SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Benchmarking Saddle Point Approximations of the Gamma Distribution

av

Elvis Wisdom

2026 - No K27

Benchmarking Saddle Point Approximations of the Gamma Distribution

Elvis Wisdom

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Yishao Zhou

2026

Abstract

Daniels'[2] saddle point approximation recovers the probability density from the moment generating function by using the method of steepest descent to approximate the inverse Laplace transform. Lugannani-Rice[3] extends this method to also recover the distribution function in a similar way. This thesis uses the Lugannani-Rice method to derive an approximation for the Gamma cumulative distribution function, which coincides with Temme's[8] approximation to the leading term. This approximation is benchmarked for accuracy and runtime against the approximation methods used in DiDonato-Morris[7], which includes Temme. The Lugannani-Rice/Temme approximation is fast to compute and achieves high accuracy when approximating tail probabilities and when the shape parameter α is large. The main downside of the approximation is that it is difficult to compute higher order terms which creates an upper bound for accuracy. The main advantage is that it fills a niche by being accurate for large α even at moderate deviations.

Sammanfattning

Daniels[2] sadelpunktmetod approximerar täthetsfunktioner utifrån fördelningens momentgenererande funktion genom att skatta den inversa laplacetransformen med hjälp av den komplexvärda förlängningen av Laplaces metod. Lugannani-Rice[3] använder samma princip för att även skatta den kumulativa fördelningsfunktionen. I denna uppsats används Lugannani-Rice metoden för att skatta gammafördelningens kumulativa fördelningsfunktion, vilket leder till ett uttryck som motsvarar Temmes[8] skattning till första ordning. Denna approximation jämförs sedan med de andra skattningsmetoderna som ingår i DiDonato-Morris[7] för att utvärdera hastighet och noggrannhet. Resultaten visar att den härledda metoden är snabb och uppvisar hög noggrannhet när den används för att skatta fördelningens svansar, samt när α-parametern är stor. Den huvudsakliga bristen med metoden är att det är svårt att beräkna termer som krävs för skattningar av högre ordning. Den största fördelen med metoden är att den uppvisar hög noggranhet för stora α-värden, vilket andra metoder inte uppvisar. På så vis fyller den en nisch.

Acknowledgements

I want to give a special thanks to my thesis supervisor Professor Yishao Zhou for her encouragement, guidance and insights. I would also like to thank Alma Söderström, Felix Hasselblad, Gabriel Paulsrud, Harald Enoksson and Jörn Falk for reading my drafts and providing valuable feedback. Finally I want to thank my family for their never ending love and support.

Contents

1 Introduction

The probability distribution of a stochastic variable can be described by a variety of different functions, such as the cumulative distribution function (CDF) $F(x) := \mathbb{P}(X \leq x)$ and the cumulant generating function (CGF) $K(s) := \log \mathbb{E}[e^{sX}]$. In general it is difficult to analytically transform $K(s) \to F(x)$, but Lugannani-Rice[3] provides a formula for approximating such a transform.

Theorem 1.1. Lugannani-Rice Approximation[3]
The CDF $F(x)$ can be approximated using the CGF $K(s)$ by the formula below. The variable s is evaluated in the *saddle point* $\hat{s}$ which is the s which solves $K'(\hat{s}) = x$

$$F(x) \approx \begin{cases} \Phi(\hat{w}) + \phi(\hat{w}) \left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}} \right) & \text{for } x \neq \mu, \\ \frac{1}{2} + \frac{K'''(0)}{6\sqrt{2\pi}\, K''(0)^{3/2}} & \text{for } x = \mu, \end{cases}$$
$$\hat{w} = \operatorname{sgn}(\hat{s}) \sqrt{2(\hat{s}x - K(\hat{s}))}, \quad \hat{u} = \hat{s} \sqrt{K''(\hat{s})}$$

where $\Phi(x)$ and $\phi(x)$ are the standard normal $\mathcal{N}(0,1)$ CDF and probability density function(PDF) respectively, μ is the mean value of the distribution. The values w and u are auxiliary variables and the hat indicates that they are evaluated at the saddle point $\hat{s}$.

Some probability distributions such as the Gamma distribution have a simple CGF, but a complicated CDF which requires numerical approximations to evaluate. The core idea of this thesis is to use the simple CGF and relatively simple Lugannani-Rice formula to produce an approximation for the gamma CDF and then evaluate its usefulness.

The Gamma distribution has CDF and CGF

$$F(x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}, \quad K(s) = -\alpha \log\left(1 - \frac{s}{\beta}\right).$$

Using Lugannani-Rice to approximate $F(x)$ with $K(s)$ gives

$$F(x) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w}) \left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}} \right) & \text{for } x \neq \frac{\alpha}{\beta}, \\ \frac{1}{2} + \frac{1}{3\sqrt{2\pi\alpha}} & \text{for } x = \frac{\alpha}{\beta}, \end{cases}$$
$$\hat{w} = \operatorname{sgn}(\beta - \frac{\alpha}{x}) \sqrt{2\left(\beta x - \alpha - \alpha \log\left(\frac{\beta x}{\alpha}\right)\right)}, \quad \hat{u} = \left(\frac{\beta x - \alpha}{\sqrt{\alpha}}\right).$$

This approximation of the gamma CDF will get benchmarked against other common approximations later in this thesis. Results show that this approximation is fast and achieves high accuracy in the tails as well as for large α-values.

Particularly the accuracy for large α-values sets it apart from other methods and fills a niche.

Temme[8] published an asymptotic approximation for the gamma distribution before the Lugannani-Rice[3] approximation was published and to the first order (at least) the Temme approximation coincides with the Lugannani-Rice approximation of the gamma distribution. This fact is not acknowledged in any of the surveyed literature.

The structure of the thesis is as follows: Section 2 will cover the basic probability theory needed to understand the problem being studied and defines terminology which will be used extensively throughout the thesis. Section 3 is the most theory-heavy chapter and constructs the framework needed to understand the method of steepest descent. Section 4 will use this theoretical foundation to derive the saddle point and Lugannani-Rice approximations. Section 5 constitutes the beginning of a more empirical focus in the second half of the thesis and will describe the current landscape of approximation methods for the gamma distribution. Section 6 is technical in nature and will cover details about methods and metrics used for benchmarking the approximation methods. Section 7 will cover the results from this benchmarking. Section 8 covers discussion, conclusions and future research.

Glossary

- PDF: Probability density function $f(x)$
- CDF: Cumulative distribution function $F(x)$
- MGF: Moment generating function $M(s)$
- CGF: Cumulant generating function $K(s)$
- SPA: Saddle point approximation $K(s) \mapsto f(x)$
- LR: Lugannani-Rice (approximation) $K(s) \mapsto F(x)$

2 Probability Theory Foundations

This thesis uses four different functions which describe probability distributions. In order to reduce the confusion this creates, all of these will be defined and explained in this section. This will be followed by a cursory explanation of what the saddle point approximation and Lugannani-Rice approximation does from a probability perspective and why it is useful. All specifics regarding how these methods work is delegated to the next two sections. The last part of this section will cover elementary properties of the gamma distribution specifically.

All distributions are assumed to be absolutely continuous and univariate.

Definition 2.1. Probability Density Function
The probability density function (PDF) $f(x)$ for a stochastic variable X is the function for which

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x)\,dx.$$

The PDF will often be the simplest function which describes a distribution and will be the starting point for many calculations.

Definition 2.2. Cumulative Distribution Function
The cumulative distribution function (CDF) $F(x)$ for stochastic variable X is such that

$$F(x) := \mathbb{P}(X \leq x).$$

It is the antiderivative of the PDF and can be calculated through

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

For the cases studied in this thesis the CDF will generally be the desired endpoint of calculations.

Definition 2.3. Moment Generating Function
The moment generating function (MGF) $M(s)$ for stochastic variable X is defined as

$$M(s) := \mathbb{E}[e^{sX}].$$

It can be calculated through the sign inverted two-sided Laplace transform of the PDF as

$$M(s) = \mathcal{L}(f(x))(-s) = \int_{-\infty}^{\infty} e^{sx} f(x)\,dx.$$

The MGF only exists if this integral converges in some neighbourhood of 0, this criteria is later described in more detail by Cramér's condition.

In several ways the MGF has much more desirable properties for analytic manipulation than the PDF or CDF, but it is also more abstract and translating an MGF back into a PDF/CDF requires solving the inverse Laplace transform, which is often difficult or impossible.

Definition 2.4. Cumulant Generating Function
The cumulant generating function (CGF) $K(s)$ of the stochastic variable X is the log-space version of the MGF, such that

$$K(s) := \log(M(s)).$$

Throughout this thesis it will generally be easier to work with this log-space

function, but it shares many properties with the MGF and it is easy to translate back and forth between CGF and MGF.

2.1 Mapping the Relationship

Putting these facts together then a map can be constructed which describes the interdependence of these four functions which all uniquely determine the distribution.

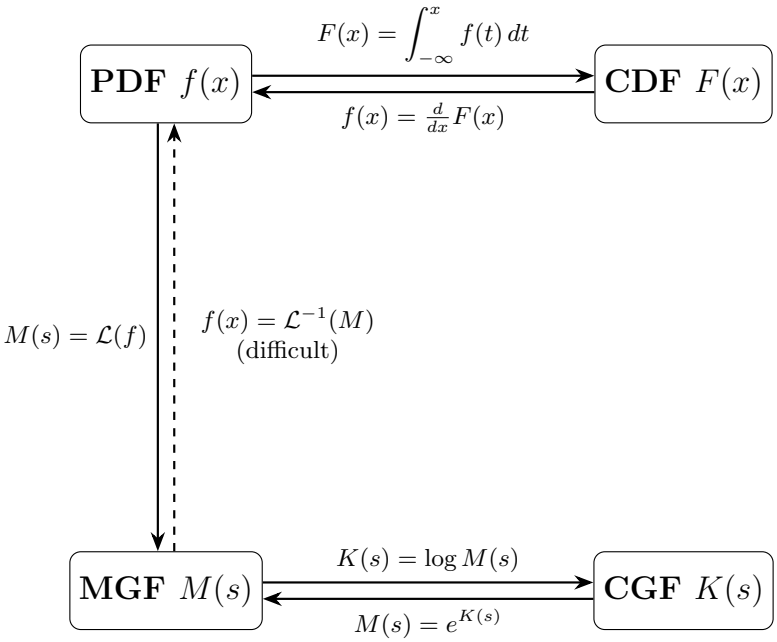

Figure 2.1: Probability function map

What becomes clear from this map is that due to the difficulty of solving the inverse Laplace transform, it is difficult to transform a CGF/MGF to a PDF/CDF. The saddle point approximation[2] uses a technique called the "method of steepest descent" in order to approximate the inverse Laplace transform in a way that becomes accurate in the tails of a distribution. This provides a way to approximate the PDF using the CGF/MGF. This PDF approximation is however difficult to integrate properly and therefore can not be easily translated into an approximation of the CDF. This is the contribution of Lugannani and Rice[3]. Their approximation is built on the same principles as the saddle point approximation, but their formula approximates the CDF instead of the PDF.

