwez Shahabuddin. Rare-event, heavy-tailed simulations using hazard function transformations, with applications to value-at-risk. In *Proceedings of the 2003 Winter Simulation Conference*, 2003.

[100] Henrik Hult and Filip Lindskog. Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34(3):587–608, 2002.

[101] European Insurance and Occupational Pensions Authority. Quantitative impact studies V: Technical specifications. Technical report, European Comission, Brussels, 2010.

[102] Tao Jin, Serge B. Provost, and Jiandong Ren. Moment-based density approximations for aggregate losses. *Scandinavian Actuarial Journal*, 2016(3):216–245, 2016.

[103] Harry Joe. *Multivariate Models and Multivariate Dependence Concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. CRC Press, 1997.

[104] Norman L. Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous Univariate Distributions (Volume 1: Models and Applications)*. Applied Probability and Statistics. John Wiley & Sons, 2nd edition, 1994.

[105] Sandeep Juneja and Perwez Shahabuddin. Simulating heavy tailed processes using delayed hazard rate twisting. *ACM Transactions on Modeling and Computer Simulation*, 12(2):94–118, 2002.

[106] Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot. *Loss Models: From Data to Decisions*, volume 715 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 3rd edition, 2012.

[107] Claudia Klüppelberg, Daniel Straub, and Isabell M. Welpe, editors. *Risk: A Multidisciplinary Introduction*. Springer, 2014.

[108] John E Kolassa. *Series Approximation Methods in Statistics*, volume 88 of *Lecture Notes in Statistics*. Springer Science & Business Media, 3rd edition, 2006.

[109] Samuel Kotz, Tomaz J. Kozubowski, and Krzysztof Podgórski. Asymmetric multivariate Laplace distribution. In *The Laplace Distribution and Generalizations*, pages 239–272. Springer, 2001.

[110] Dirk P. Kroese, Thomas Taimre, and Zdravko I. Botev. *Handbook of Monte Carlo Methods*, volume 706 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2013.

[111] Patrick J. Laub, Søren Asmussen, Jens Ledet Jensen, and Leonardo Rojas-Nandayapa. *Online accompaniment for "Approximating the Laplace transform of the sum of dependent lognormals"*, 2016. Available at `https://github.com/Pat-Laub/SLNLaplaceTransformApprox`.

[112] Patrick J. Laub, Robert Salomone, and Zdravko I. Botev. Monte Carlo estimation of the density of the sum of dependent random variables. *Mathematics and Computers in Simulation*, 2018. Under Revision.

[113] Anthony W. Ledford and Jonathan A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187, 1996.

[114] Anthony W. Ledford and Jonathan A. Tawn. Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B*, 59(2):475–499, 1997.

[115] Anthony W. Ledford and Jonathan A. Tawn. Concomitant tail behaviour for extremes. *Advances in Applied Probability*, 30(1):197–215, 1998.

[116] Simon C. K. Lee and X. Sheldon Lin. Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, 14(1):107–130, 2010.

[117] Claude Lefèvre and Philippe Picard. A new look at the homogeneous risk model. *Insurance: Mathematics and Economics*, 49(3):512–519, 2011.

[118] Claude Lefèvre, Julien Trufin, and Pierre Zuyderhoff. Some comparison results for finite-time ruin probabilities in the classical risk model. *Insurance: Mathematics and Economics*, 77(Supplement C):143–149, 2017.

[119] Eckhard Limpert, Werner A. Stahel, and Markus Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352, 2001.

[120] Filip Lundberg. *I. Approximerad Framställning af Sannolikhetsfunktionen, II. Återförsäkring af Kollektivrisker*. Almqvist & Wiksell, Uppsala, 1903.

[121] Arnaud Mallet. *Numerical inversion of Laplace transform*, 2000. Mathematica package, accessed online on 26th August 2015 at `http://library.wolfram.com/infocenter/MathSource/2691`.

[122] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[123] Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2nd edition, 2015.

[124] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.

[125] Thomas Mikosch. Copulas: tales and facts. *Extremes*, 9(3), 2006.

[126] Moshe Arye Milevsky and Steven E Posner. Asian options, the sum of lognormals, and the reciprocal gamma distribution. *Journal of Financial and Quantitative Analysis*, 33(03):409–422, 1998.

[127] Robert M. Mnatsakanov and Khachatur Sarkisian. A note on recovering the distributions from exponential moments. *Applied Mathematics and Computation*, 219(16):8730–8737, 2013.

[128] Robert M. Mnatsakanov, Khachatur Sarkisian, and A. Hakobyan. Approximation of the ruin probability using the scaled Laplace transform inversion. *Applied Mathematics and Computation*, 268:717–727, 2015.

[129] Carl N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80, 1982.

[130] Saralees Nadarajah. A review of results on sums of random variables. *Acta Applicandae Mathematicae*, 103(2):131–140, 2008.

[131] Saralees Nadarajah, Jeffrey Chu, and Xiao Jiang. On moment based density approximations for aggregate losses. *Journal of Computational and Applied Mathematics*, 298:152–166, 2016.

[132] Roger B Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer Science & Business Media, 2nd edition, 2006.