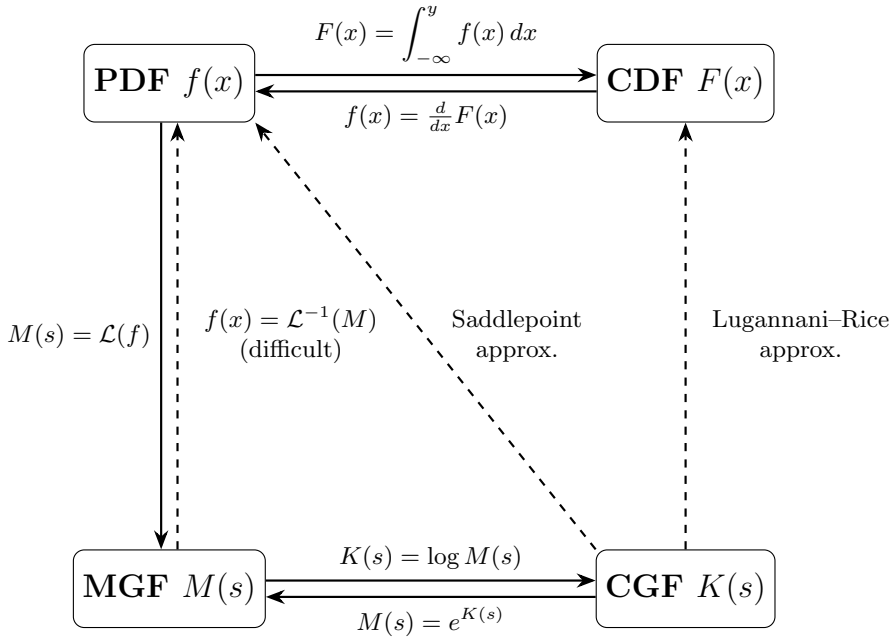

Figure 2.2: Probability function map with approximations

A quick note about terminology, sometimes the Lugannani-Rice approximation is described as "a saddle point approximation" since it is a natural extension of the technique developed by Daniels and since it is more descriptive of the underlying mechanism. Throughout this thesis however "saddle point approximation" (SPA) refers exclusively to Daniels' 1954 method and the later 1980 extension will always be called "Lugannani-Rice" (LR). This is purely for the purpose of clarity.

2.2 Gamma Distribution

The gamma distribution is a well known and thoroughly studied distribution which is described by the following functions:

$$\text{PDF:} \quad f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$\text{CDF:} \quad F(x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$$

$$\text{MGF:} \quad M(s) = \left(1 - \frac{s}{\beta}\right)^{-\alpha}, \quad s < \beta$$

$$\text{CGF:} \quad K(s) = -\alpha \log\left(1 - \frac{s}{\beta}\right), \quad s < \beta$$

$$\text{where} \quad \Gamma(z) := \int_0^\infty t^{z-1} e^{-t}\, dt \quad \text{and} \quad \gamma(z, a) := \int_0^a t^{z-1}\, e^{-t}\, dt.$$

The function $\Gamma(z)$ is known as the complete gamma function and is an extension of the factorial operator to the complex plane. When the factorial is well defined $n! = \Gamma(n+1)$. The function $\gamma(z, a)$ is known as the lower incomplete gamma function. Both of these functions are known as "special functions" because they are defined by integrals which lack a closed form solution, they can only be evaluated numerically except in special cases.

What makes the gamma distribution particularly well suited for LR approximations is that the CGF has a very simple expression, while the CDF has no closed form solution. The question being examined in this thesis is whether the LR approximation of the CDF is fast and accurate compared to other approximation methods.

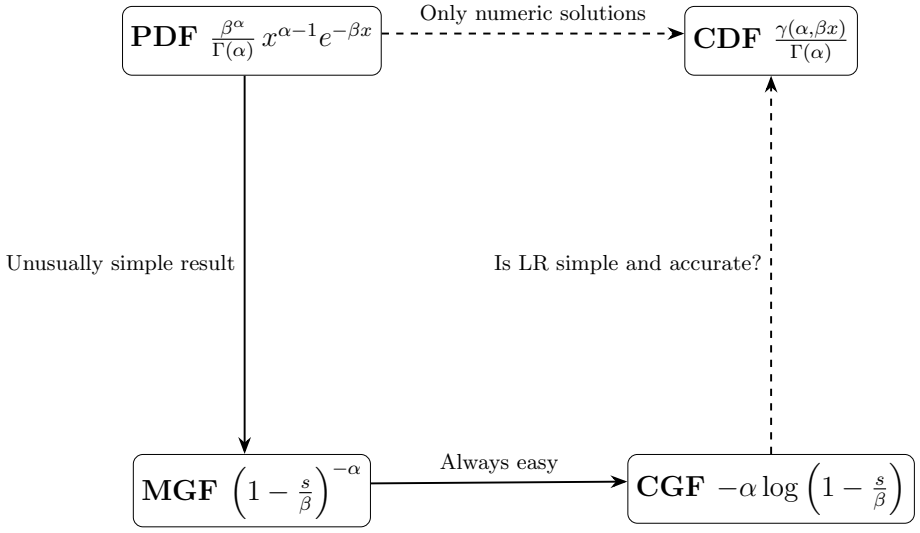

Figure 2.3: Map of intended procedure

This concludes the overview of what the LR approximation does and why it is

useful. The next two sections will attempt to describe how the method works.

3 Method of Steepest Descent

As mentioned in the previous section, the inverse Laplace transform is an integral which is in general difficult to solve. The SPA and LR use the "method of steepest descent" to approximate this transform. Explaining how steepest descent works is the primary goal of this section. The core underlying approximation is "Laplace's method" which is defined on the real number line $\mathbb{R}$. The method of steepest descent is an extension of Laplace's method to the complex plane $\mathbb{C}$. This extension will require an understanding of some concepts from complex analysis. Once the method of steepest descent is understood it is relatively trivial to derive the SPA, which will be done in the next section.

3.1 Laplace's Method

Laplace's method is a technique for estimating an integral of form

$$\int_a^b e^{Mf(x)}\,dx,$$

where M is some large number and $f(x)$ is a twice differentiable function which has some dominant interior maxima. The integration limits a and b may be finite or infinite.

Since the dominant maxima will be greatly amplified by the factor M and by the exponential function, the region surrounding this maxima will contribute the vast majority of the value of the integral.

When the value of an integral is dominated by a single region, an approximation which is accurate in that region is also fairly accurate for the whole integral. This justifies a Taylor expansion of $f(x)$ around that maxima at $x = \hat{x}$ identified through $f'(\hat{x}) = 0$,

$$f(x) = \sum_{n=0}^{\infty} f^{(n)}(\hat{x})\frac{(x-\hat{x})^n}{n!}.$$

The constant term $f(\hat{x})$ captures the maximum value of $f(x)$, which is important but doesn't describe how sharply $f(x)$ decays surrounding $\hat{x}$. The first order term is always 0 since $f'(\hat{x})$ is the first derivative of a stationary point. The second order term $\frac{1}{2}f''(\hat{x})(x-\hat{x})^2$ describes the curvature around the maxima, and precisely because this point is a maxima, the neighbourhood around the point will be a concave parabola, meaning that a second order term will describe this neighbourhood quite well. Including any higher order terms will make the integral much more difficult to evaluate and the second degree polynomial is already a good estimate for the immediate neighbourhood. The integral is approximated as

$$\int_a^b e^{Mf(x)}\,dx \approx \int_a^b e^{M(f(\hat{x})+\frac{1}{2}f''(\hat{x})(x-\hat{x})^2)}\,dx,$$
$$\approx e^{Mf(\hat{x})}\int_a^b e^{-M\frac{1}{2}|f''(\hat{x})|(x-\hat{x})^2}\,dx,$$
$$\text{where } f(\hat{x}) = \max(f(x)) \text{ and } \hat{x} \in (a,b).$$

The constant can be factored out of the integral, while the second order term takes the form of a Gaussian integral, which has an associated formula.

Theorem 3.1. Gaussian Integral
The integral of an arbitrary Gaussian function is

$$\int_{-\infty}^{\infty} e^{-a(x+b)^2}\,dx = \sqrt{\frac{\pi}{a}}.$$

The integration limits here are $\pm\infty$ and this is the only case which will actually show up in this thesis. As an aside however, as long as the maxima is sufficiently dominant and interior to the integration bounds only a minor error is introduced even if the integral's limits are finite.

The remaining integral from the expansion is

$$\int_a^b e^{-M\frac{1}{2}f''(\hat{x})(x-\hat{x})^2}\,dx \approx \sqrt{\frac{2\pi}{M|f''(\hat{x})|}}.$$

By including the constant which was factored out of the integral previously the expression for Laplace's method is obtained.

Theorem 3.2. Laplace's Method[6]
If M is a large number, $f(\hat{x})$ is a dominant maxima to the twice differentiable function $f(x)$ and $\hat{x} \in (a,b)$, then

$$\int_a^b e^{Mf(x)}\,dx \approx e^{Mf(\hat{x})}\sqrt{\frac{2\pi}{M|f''(\hat{x})|}}.$$

When $M \to \infty$ the approximation becomes an equality, which means this is an asymptotic expansion.

This may seem abstract, so an example is included for clarification.

Example 3.3. Consider the integral

$$\int_{-\infty}^{\infty} e^{Mf(x)}\,dx, \quad \text{where } f(x) = 1 - \log(1+x^2).$$

The first two derivatives of $f(x)$ are

$$f'(x) = -\frac{2x}{1+x^2},$$
$$f''(x) = -\frac{2(1-x^2)}{(1+x^2)^2},$$

and its maximum is located at $x = 0$. The Taylor expansion of $f(x)$ around its maximum is thus

$$\begin{aligned} f(x) &= f(0) + f'(0)x + f''(0)\frac{x^2}{2} + \mathcal{O}(x^3) \\ &= 1 - \log(1+0^2) - \frac{2\cdot 0}{1+0^2}x - \frac{2(1-0^2)}{(1+0^2)^2}\frac{x^2}{2} + \mathcal{O}(x^3) \\ &= 1 - x^2 + \mathcal{O}(x^3). \end{aligned}$$

This second order Taylor series will be named $g(x) = 1 - x^2$.

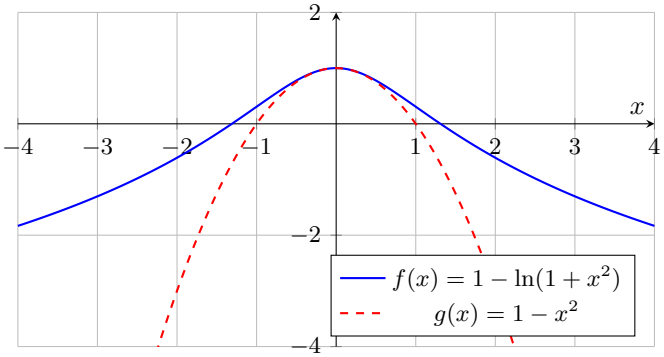

Figure 3.1: $f(x)$ vs $g(x)$

Comparing $f(x)$ and $g(x)$ in figure 3.1 shows a fairly standard result when comparing a function to a low order Taylor expansion. It is locally accurate around the point of expansion, but the functions quickly depart from each other outside of this region.

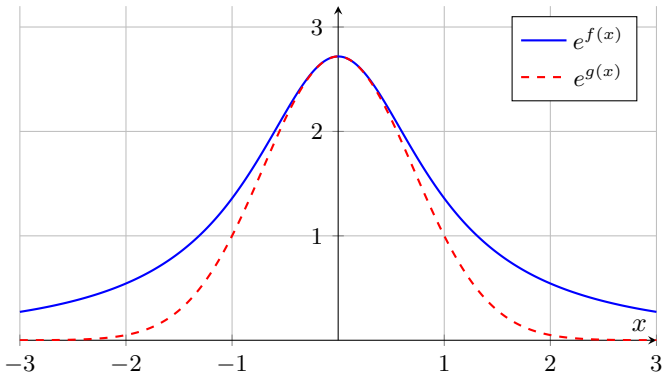

Figure 3.2: $e^{f(x)}$ vs $e^{g(x)}$

Next consider what happens when the exponential functions of $f(x)$ and $g(x)$ are compared, as in figure 3.2. The long term trends now align as both exponentials will tend to 0 as $x \to \pm\infty$. The 2 functions match each other a bit better in this example compared to last, but the precision is by no means impressive.

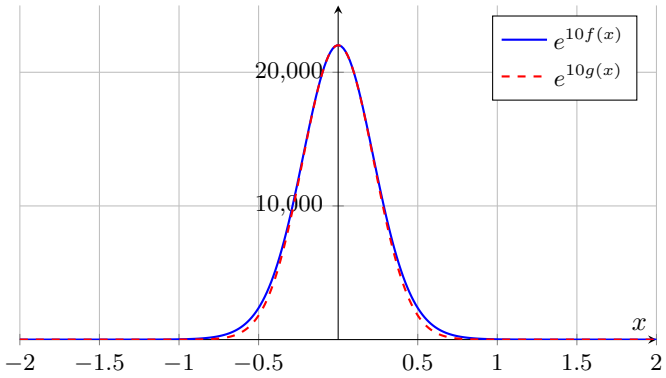

Figure 3.3: $e^{10 \cdot f(x)}$ vs $e^{10 \cdot g(x)}$

In figure 3.3 the factor $M = 10$ is added, causing the functions to explode in magnitude around the maxima at $x = 0$. The functions now overlap to such an extent that it becomes difficult to tell them apart. The second order Taylor expansion is still only accurate in the region surrounding the expansion point, what has changed is that this region has become the only region which is significant when evaluating the integral. Comparing the value of each integral gives

$$\int_{-\infty}^{\infty} e^{1\cdot f(x)}\,dx \approx 8.5397$$
$$\int_{-\infty}^{\infty} e^{1\cdot g(x)}\,dx \approx 4.8180$$
$$\int_{-\infty}^{\infty} e^{10\cdot f(x)}\,dx \approx 12834$$
$$\int_{-\infty}^{\infty} e^{10\cdot g(x)}\,dx \approx 12346$$

which means that the relative error is roughly 0.44 for $M = 1$ and 0.038 for $M = 10$, a stark difference.

This subsection will be concluded with one simple but important remark,

> $M_1 \cdot f_1(x)$ with $f_1(x) = 1 - \log(1 + x^2)$ and $M_1 = 10$,
> is the exact same expression as
> $M_2 \cdot f_2(x)$ with $f_2(x) = 10(1 - \log(1 + x^2))$ and $M_2 = 1$.

This means that a factor M in the exponent is not actually necessary. An alternate but sufficient condition is that $e^{f(\hat{x})} \gg e^{f(x)}$ for all x which are not equal to (or almost equal to) $\hat{x}$. In this thesis no factor M will be present, instead this stricter demand on the dominance of the maxima will be relied upon.

3.2 Complex Analysis Fundamentals

On the real number line $\mathbb{R}$, an integral from $x = a$ to $x = b$ is unambiguous as there exists only one path between these two points. On the complex plane $\mathbb{C}$ however there exists an infinite number of possible paths connecting two points $z = a+ib$ and $z = c+id$. Some path between these points must be chosen in order to evaluate an integral from one point to the other. In multivariable analysis this is called a line integral, but in complex analysis this is more commonly known as a contour integral. Quite a few common complex analysis concepts will introduced to define such an integral, most of the properties are technical and are stated mostly to exclude problematic edge cases.

Definition 3.4. Curve and Contour[5]
A curve γ is a subset of the complex plane $\mathbb{C}$ which is parametrized by a continuous function z(t) from a closed interval $t \in [a, b] \in \mathbb{R}$ to the complex plane $\mathbb{C}$.
A curve is smooth if it has a parametrization with a non-vanishing continuous derivative and it is injective, with a possible exception of the start/end points which may overlap.
A curve is directed if it has a specified direction of traversal.
A contour Γ is a finite sequence of directed smooth curves γ_n where the following curve in the sequence begins where the previous ended $\gamma_n(b_n) = \gamma_{n+1}(a_{n+1})$.

Since the definition is rather technical, it is worth stating that a contour is simply some path which connects the starting point and the end point in a way that preserves important properties.

Definition 3.5. Contour Integral[5]
A contour integral is an integral

$$\int_\Gamma f(z)\,dz := \int_a^b f(z(t))z'(t)\,dt$$

where $z(t)$ is a parametrization of the contour Γ and $\{a, b\} \in \mathbb{R}$ are the bounds of the domain of t.

This disambiguates the concept of complex valued integration limits by mapping them to real valued ones. Note that the upper and lower case gamma symbols here are completely unrelated to the previously discussed gamma functions and gamma distribution, this is merely an unfortunate consequence of overlapping notation conventions.

The next classic theorem from complex analysis which will be relied upon is the Cauchy-Riemann equations.

Theorem 3.6. Cauchy-Riemann Equations[5]
A function $f(z) = f(x + iy) = u(x, y) + iv(x, y)$ with $\{u(x, y), v(x, y)\} \in \mathbb{R}$ is holomorphic in a point iff u and v have well defined continuous partial derivatives and the Cauchy-Riemann equations

$$\frac{du}{dx} = \frac{dv}{dy}, \quad \frac{du}{dy} = -\frac{dv}{dx}$$

hold true.

The word holomorphic is unique to complex analysis, but is equivalent to a function being analytic on the complex plane. Holomorphic functions have many useful properties, but the one which is important in this context is that

if the domain is *simply connected*, that is to say has no holes in it, then the contour integral is path independent.

Theorem 3.7. Holomorphic Path Independence[5]
For a function $f(z)$ which is holomorphic on the simply connected domain U, the contour integrals over contours Γ_1 and Γ_2 are equal

$$\int_{\Gamma_1} f(z)\,dz = \int_{\Gamma_2} f(z)\,dz$$

if $\{\Gamma_1, \Gamma_2\} \in U$ and they have the same start and end points: $\Gamma_1(a) = \Gamma_2(a)$ and $\Gamma_1(b) = \Gamma_2(b)$.

In this context, the main reason why a domain might not be simply connected is due to the presence of an interior singularity. In simplified terms this all means that for any contour integral of a holomorphic function the path between the points can be modified in any way, as long as no poles are crossed, without the value of the integral changing.

One important property of holomorphic functions is that they have no internal local or global maximum or minimum.

Theorem 3.8. Maximum Modulus Principle[5]
For a function $f(z)$ which is holomorphic on the domain U there exists no interior point z_0 in U for which $|f(z_0)| \geq |f(z)|$ for all z in any neighbourhood of z_0 unless $f(z)$ is constant on the domain.

This means that if $f(z)$ is non-constant and holomorphic in z_0 as well as its neighbourhood, and z_0 is a stationary point, then z_0 must be saddle point. With this theory covered Laplace's method can now be extended to the complex plane.

3.3 Path of Steepest Descent

With Laplace's method it is possible to approximate an integral

$$\int_a^b e^{f(x)}\,dx$$

by Taylor expanding $f(x)$ around the maxima. The objective is to do something similar in the complex plane for a contour integral

$$\int_\Gamma e^{f(z)}\,dz$$

despite the fact that there exists no local or global maximum which is interior to the domain where $f(z)$ is holomorphic. The goal is instead to choose a contour Γ such that the integral resembles the $\mathbb{R}$ case along this contour. In the same way that Laplace's method assumes there to exist a uniquely dominant maxima,

the method of steepest descent assumes there exists a uniquely dominant saddle point. The method is called steepest descent because the contour is defined as the direction away from the saddlepoint in which the amplitude decreases fastest. A saddle point will have two such directions, each eventually heading towards the respective integration limits. An illustration will hopefully clarify what is meant by this.

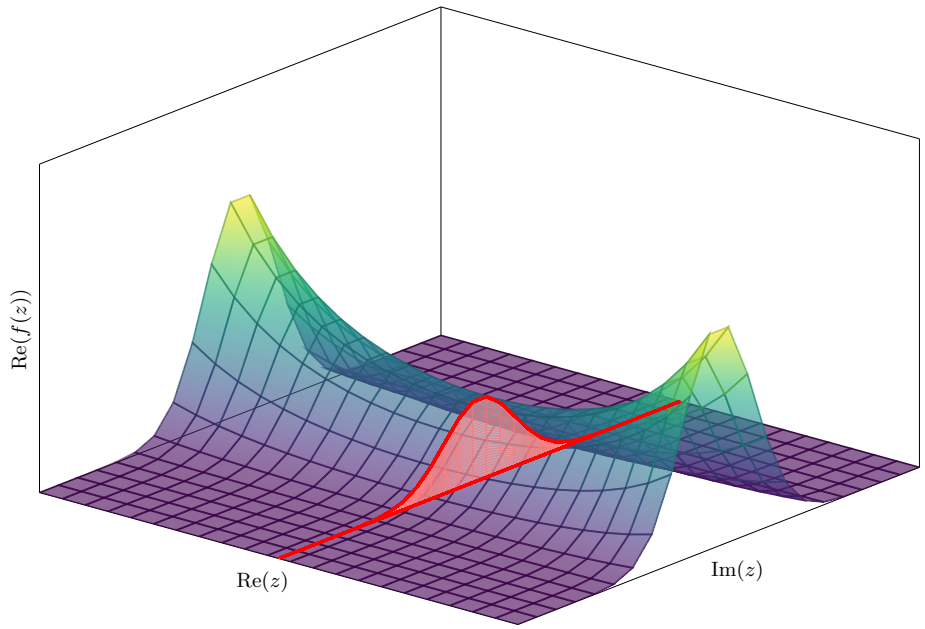

Figure 3.4: Contour tracing the path of steepest descent through a saddle point

Note that plotting $\mathbb{C} \to \mathbb{C}$ functions is inherently limited by the fact that 4 dimensions are involved and the geometry can be complicated. Nevertheless figure 3.4 illustrates some important points, namely that the saddle point constitutes the maximum amplitude along the path of integration, despite this point not being a maximum in the plane. The light red area below the bright red contour is the area which is being integrated and has begun to resemble the example in figure 3.3 from the Laplace's method section.