[133] Ryuei Nishii. *Orthogonal Functions of Inverse Gaussian Distributions*, pages 243–250. Springer, 1996.

[134] Natalia Nolde. Geometric interpretation of the residual dependence coefficient. *Journal of Multivariate Analysis*, 123:85–95, 2014.

[135] Marita Olsson. Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, 23:443–460, 1996.

[136] Adrian Pagan and Aman Ullah. *Nonparametric Econometrics*. Cambridge University Press, 1999.

[137] Harry H. Panjer and Gordon E. Willmot. Finite sum evaluation of the negative binomial-exponential model. *ASTIN Bulletin: The Journal of the IAA*, 12(2):133–137, 1981.

[138] Dmitry E. Papush, Gary S. Patrik, and Felix Podgaits. Approximations of the aggregate loss distribution. *CAS Forum (Winter)*, pages 175–186, 2001.

[139] Steven E. Pav. *PDQutils: PDQ Functions via Gram Charlier, Edgeworth, and Cornish Fisher Approximations*, 2017. R package version 0.1.6.

[140] Ray Popovic and David Goldsman. Easy Gram–Charlier valuations of options. *Journal of Derivatives*, 20(2):79–97, 2012.

[141] Serge B. Provost. Moment-based density approximants. *Mathematica Journal*, 9(4):727–756, 2005.

[142] Marvin Rausand and Arnljot Høyland. *System Reliability Theory: Models, Statistical Methods, and Applications*, volume 396 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2nd edition, 2003.

[143] Sidney I. Resnick. Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5(4):303–336, 2002.

[144] Sidney I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. Springer, 2013.

[145] Christian P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5(2):121–125, 1995.

[146] Leonardo Rojas-Nandayapa and Wangyue Xie. Asymptotic tail behaviour of phase-type scale mixture distributions. *Annals of Actuarial Science*, 12(2):412–432, 2018.

[147] Tomasz Rolski, Hanspeter Schmidli, Volker Schmidt, and Jozef L. Teugels. *Stochastic Processes for Insurance and Finance*, volume 505 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2009.

[148] Gerardo Rubino and Bruno Tuffin, editors. *Rare Event Simulation using Monte Carlo Methods*. John Wiley & Sons, 2009.

[149] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*, volume 707 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2nd edition, 2011.

[150] Ludger Rüschendorf. *Mathematical Risk Analysis*. Springer Series in Operations Research and Financial Engineering. Springer, 2013.

[151] Felix Salmon. Recipe for disaster: The formula that killed Wall Street, 2009. News article on wired.com from 23/02/2009.

[152] I. Richard Savage. Mills' ratio for multivariate normal distributions. *Journal of research of the National Bureau of Standards: Section B*, 66:93–96, 1962.

[153] Stuart C. Schwartz and Yu-Shuan Yeh. On the distribution function and moments of power sums with log-normal components. *Bell System Technical Journal*, 61(7):1441–1462, 1982.

[154] Damith Senaratne and Chintha Tellambura. Numerical computation of the log-normal sum distribution. In *Proceedings of the 28th IEEE conference on Global Telecommunications*, pages 3966–3971. IEEE Press, 2009.

[155] Masaaki Sibuya. Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210, 1960.

[156] Harald Stehfest. Algorithm 368: Numerical inversion of Laplace transforms [d5]. *Communications of the ACM*, 13(1):47–49, 1970.

[157] Thomas Joannes Stieljes. *Recherches sur les fractions continues*. Annales de la Faculté des sciences de Toulouse: Mathématiques, 1894.

[158] Gabor Szegö. *Orthogonal Polynomials*, volume XXIII. American Mathematical Society Colloquium Publications, 1939.

[159] Béla Szökefalvi-Nagy. *Introduction to Real Functions and Orthogonal Expansions*. Akadémiai Kiadó, 1965.

[160] Stuart M. Turnbull and Lee Macdonald Wakeman. A quick algorithm for pricing European average options. *Journal of Financial and Quantitative Analysis*, 26(3):377–389, 1991.

[161] Matt P. Wand and M. Chris Jones. *Kernel Smoothing*. Number 60 in Monographs on Statistics and Applied Probability. CRC Press, 1994.

[162] Carl Severin Wigert. Sur les polynomes orthogonaux et l'approximation des fonctions continues. *Almqvist and Wiksell*, 1923.

[163] Gordon E. Willmot and X. Sheldon Lin. Risk modelling with the mixed Erlang distribution. *Applied Stochastic Models in Business and Industry*, 27(1):2–16, 2011.

[164] Gordon E. Willmot and Jae-Kyung Woo. On the class of Erlang mixtures with risk theoretic applications. *North American Actuarial Journal*, 11(2):99–115, 2007.

[165] Mario V. Wüthrich. Asymptotic value-at-risk estimates for sums of dependent random variables. *Astin Bulletin*, 33(01):75–92, 2003.

[166] Hui Yao, Leonardo Rojas-Nandayapa, and Thomas Taimre. Estimating tail probabilities of random sums of infinite mixtures of phase-type distributions. In *Proceedings of the 2016 Winter Simulation Conference*, pages 347–358. IEEE Press, 2016.