Note that the phrase "steepest descent" is from the perspective of viewing the contour from the saddle point, following the contour from start to finish the function indeed ascends rapidly before reaching the saddle point.

A key consequence of the Cauchy-Riemann equations is that all holomorphic functions have real and imaginary components which grow most rapidly in orthogonal directions, which can be shown through

$$\nabla u \cdot \nabla v = \frac{du}{dx}\frac{dv}{dx} + \frac{du}{dy}\frac{dv}{dy} = 0 \implies \nabla u \perp \nabla v.$$

For a complex exponential function $e^{f(z)}$, splitting the real and imaginary components

$$e^{f(z)} = e^{f(x+yi)} = e^{u(x,y)+iv(x,y)} = e^{u(x,y)}(\cos(v(x,y)) + i\sin(v(x,y)))$$

reveals that the real portion $u(x,y)$ controls the amplitude and the imaginary portion $iv(x,y)$ controls the phase. In order to achieve the steepest descent the contour should be chosen such that the amplitude decays as fast as possible $\nabla u_{\max}$. Since $\nabla u \perp \nabla v$ this is achieved precisely when $\nabla v = \mathbf{0}$ which in turn means that $v(x,y)$ must be constant. The path where this holds true is the path of steepest descent.

Theorem 3.9. Path of Steepest Descent[10]
For a contour integral

$$\int_\Gamma e^{f(z)}\,dz = \int_\Gamma e^{u(x,y)+iv(x,y)}\,dz$$

with contour $\Gamma : [a,b] \to [\Gamma(a), \Gamma(b)]$ and with the property that the imaginary portion of the function $f(z)$ is equal to some constant c at both integration bounds

$$\mathrm{Im}(f(\Gamma(a))) = \mathrm{Im}(f(\Gamma(b))) = c,$$

The path of steepest descent is the contour Γ for which the imaginary component remains constant $f(\Gamma(t)) = \mathrm{Re}(f(\Gamma(t))) + c$ for every point $t \in [a,b]$. Along this path

$$\int_\Gamma e^{f(z)}\,dz = \int_\Gamma e^{u(x,y)+ic}\,dz.$$

Now that the path of steepest descent has been defined, the saddle point will be studied in order to understand the local behaviour of the path in this crucial point.

3.4 Expansion around Saddle Point

Just as in the real valued case the stationary point is identified through the property that the function's first derivatives are zero at that location, $f'(z) = 0$. In this thesis the convention is that every parameter which is evaluated in the saddle point is noted with a hat, so $f(\hat{z})$ is the function evaluated in the saddle point.

In the general case there is nothing which states that only one saddle point can exist. If one saddle point is far more dominant than all others it is possible to simply base the approximation on the most dominant saddle point. If several saddle points are roughly equally dominant the standard approach outlined in this section fails and more complicated solutions are required. The only case which is considered henceforth is that where only one dominant saddle point needs to be considered. Moreover it is assumed that this saddle point is simple, such that

$$f'(\hat{z}) = 0, \quad f''(\hat{z}) \neq 0.$$

The function $f(z)$ is Taylor expanded around the saddle point in the exact same way as in the one-dimensional case, keeping in mind that the first order terms are 0 there and third order terms and higher are disregarded

$$f(z) \approx f(\hat{z}) + \frac{1}{2} f''(\hat{z})(z - \hat{z})^2.$$

The second order terms can be rewritten in exponential form through

$$f''(\hat{z}) = |f''(\hat{z})|e^{i\theta}, \quad z - \hat{z} = |z - \hat{z}|e^{i\psi}$$

which will assist in identifying the path of steepest descent in the saddlepoint. This rewritten form results in

$$\begin{aligned} f(z) &\approx f(\hat{z}) + \frac{1}{2}|f''(\hat{z})|e^{i\theta}(|z - \hat{z}|e^{i\psi})^2 \\ &= f(\hat{z}) + \frac{1}{2}|f''(\hat{z})||z - \hat{z}|^2 e^{i(\theta + 2\psi)} \\ &= f(\hat{z}) + \frac{1}{2}|f''(\hat{z})||z - \hat{z}|^2(\cos(\theta + 2\psi) + i\sin(\theta + 2\psi)) \end{aligned}$$

Note that θ is a constant while ψ depends on z. Since the path of steepest descent is defined by keeping the imaginary part constant, and since the $f(\hat{z})$ term already contains this constant, the second order component must be exclusively real, which means that

$$\begin{aligned} &\sin(\theta + 2\psi) = 0 \\ &\theta + 2\psi = k\pi, \quad k \in \mathbb{Z} \\ &\psi = \frac{k\pi - \theta}{2}, \quad k \in \mathbb{Z} \end{aligned}$$

This expression has four solutions in any 2π radian branch of ψ, such as

$$\begin{cases} \psi_0 = \frac{-\theta}{2} \\ \psi_1 = \frac{\pi - \theta}{2} \\ \psi_2 = \frac{2\pi - \theta}{2} \\ \psi_3 = \frac{3\pi - \theta}{2}. \end{cases}$$

Setting these values into the cosine function gives

$$\begin{cases} \cos(\theta + 2\psi_0) = \cos(0) = 1 \\ \cos(\theta + 2\psi_1) = \cos(\pi) = -1 \\ \cos(\theta + 2\psi_2) = \cos(2\pi) = 1 \\ \cos(\theta + 2\psi_3) = \cos(3\pi) = -1. \end{cases}$$

These are the four different directions leading away from the saddle point which will only affect amplitude and not phase. Since the choice of angle here decides

the sign of the second derivative, this decides whether parabola will be concave or convex. The saddle point is meant to be the maxima along the path of integration which means that the parabola should be concave and thus the sign negative. The positive cosine directions then are the paths of steepest ascent in the saddle point, which is the opposite of what is desired. The path of steepest descent in the neighbourhood of the saddle point then will be the line passing through $|z-\hat{z}|e^{i\psi_1}$ and $|z-\hat{z}|e^{i\psi_3}$.

In order to determine a local parametrization of $z(t)$ with this property, set

$$\begin{aligned} z(t) &= z(\hat{t}) + t\exp i\psi_1 \\ &= z(\hat{t}) + t\exp\left(\frac{i}{2}(\pi-\theta)\right) \end{aligned}$$

where $\theta = \arg(f''(\hat{z}))$. This is the local parametrization of the contour which gives steepest descent around the saddle point. With this information it is now possible to expand the contour integral around the saddle point. This expansion is

$$\begin{aligned} &\int_\Gamma \exp(f(z))\,dz \\ =&\int_a^b \exp(f(z(t)))z'(t)\,dt \\ \approx&\int_a^b \exp\left(f(\hat{z}) + \frac{1}{2}f''(\hat{z})(z(t)-\hat{z})^2\right)z'(t)\,dt \\ =&\int_a^b \exp\left(f(\hat{z}) + \frac{1}{2}f''(\hat{z})\left(\hat{z} + t\exp\left(\frac{i}{2}(\pi-\theta)\right) - \hat{z}\right)^2 + \frac{i}{2}(\pi-\theta)\right)\,dt \\ =&\exp\left(f(\hat{z}) + \frac{i}{2}(\pi-\theta)\right)\int_a^b \exp\left(\frac{1}{2}f''(\hat{z})\left(t^2\exp\left(i(\pi-\theta)\right)\right)\right)\,dt \\ =&\frac{i}{\sqrt{\exp(i\theta)}}\exp(f(\hat{z}))\int_a^b \exp\left(-\frac{1}{2}|f''(\hat{z})|t^2\right)\,dt \\ &\text{the remaining integral is a real valued Gaussian} \\ =&\frac{i}{\sqrt{\exp(i\theta)}}\exp(f(\hat{z}))\sqrt{\frac{2\pi}{|f''(\hat{z})|}} \\ =&\exp(f(\hat{z}))\sqrt{\frac{2\pi}{-f''(\hat{z})}} \end{aligned}$$

which constitutes the standard formula for this method.

Theorem 3.10. Method of Steepest Descent[10]
A holomorphic function $f(z)$ with a single simple dominant saddle point $\hat{z}$, such that $\nabla f(\hat{z}) = \mathbf{0}$ and a contour Γ passing through $\hat{z}$ along the path of steepest descent allows for the approximation

$$\int_\Gamma e^{f(z)}\,dz \approx e^{f(\hat{z})}\sqrt{\frac{2\pi}{-f''(\hat{z})}}.$$

Extending Laplace's method to the complex plane requires significant and delicate analysis work but the resulting formula is near identical to real axis version. The key differences are that it has a negative denominator and much stricter requirements. The above formula will however be immensely useful in approximating the inverse Laplace transform, which is the key to deriving the SPA.

4 Deriving the Saddle Point Approximation

With the theoretical framework laid out in the previous section, deriving the SPA will be straight forward. After that is done, some extra work will be needed to integrate the SPA into LR which is the main topic of this section. Some important criteria and useful examples will also be discussed in this section.

Recall that the MGF can be calculated using the two-sided, sign inverted, Laplace transform of the PDF.

Definition 4.1. Two-sided Laplace Transform

$$\mathcal{L}\left\{f(x)\right\}(s) := \int_{-\infty}^{\infty} e^{-sx} f(x)\,dx$$

Theorem 4.2. MGF Transform[2]
The MGF $M(s)$ for the distribution with PDF $f(x)$ is

$$M(s) = \mathcal{L}\left\{f(x)\right\}(-s) = \int_{-\infty}^{\infty} e^{sx} f(x)\,dx$$

if the integral always converges in some neighbourhood of $s = 0$.

For the remainder of this thesis, the Laplace transform will always refer to this two-sided and sign inverted version, even if not explicitly stated. In order to transform the MGF to a PDF, the inverse Laplace transform, also known as the Bromwich integral, is required.

Definition 4.3. Inverse Laplace Transform[2]

$$\mathcal{L}^{-1}\left\{g(s)\right\}(x) := \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{sx} g(s)\,ds$$

Thus the PDF can be gained from the MGF through

$$\begin{aligned} f(x) &= \mathcal{L}^{-1}\{(\mathcal{L}\{f(x)\}\{-s\})\}\{-x\} \\ &= \mathcal{L}^{-1}\{M(s)\}\{-x\} \\ &= \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{-sx}M(s)\,ds \\ &= \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{-sx}e^{\log(M(s))}\,ds \\ &= \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{\log(M(s))-sx}\,ds \\ &\text{since the CGF } K(s) = \log(M(s)) \\ &= \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{K(s)-sx}\,ds. \end{aligned}$$

Daniels'[2] innovation is applying the method of steepest descent to this problem. Note that the complex exponential function is *entire*, meaning that it is holomorphic everywhere. If the exponent is well defined no risk of singularities exist, which might otherwise disrupt the path of steepest descent. For convenience the exponent is named $h(s) = K(s) - sx$. The first two derivatives of this function are

$$h'(s) = K'(s) - x, \quad h''(s) = K''(s).$$

The saddle point $\hat{s}$ is identified by

$$\begin{aligned} h'(s) = K'(s) - x &= 0 \\ K'(\hat{s}) &= x. \end{aligned}$$

Inserting these values into the method of steepest descent formula from the previous section gives

$$\begin{aligned} f(x) &= \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{K(s)-sx}\,ds \\ &= \frac{1}{2\pi i}e^{K(\hat{s})-\hat{s}x}\sqrt{\frac{2\pi}{-K''(\hat{s})}} \\ &= \frac{e^{K(\hat{s})-\hat{s}x}}{\sqrt{2\pi K''(\hat{s})}}. \end{aligned}$$

This is the saddle point approximation.

Theorem 4.4. Saddle Point Approximation[2]
The saddle point approximation of the PDF for a distribution with a twice differentiable CGF $K(s)$ and saddle point $\hat{s}$ solving to $K'(\hat{s}) = x$ is

$$f(x) = \frac{e^{K(\hat{s})-\hat{s}x}}{\sqrt{2\pi K''(\hat{s})}}.$$

As noted, the difficult work here is done by the method of steepest descent, once the formula is applied the derivation is mostly trivial. In the next subsection the requirements for a well defined MGF will be discussed.

4.1 Cramér's Condition

Convergent Laplace-/inverted Laplace transforms don't always exist, thus a brief discussion surrounding the necessary conditions is warranted. It is not necessary for the Laplace transform

$$\mathcal{L}\{f(x)\}(s) = \int_{-\infty}^{\infty} e^{-sx} f(x)\, dx$$

to be convergent for all s in order for the MGF to be well defined. The important region will instead turn out to be the neighbourhood of 0. An intuition of this can be gained by considering the fact that a key property of the MGF is its ability to produce all raw moments m_n through

$$M^{(n)}(0) = m_n.$$

This property becomes clear by considering the Maclaurin series

$$\begin{aligned} M(s) &= \mathrm{E}[e^{sX}] \\ &= 1 + s\,\mathrm{E}[X] + \frac{s^2\,\mathrm{E}[X^2]}{2!} + \frac{s^3\,\mathrm{E}[X^3]}{3!} + \cdots + \frac{s^n\,\mathrm{E}[X^n]}{n!} + \cdots \\ &= 1 + sm_1 + \frac{s^2 m_2}{2!} + \frac{s^3 m_3}{3!} + \cdots + \frac{s^n m_n}{n!} + \cdots . \end{aligned}$$

It is the fact that this information is encoded in the MGF which means that it uniquely determines a distribution. A Maclaurin series in general does not only require a function to be well defined in the point 0, it must also be well defined in a neighbourhood of 0. This is not a definitive proof of the importance of this neighbourhood for the properties of a MGF, but is only intended to give an intuition.

The criteria that the MGF must be always finite in some neighbourhood of 0 is sometimes known as Cramér's condition after Harald Cramér.

Theorem 4.5. Cramér's Condition[1]
An MGF is well defined iff

$$\exists h > 0 \text{ such that } M(s) = \mathbb{E}[e^{sX}] < \infty, \quad \forall s \in (-h, h).$$

A distribution which does not satisfy Cramér's condition is known as heavy-tailed. By investigating the convergence criteria for the MGF it will be possible to describe in further detail which properties the PDF must have in order to fulfil Cramér's condition. Consider then the convergence for

$$\mathbb{E}[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f(x)\, dx.$$

First the special case $s = 0$ is handled,

$$\int_{-\infty}^{\infty} e^{0x} f(x)\, dx = \int_{-\infty}^{\infty} f(x)\, dx = 1, \quad \text{by law of total probability.}$$

Proceeding with cases where $s \neq 0$, the two tails are handled separately,

$$\int_{-\infty}^{\infty} e^{sx} f(x)\, dx = \int_{-\infty}^{0} e^{sx} f(x)\, dx + \int_{0}^{\infty} e^{sx} f(x)\, dx$$

The first of these integrals will be well behaved if $s > 0$ as $e^{-sx} \to 0$ rapidly. For the same reason the second integral will be well behaved if $s < 0$. The problematic integral however diverges unless the PDF decays faster than e^{sx} grows. The criteria once discovered will through symmetry apply to whichever integral risks divergence. The condition for convergence is that $\exists s > 0$ such that

$$\int_{0}^{\infty} e^{sx} f(x)\, dx < \infty.$$

Assume that $f(x) \leq Ce^{-ax}$ for some constants $C > 0$ and $a > 0$ on $x \geq 0$, then

$$\int_{0}^{\infty} e^{sx} f(x)\, dx \leq \int_{0}^{\infty} e^{sx} Ce^{-ax}\, dx$$
$$\int_{0}^{\infty} e^{sx} f(x)\, dx \leq C \int_{0}^{\infty} e^{(s-a)x}\, dx$$

which converges iff $a > s$. This only has to hold for some $s > 0$, and if $a > 0$ there must exist some value s such that $a > s > 0$.

Corollary 4.6. Criterion for Cramér's Condition
If the PDF $f(x) \leq Ce^{-ax}$ for some $C > 0$, $a > 0$, then Cramér's condition is met and the MGF exists.

The inverse Laplace transform

$$\frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{K(s) - sx}\, ds$$

also requires certain conditions in order to converge. Since the saddle point is real valued and since the path of steepest descent keeps the imaginary part

constant, after correct parametrization the contour will in fact be real valued at every point along the path of integration.

The strip of convergence for this integral is therefore the domain where the real component allows for a convergent MGF. In other words, this condition is the same as the previous condition, meaning that as long as Cramér's condition is fulfilled the inverse Laplace transform adds no additional constraints. It is quite intuitive that the conditions which allow a Laplace transform to turn a PDF into a well defined MGF are the same conditions which allow an inverse Laplace transform to turn an MGF into a well defined PDF.

In summary then, a SPA is only possible if the PDF has at least an exponential decay rate, anything slower will lead to divergent integrals.

4.2 Identification of Saddle Point

This brief subsection will discuss some consequences stemming from how saddle points are identified. The equation used to find the saddle point, $K'(\hat{s}) = x$ is somewhat backwards. It would be preferable to define it explicitly in terms of x, such that some function $f(x) = \hat{s}$ since x is used to find the saddle point and not the other way around. The relationship between x and $K'(\hat{s})$ is one-to-one, meaning they form a bijection and thus the function $K'(s)$ must be invertible at this point. However, the fact that an inverse function exists does not guarantee that the inverse function can be written as a closed form expression. This raises the question, when can the saddle point equation be inverted in order to provide a closed form expression?

The answer is that unfortunately it is difficult to determine ahead of time which functions have closed form inversions, some classes of functions like linear and quadratic equations are naturally easy to invert as well as most expressions containing a single elementary function. Third and fourth degree polynomials are more complicated to solve, but have associated formulas, while fifth degree or above are proved to sometimes be impossible. Expressions containing a mix of different functions are in general difficult to invert, such as an expression containing both logarithms and polynomials. When it is not possible to invert explicitly, the saddle point must instead be identified with a root finding algorithm, such as Newton's method. This is naturally more computationally demanding than evaluating a closed form expression.

One reason why the Gamma distribution is well suited for SPA and LR is that its CGF's first derivative is easily invertible

$$K'_\Gamma(\hat{s}) = \frac{\alpha}{\beta - \hat{s}} = x \implies \hat{s} = \beta - \frac{\alpha}{x}.$$

This permits substituting all $\hat{s}$ expressions using the identity above. In summary, the saddle point is not always easy to find, and the efficiency of a SPA depends on how challenging this is.

4.3 Deriving Stirling's Approximation

This subsection is somewhat tangential to the derivation of the approximations and is not required in order to understand the following steps. It is however a useful example as it both shows how the gamma SPA is calculated and reproduces the most famous approximation of the complete gamma function, Stirling's approximation.

Example 4.7. The SPA requires the CGF and its first two derivatives, for the gamma distribution these are

$$K_\Gamma(s) = -\alpha \log\left(1 - \frac{s}{\beta}\right), \quad K'_\Gamma(s) = \frac{\alpha}{\beta - s}, \quad K''_\Gamma(s) = \frac{\alpha}{(\beta - s)^2}.$$

The saddle point is, as seen in the previous subsection, identified through

$$K'_\Gamma(\hat{s}) = \frac{\alpha}{\beta - \hat{s}} = x \implies \hat{s} = \beta - \frac{\alpha}{x}.$$

The s-variable in the CGF and its second derivative can be substituted for the value of the saddle point, giving

$$K_\Gamma(\hat{s}) = -\alpha \log\left(\frac{\alpha}{\beta x}\right), \quad K''_\Gamma(\hat{s}) = \frac{x^2}{\alpha}.$$

Inserting these values into the SPA formula results in

$$f(x) \approx e^{-\alpha \log\left(\frac{\alpha}{\beta x}\right) - \beta x + \alpha} \frac{1}{\sqrt{2\pi \frac{x^2}{\alpha}}}.$$

This concludes deriving the gamma SPA. Now by setting the approximation to be equivalent to the exact gamma PDF and solving for $\Gamma(\alpha)$ will result in

$$\begin{aligned}
\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} &= e^{-\alpha \log\left(\frac{\alpha}{\beta x}\right) - \beta x + \alpha} \frac{1}{\sqrt{2\pi \frac{x^2}{\alpha}}} \\
\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} &= \left(\frac{\beta x}{\alpha}\right)^\alpha e^\alpha \frac{\sqrt{\alpha}}{x\sqrt{2\pi}} \\
\frac{1}{\Gamma(\alpha)} &= \left(\frac{e}{\alpha}\right)^\alpha \frac{\sqrt{\alpha}}{\sqrt{2\pi}} \\
\Gamma(\alpha) &= \left(\frac{\alpha}{e}\right)^\alpha \sqrt{\frac{2\pi}{\alpha}}.
\end{aligned}$$

where the final row is now the Stirling approximation.

This example connects the SPA to a well known approximation and shows what this thesis is aiming to achieve through these approximations, namely replacing special functions with elementary functions through SPA/LR.

4.4 Lugannani-Rice Approximation

Next the Lugannani-Rice approximation will finally be derived. In their original paper Lugannani and Rice[3] define their approximation using an asymptotic series with implicitly defined terms of rapidly increasing intricacy. In this section the simpler, but more common, first order approximation will be derived in a similar way as in Butler[6] (Chapter 2.3).

Accurately integrating the SPA

$$F(x) = \int_{-\infty}^{x} \frac{e^{K(\hat{s})-\hat{s}t}}{\sqrt{2\pi K''(\hat{s})}}\, dt$$

is deceptively difficult, as $\hat{s}$ has a complicated dependency on the integration variable t. Attempting to integrate over t before performing the inverse Laplace transform is easy, but creates a pole which invalidates the method of steepest descent. For these reasons the integral will be approximated using the same method as Butler[6] uses, the confusingly named Temme approximation. This is not the same 1979 Temme approximation which coincides with the LR gamma approximation, this is a 1982 technique used for approximating an integral of form

$$\int_{-\infty}^{x} \phi(t) f(t)\, dt.$$

In order to differentiate this method from the gamma approximation, this method will be called "Temme's integral approximation". This method will only be used in this specific section, any other references in the text to Temme will refer to the 1979 gamma approximation. A truncated derivation of Temme's integral approximation will now be demonstrated.

Definition 4.8. The standard normal $\mathcal{N}(0,1)$ distribution is the probability distribution with

$$\begin{array}{cc} \text{PDF:} & \text{CDF:} \\ \phi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2} & \Phi(x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \end{array}$$

Let $\phi(x)$ be a standard normal PDF, this function has the property

$$\phi'(x) = -\phi(x)x.$$

Temme's integral approximation can be derived by utilizing this property in combination with integration by parts

$$
\begin{aligned}
&\int_{-\infty}^{x} \phi(t) f(t)\, dt \\
=&\int_{-\infty}^{x} \phi(t) \left(f(t) - f(0) + f(0)\right)\, dt \\
=&\int_{-\infty}^{x} \phi(t) f(0)\, dt + \int_{-\infty}^{x} \phi(t) \left(f(t) - f(0)\right)\, dt \\
=&\left[\Phi(t) f(0)\right]_{-\infty}^{x} - \int_{-\infty}^{x} \phi'(t) \frac{f(t) - f(0)}{t}\, dt \\
=&\Phi(x) f(0) - \left[\phi(t) \frac{f(t) - f(0)}{t}\right]_{-\infty}^{x} + \int_{-\infty}^{x} \phi(t) \frac{f'(t)t - f(t) + f(0)}{t^2}\, dt, \\
\approx&\Phi(x) f(0) + \phi(x) \left(\frac{f(0)}{x} - \frac{f(x)}{x}\right).
\end{aligned}
$$

Which is Temme's integral approximation. The remaining integral was dropped entirely in the last step, as it is small. A proof for this is beyond the scope of this thesis, but it can be somewhat justified by the fact that both $\phi(t)$ and $\frac{1}{t^2}$ will suppress the expression for large $|t|$ and $f'(t)t - f(t) + f(0) \approx 0$ when $t \approx 0$ because $f'(t)t + f(0)$ is similar to the first two terms of the Maclaurin series of $f(t)$.

Theorem 4.9. Temme's Integral Approximation[6]

$$
\int_{-\infty}^{x} \phi(t) f(t)\, dt \approx \Phi(x) f(0) + \phi(x) \left(\frac{f(0)}{x} - \frac{f(x)}{x}\right)
$$

where $\phi(x)$ is the standard normal PDF and $\Phi(x)$ is the standard normal CDF.

Next the SPA will be rearranged and $f(0)$ calculated so that this theorem may be utilized. First a variable substitution in the integral

$$
F(x) = \int_{-\infty}^{x} \frac{e^{K(\hat{s}) - \hat{s}t}}{\sqrt{2\pi K''(\hat{s})}}\, dt
$$

of the SPA is required. Introduce

$$
\begin{aligned}
w(t) &= \operatorname{sgn}(\hat{s}) \sqrt{2(\hat{s}t - K(\hat{s}))} \\
-\frac{w(t)^2}{2} &= K(\hat{s}) - \hat{s}t \\
t(w) &= \frac{w^2}{2\hat{s}} - \frac{K(\hat{s})}{\hat{s}} \\
dt &= \frac{w}{\hat{s}}\, dw
\end{aligned}
$$

and begin integration by substitution

$$\begin{aligned}\int_{-\infty}^{x} \frac{e^{K(\hat{s})-\hat{s}t}}{\sqrt{2\pi K''(\hat{s})}}\,dt &= \int_{w(-\infty)}^{w(x)} \frac{e^{-\frac{w^2}{2}}}{\sqrt{2\pi K''(\hat{s})}}\,\frac{w}{\hat{s}}\,dw \\ &= \int_{-\infty}^{w(x)} \phi(w)\frac{w}{\hat{s}\sqrt{K''(\hat{s})}}dw \\ &= \int_{-\infty}^{w(x)} \phi(w)f(w)\,dw.\end{aligned}$$

Next $f(0)$ must be calculated, note that inserting $w = 0$ into the expression is insufficient since $\hat{s}$ is now dependent on w. If $w = 0$ then $\hat{s}x - K(\hat{s}) = 0$ which can be solved by $\hat{s} = 0$ since $K(0) = \log(\mathbb{E}[e^{0X}] = \log(1) = 0$. Therefore

$$f(0) = \left.\frac{\operatorname{sgn}(\hat{s})\sqrt{2(\hat{s}t - K(\hat{s}))}}{\hat{s}\sqrt{K''(\hat{s})}}\right|_{s=0}$$

which creates a division by 0 issue, it will however be sufficient to calculate the value as $\lim_{s\to 0+}$. The CGF is expanded with a Maclaurin series

$$\begin{aligned}f(s) &= \frac{\operatorname{sgn}(\mathrm{s})\sqrt{2(st - K(0) - K'(0)s - K''(0)s^2\frac{1}{2} - \sum_{n=3}^{\infty} K^{(n)}(0)s^n\frac{1}{n!})}}{s\sqrt{K''(0) - \sum_{n=3}^{\infty}K^{(n)}(0)s^{n-2}\frac{1}{(n-2)!}}} \\ &= \sqrt{\frac{-2(st - 0 - st - K''(0)s^2\frac{1}{2} - \sum_{n=3}^{\infty} K^{(n)}(0)s^n\frac{1}{n!})}{K''(0)s^2 + \sum_{n=3}^{\infty} K^{(n)}(0)s^n\frac{1}{(n-2)!}}} \\ &= \sqrt{\frac{K''(0)s^2 + 2\sum_{n=3}^{\infty} K^{(n)}(0)s^n\frac{1}{n!}}{K''(0)s^2 + \sum_{n=3}^{\infty} K^{(n)}(0)s^n\frac{1}{(n-2)!}}} \\ &= \sqrt{\frac{K''(0)s^2(1 + \sum_{n=3}^{\infty} K^{(n)}(0)s^{n-2}\frac{2}{n!K''(0)})}{K''(0)s^2(1 + \sum_{n=3}^{\infty} K^{(n)}(0)s^{n-2}\frac{1}{K''(0)(n-2)!})}} \\ &= \sqrt{\frac{1 + \sum_{n=3}^{\infty} K^{(n)}(0)s^{n-2}\frac{2}{n!K''(0)}}{1 + \sum_{n=3}^{\infty} K^{(n)}(0)s^{n-2}\frac{1}{K''(0)(n-2)!}}}.\end{aligned}$$

Finally taking the limit

$$\lim_{s\to 0^+} f(s) = \sqrt{\frac{1}{1}} = 1$$

the pole is therefore removable and through continuity $f(0) := 1$. The Temme integral approximation can now be constructed

$$\begin{aligned}\int_{-\infty}^{w(x)} \phi(w)f(w)\,dw &\approx \Phi(w(x))f(0) + \phi(w(x))\left(\frac{f(0)}{w(x)} - \frac{f(w(x))}{w(x)}\right) \\ &= \Phi(w(x)) + \phi(w(x))\left(\frac{1}{w(x)} - \frac{\frac{w(x)}{\hat{s}\sqrt{K''(\hat{s})}}}{w(x)}\right) \\ &= \Phi(w(x)) + \phi(w(x))\left(\frac{1}{w(x)} - \frac{1}{\hat{s}\sqrt{K''(\hat{s})}}\right).\end{aligned}$$

Introducing the naming convention

$$\begin{aligned}\hat{w} &:= w(x) = \operatorname{sgn}(\hat{s})\sqrt{2(\hat{s}x - K(\hat{s}))} \\ \hat{u} &:= \hat{s}\sqrt{K''(\hat{s})}\end{aligned}$$

yields the standard first order LR expression

$$F(x) = \Phi(\hat{w}) + \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right), \quad \hat{w} \neq 0.$$

The specific case where $\hat{w} = 0$ corresponds to evaluating the CDF at the mean value of the distribution, which leads to catastrophic cancellation and division by 0 issues. The expression must therefore be expanded with additional terms around this point which then gives

$$F(\mu) = \frac{1}{2} + \frac{K'''(0)}{6\sqrt{2\pi}K''(0)^{\frac{3}{2}}}.$$

This concludes deriving the LR approximation.

Theorem 4.10. Lugannani-Rice Approximation[3]
The Lugannani-Rice approximation of the CDF $F(x)$ from CGF $K(s)$ is

$$F(x) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right) & \text{for } x \neq \mu \\ \frac{1}{2} + \frac{K'''(0)}{6\sqrt{2\pi}K''(0)^{3/2}} & \text{for } x = \mu \end{cases}$$

where $\hat{s}$ solves $K'(\hat{s}) = x$, $\hat{w} = \operatorname{sgn}(\hat{s})\sqrt{2(\hat{s}x - K(\hat{s}))}$, $\hat{u} = \hat{s}\sqrt{K''(\hat{s})}$, μ is the mean of distribution and Φ/ϕ is the standard normal $\mathcal{N}(0,1)$ CDF/PDF.

Before continuing to the empirical portion of the thesis an older simpler version of this approximation will briefly be explored.

4.5 Bahadur-Rao

A simplified version of LR can be derived through the following steps

$$\begin{aligned}
\Phi(\hat{w}) + \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right) &= \int_{-\infty}^{\hat{w}} \phi(t)dt + \frac{\phi(\hat{w})}{\hat{w}} - \frac{\phi(\hat{w})}{\hat{u}} \\
&= 1 - \int_{\hat{w}}^{\infty} \phi(t)dt + \frac{\phi(\hat{w})}{\hat{w}} - \frac{\phi(\hat{w})}{\hat{u}} \\
&= 1 - \int_{\hat{w}}^{\infty} -\frac{\phi'(t)}{t}dt + \frac{\phi(\hat{w})}{\hat{w}} - \frac{\phi(\hat{w})}{\hat{u}} \\
&= 1 - \left[-\frac{\phi(t)}{t}\right]_{\hat{w}}^{\infty} + \int_{\hat{w}}^{\infty} -\frac{\phi(t)}{t}dt + \frac{\phi(\hat{w})}{\hat{w}} - \frac{\phi(\hat{w})}{\hat{u}} \\
&= 1 - \frac{\phi(\hat{w})}{\hat{w}} + \frac{\phi(\hat{w})}{\hat{w}} - \frac{\phi(\hat{w})}{\hat{u}} - \int_{\hat{w}}^{\infty} \frac{\phi(t)}{t}dt \\
&= 1 - \frac{\phi(\hat{w})}{\hat{u}} - \int_{\hat{w}}^{\infty} \frac{\phi(t)}{t}dt.
\end{aligned}$$

For large $\hat{w}$ which occurs in the tails,

$$\int_{\hat{w}}^{\infty} \frac{\phi(t)}{t}dt \approx 0.$$

This gives the much simplified approximation

$$F(x) \approx 1 - \frac{\phi(\hat{w})}{\hat{u}} = 1 - \frac{e^{K(\hat{s})-\hat{s}x}}{\hat{s}\sqrt{2\pi|K''(\hat{s})|}}.$$

With just very minor alterations then it is possible to go from the SPA to an approximation of the CDF without having to evaluate the full LR expression. This formula was actually discovered by Bahadur and Rao[4], 20 years before the LR approximation was derived, though naturally through different methods. This approximation only captures the leading order decay rate without any corrections which makes it less accurate than LR, but it is still asymptotically accurate. This method will also be included in the computational efficiency analysis.

This concludes not only this section, but also the theoretical portion of the thesis. The following sections are instead intended to lay the foundation for an empirical comparison between the LR approximation of the gamma distribution and other approximation methods.

5 Gamma Approximations

Special function algorithms do not generally use one single method to approximate the gamma CDF, instead a variety of different methods are generally used depending on the parameter values. This section will explain how the various component methods work. The main topic of this thesis is evaluating the LR approximation, while these other methods are mainly references which LR will be benchmarked against. For this reason they will be covered with far

less depth than LR.

Since there exists no closed form solution to the gamma CDF $\frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)}$ some type of approximation is required. The naive straightforward approach is to numerically integrate the complete and incomplete gamma functions. This can be done with a step function or with optimized modern variants such as adaptive quadrature. This approach is however very computationally inefficient, which is why a plethora of other methods have been developed over time.

It is common to approximate the complete and incomplete gamma functions separately. Out of the two, the complete gamma function is easier to approximate and there exists methods which provide high accuracy for all parameters. The incomplete gamma function is the problematic component as there exists no method which is fast and efficient for all possible parameters, creating the need for algorithms which use different methods for different parameter ranges.

5.1 Complete Gamma Function

The most well known approximation of the complete gamma function is the one which was derived in the previous section, Stirling's approximation

$$\Gamma(n+1) \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

It becomes very accurate for large n, but can be unreliable for small n which is why it will not be used in any benchmarking tests. A more complicated but highly accurate alternative to Stirling's method is the Lanczos approximation.

Theorem 5.1. Lanczos Approximation[9] The Lanczos approximation of the gamma function holds that

$$\Gamma(z+1) = \sqrt{2\pi}\big(z+g+\tfrac{1}{2}\big)^{z+1/2} e^{-(z+g+1/2)} A_g(z)$$

$$\text{where } A_g(z) = \frac{1}{2}p_0(g) + p_1(g)\frac{z}{z+1} + p_2(g)\frac{z(z-1)}{(z+1)(z+2)} + \cdots$$

and $p_n(g)$ are constants which are precalculated and catalogued from a somewhat complicated series involving Chebyshev polynomials.

This method achieves high accuracy even with a moderate number of $A_g(z)$ terms, contains only elementary functions and allows for pre-calculation of the computationally heavier $p_n(g)$ constants. The specific inner workings of this method is beyond the scope of this thesis. The important facts are that it is fast and reliable, for this reason it will be used as an auxiliary function for all three methods which only approximate the incomplete gamma function.

5.2 Incomplete Gamma Function

Firstly it is appropriate to note that there in fact exists two different incomplete gamma functions, an upper and a lower incomplete gamma function.

Definition 5.2.

Lower incomplete gamma function

$$\gamma(\alpha, \beta x) = \int_0^{\beta x} t^{\alpha-1} e^{-t}\, dt$$

Upper incomplete gamma function

$$\Gamma(\alpha, \beta x) = \int_{\beta x}^{\infty} t^{\alpha-1} e^{-t}\, dt$$

It is the lower incomplete gamma function which is represented in the gamma CDF and is thus of interest here. However, the ability to approximate either is equally useful in determining the CDF since

$$\frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} = 1 - \frac{\Gamma(\alpha, \beta x)}{\Gamma(\alpha)}$$

which can be proven analytically, or perhaps more intuitively through the law of total probability.

Neither incomplete gamma function has any known approximation with high accuracy for all values, instead different methods are typically utilized for different parameter values. Three different gamma algorithms have been studied in this project, namely the DiDonato-Morris (1986) algorithm and the source code for the special functions libraries in Julia and Python (accessed 2026). The exact implementation in each differs slightly, but four distinct primary methods dominate in importance:

- Taylor expansion
- Continued fractions
- Integration by parts expansion
- Temme approximation (Same as LR)

Taylor[7]
One fairly straightforward approach to integrating a function which has no closed form anti-derivative is to replace the integrand with its Taylor expansion, since polynomials are easy to integrate. Naturally the downside of this approach is that it will not be possible to include an infinite number of terms and hence the approximation will begin to fail when evaluated at a point far away from where the expansion was performed. This point of expansion is typically 0, since the gamma distribution tends to have its mass concentrated close to 0. This makes the method ideal for approximating probabilities in the central region, as they will be close to the point of expansion. While the method is most easily understood as a Taylor expansion, it is worth noting that due to lack

of analyticity in the origin for some complex parameters, it is technically a generalized power series. The series is given by

$$\gamma(\alpha, \beta x) = \sum_{k=0}^{\infty} \frac{(\beta x)^{\alpha} e^{-\beta x} (\beta x)^k}{\alpha(\alpha+1)\dots(\alpha+k)}.$$

The structure of the series allows for the terms to be recursively generated which increases its computational efficiency. At some point the series must be truncated, exactly where this truncation occurs depends both on the desired accuracy and the parameters used.

Continued Fractions[7]
A method which is primarily used to evaluate moderate deviations is based on Gauss's continued fractions which provides the identity

$$\gamma(\alpha, \beta x) = \cfrac{(\beta x)^{\alpha} e^{-\beta x}}{\alpha - \cfrac{\alpha\beta x}{\alpha + 1 + \cfrac{\beta x}{\alpha + 2 - \cfrac{(\alpha+1)\beta x}{\alpha + 3 + \cfrac{2\beta x}{\alpha + 4 - \cfrac{(\alpha+2)\beta x}{\alpha + 5 + \cfrac{3\beta x}{\alpha + 6 - \ddots}}}}}}}.$$

In order to use this identity a modified Lentz algorithm is implemented[11, Chapter 5.2]. This algorithm is iterative and forward recursive, which allows the calculation to "start from the top" rather than "from the bottom", which would have been impossible since the fraction is infinite. Once again the details are technical and best understood by viewing the source material. Much of the value of this method comes from the fact that when the shape parameter α is small, and the deviation βx is at least moderate, the ratio of the iterated fraction approaches 1 rather quickly, meaning that truncation after a handful of terms does not exclude much useful data in these circumstances. Through this it provides accuracy in a specific parameter domain which other methods struggle with.

Integration by Parts[7]

By repeatedly integrating the upper incomplete gamma function by parts the integrand grows closer and closer to an exponential. If $\alpha \in \mathbb{N}$, integration by parts $\alpha - 1$ times will make the integrand exactly exponential. The integral of an exponential is naturally easy to evaluate exactly. Through the process of repeated integration by parts, a sum is generated by the exponents leading to the asymptotic expression

$$\gamma(\alpha, \beta x) \sim 1 - (\beta x)^{\alpha-1} e^{-\beta x} \sum_{k=0}^{\infty} \frac{(\alpha-1)(\alpha-2)\cdots(\alpha-k)}{(\beta x)^k}.$$

Any truncation error will be scaled down in importance if the deviation is large, and fewer terms will be required to avoid a large truncation error if α is small. If α is not an integer the optimal truncation problem becomes more complicated, but the error can begin to grow if not truncated appropriately.

5.3 Temme Approximation

The Temme approximation[8] was published before the LR approximation but also utilizes Daniels' SPA. One of the primary results of this thesis is that the first order LR gamma approximation is identical to the first order Temme approximation. It is reasonable to suspect, but has not been verified that higher order terms also coincide. This finding has dubious novelty, on the one hand none of the surveyed literature draws this connection, on the other hand the common underlying theory makes this connection potentially obvious to subject matter experts. From the perspective of this thesis' author however this overlap was a discovery.

An appropriate way of framing the connection between Temme and LR is that Temme solved the specific problem of using the SPA to derive an approximation specifically for the gamma CDF, while LR later managed to solve the general case problem of using the SPA to approximate any CDF, assuming that Cramér's condition is met. LR is also specifically framed as a way of approximating the CDF of a sample mean of many stochastic variables and puts the transform from MGF/CGF to CDF front and center, while Temme is narrowly focused on developing an approximation for the gamma CDF.

One of the great contributions to this thesis from Temme is that he was able to numerically derive approximations of higher order terms, which only have implicit definitions in LR. Generating these higher order terms requires evaluating differential equations without closed form solutions, but since Temme has approximated these it will be possible to benchmark them in the computational comparison in the results section.

Temme's own formulation is somewhat different to the LR formulation used so far. For example he chooses different auxiliary variables, utilizes the error function instead of the standard normal CDF and rather than stating it as an approximation his version is an asymptotic expansion. He also uses implicitly defined corrective terms for the complete gamma function in order to not introduce any approximation error and prefers the terminology of "(regularized) incomplete lower gamma function" instead of "gamma CDF". It is not important for the purposes of this thesis that the reader comprehends all the nuances of his formulation, but it may be interesting to compare some surface level similarities and differences between the first order LR gamma expression and the asymptotically accurate Temme expression.

Definition 5.3. Temme's Asymptotic Gamma Expansion[8]
The lower incomplete gamma function ratio

$$P(\alpha, \beta x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}, \quad \beta x \geq 0 \text{ and } \alpha > 0$$

with auxiliary variables

$$\mu = \frac{\beta x}{\alpha} - 1, \quad \eta = \text{sgn}(\mu)\sqrt{2[\mu - \log(1+\mu)]},$$

has uniform asymptotic expansion

$$P(\alpha, \beta x) = \tfrac{1}{2}\,\text{erfc}\Big(-\eta\sqrt{\alpha/2}\Big) - R_\alpha(\eta), \quad \eta \in \mathbb{R}$$
$$R_\alpha(\eta) \sim \frac{1}{\sqrt{2\pi\alpha}}\, e^{-\alpha\eta^2/2} \sum_{k=0}^{\infty} c_k(\eta)\, \alpha^{-k}, \quad \alpha \to \infty.$$

The coefficients c_k are defined through the recurrence

$$c_0(\eta) = \frac{1}{\mu} - \frac{1}{\eta},$$
$$\eta\, c_k(\eta) = \frac{dc_{k-1}(\eta)}{d\eta} + \frac{\eta}{\mu}\,\gamma_k, \qquad k \geq 1,$$

where γ_k denotes the coefficients of the asymptotic expansion

$$\frac{1}{\Gamma^*(\alpha)} \sim \sum_{k=0}^{\infty} \gamma_k\, \alpha^{-k}, \qquad \alpha \to \infty,$$
$$\Gamma^*(\alpha) = \sqrt{\alpha/(2\pi)}\, e^{\alpha} \alpha^{-\alpha}\, \Gamma(\alpha), \qquad \alpha > 0.$$

Apart from the previously mentioned numerically derived corrective terms $c_k(\eta)$ a very interesting aspect of Temme's expansion is that it is asymptotic as $\alpha \to \infty$. Because while the LR approximation is also asymptotically accurate, it is asymptotic as the number of stochastic variables $n \to \infty$ and this thesis exclusively uses $n = 1$. This was explained previously by the fact that a sufficiently dominant saddle point can compensate for this fact, which is true, but Temme's expansion shows that the expression is also asymptotic as α grows large. This is actually due to the same fact as in the LR case, namely that the gamma distribution becomes increasingly similar to a normal distribution as α grows large, and that distributions in general grow increasingly similar to the normal distribution as the sample size n grows large.

In summary, Temme's expansion is more technically intricate than the definitions used so far, but it provides precomputed higher order terms which can be used for benchmarking and shows that the expansion has asymptotic accuracy

as α grows large. The first order Temme approximation which will be referred to in the result section as Temme-1, is identical to the LR approximation of the gamma CDF.

6 Computational Benchmarking

This section will cover key methodological aspects about the benchmarking process. This includes primarily descriptions of how accuracy and runtime will be measured and which parameters will be tested. After this is covered the results of the benchmarking will be presented.

All methods being tested have been coded as separate functions in Julia. The Julia programming language was used because it is built from the ground up for numerical efficiency and is in general well regarded by mathematicians.

No special functions library for gamma approximations is used, as this would create a layer of opacity as to which of the algorithm's component methods is actually being utilized at any given time. All methods do however closely resemble their library versions. One key difference is however that the library implementations are set to iterate with additional terms until a certain desired accuracy is achieved, while the manual implementations of these methods are set to fixed iteration counts in order to be able to distinguish how different parameters effect the accuracy level for each method with a given number of terms.

Another difference of sorts is that in the library implementation each method is only ever used in conditions where it is considered an optimal choice, while the manually implemented methods will be tested for all parameter combinations, in order to be able to study the suitability of each method for a wide array of different parameters, including those where they are considered sub-optimal.

6.1 Runtime

In order to compare the computational cost of various approximation methods the Julia library BenchmarkTools is utilized. Every method runs a benchmarking test with every combination of parameters. The runtime is measured again and again during these benchmarks and the median time is catalogued. Median is chosen instead of mean since almost every cycle has the exact same runtime except a small number of outliers. This is presumably due to the CPU performing some other task in the middle of a benchmark, which slows the process considerably. These outlier values are therefore considered inaccurate and would skew the mean value notably.

Measuring each method's efficiency by the runtime of the script is imperfect as

it will vary greatly by hardware and can be prone to disruptions from background processes. Measuring this instead by floating point operations (FLOPS) or clock cycles was considered but turned out to be technically unfeasible with the hardware and operating system at hand. It is an open question if the most important measure is number of operations or number of cycles. The former more accurately represents the amount of information processing required in total while the latter advantages calculations which can be parallelized in modern CPU architecture.

For the purpose of transparency and future comparisons these tests were run on a 2022 MacBook Air M2 8 GB, always after a system restart with no other software running simultaneously.

6.2 Accuracy

All estimates are calculated using float64 type values, meaning that the maximum attainable precision is 15 accurate digits. In order to determine accuracy a precise reference value is needed. This reference has been calculated using adaptive quadrature and BigFloat data types until full precision is achieved in this bigger data type, the value is then truncated to a regular float. This method is highly computationally inefficient, but gives a reference which is precise and completely independent of any of the methods being evaluated.

The accuracy is measured in the unit "number of correct digits" which is calculated as

$$\text{Accuracy} = -\log_{10}(\text{Rel.Err.}) = -\log_{10}\left(\frac{|\text{Measured Val.} - \text{True Val.}|}{|\text{True Val.}|}\right).$$

For example if the true value is 0.14782 and the approximation is 0.14779 the accuracy will be measured as ≈ 3.69, where the number 3 means that the first 3 figures in the estimate are correct and the decimals indicate that the 4th digit is quite close to being accurate.

Something to keep in mind then with this measurement is that even an accuracy score of 2 or 3 will be sufficient for many applications, 15 digit representations are often overkill. It is also worth noting that this is a logarithmic scaling, and that each increased accuracy digit means that the relative error is 10 times smaller.

6.3 Parameters

Since the gamma CDF is

$$\frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$$

the x variable and the β parameter always appear together, which means that there is no point in letting them vary independently of each other. The only

value that matters is the product βx. The mean value of the gamma distribution is $\frac{\alpha}{\beta}$ and the variance is $\frac{\alpha}{\beta^2}$, hence in order to evaluate the distribution at k standard deviations from the mean the formula

$$x = \frac{\alpha}{\beta} + k\sqrt{\frac{\alpha}{\beta^2}}$$
$$x = \frac{\alpha + k\sqrt{\alpha}}{\beta}$$
$$x\beta = \alpha + k\sqrt{\alpha}$$

gives a method for determining βx using α and k. Describing number of standard deviations from the mean is more descriptive of the probability being evaluated than the raw numerical value of $x\beta$, therefore the benchmarks will be run with two independently varying parameters, α and k.

Some methods, most notably continued fractions and integration by parts, lead to special cases for integer value parameters. This is not the primary topic of interest in this comparison, which is why "messy" parameter values have been chosen. These are

$$k = [0.10745, 0.98423, 2.0197, 2.9487, 4.1321, 5.0281]$$
$$\alpha = [0.11345, 0.19472, 0.52396, 1.0768, 2.1188,$$
$$5.0365, 9.9764, 19.938, 49.871, 99.798].$$

This avoids accidentally testing simple special cases.

All of the tested parameters are to the right of the mean, that is to say, only right tail estimates are considered in this comparison. There are a few reasons for this, firstly the gamma distribution is defined on $[0, \infty)$ so x grows arbitrarily large in the right tail while the left tail always ends at $x = 0$, making it less interesting. Several of the methods also simply fail in the left tail, making for uninteresting results. LR is known to be a large deviation method and so it is natural that it would thrive in the right tail and this thesis is primarily about evaluating the use-case for the LR approximation of the gamma distribution. This is not to suggest that the left tail is unimportant or that it does not come with its own challenges, it is simply not the subject of this thesis.

7 Results

The purpose of the benchmarking comparison is to understand how accuracy and runtime depends on the deviation size, the α parameter and the number of terms used for each of the four primary methods. Bahadur-Rao is included as an aside. First the results from each method will be presented individually, at the end of the section all methods are compared. The interpretation of each graph will be kept short and to the point in this section and analyzed in more detail in the discussion.

7.1 Lugannani-Rice/Temme

All methods will be compared using heat maps in order to clearly display which parameter combinations lead to accurate contra inaccurate estimates. Different columns represent different α values while rows represent the number of standard deviations to the right of the mean. The value in each cell then is the accuracy measured in number of significant figures, the colour represents this same value and is meant to increase readability. Red colour then means low accuracy, yellow means moderate accuracy and blue means high accuracy. If a cell is empty it means that the approximated result was not a valid value for a probability.

std-alpha	0.11	0.19	0.52	1.08	2.12	5.04	9.98	19.94	49.87	99.8
0.11	1.11	1.44	2.26	3.13	4.46	4.16	4.47	4.87	5.44	5.88
0.98	1.63	2.02	3.24	3.61	3.74	4.2	4.64	5.1	5.72	6.19
2.02	2.07	2.6	3.56	3.61	3.97	4.56	5.06	5.59	6.28	6.79
2.95	2.43	3.14	3.55	3.84	4.31	5.01	5.6	6.2	6.99	7.57
4.13	2.85	4.42	3.77	4.23	4.82	5.7	6.43	7.19	8.17	8.88
5.03	3.19	4.03	3.99	4.56	5.25	6.28	7.15	8.05	9.23	10.08

Table 7.1: Significant figures of the Temme-1/LR approximation. Rows: σ, columns: α.

In table 7.1 which depicts the main approximation derived in the thesis the accuracy is very low for small α values and small deviations, but increases significantly for larger α values and larger deviations.

std-alpha	0.11	0.19	0.52	1.08	2.12	5.04	9.98	19.94	49.87	99.8
0.11			1.25	4.17	7.41	9.72	11.34	12.99	14.16	13.84
0.98			1.34	4.06	6.92	8.34	9.51	10.73	12.41	13.72
2.02			1.52	4.23	6.18	7.45	8.63	9.88	11.59	12.9
2.95			1.71	4.62	5.81	7.26	8.51	9.82	11.59	12.94
4.13				5.35	5.78	7.4	8.77	10.2	12.13	13.58
5.03				4.92	5.9	7.67	9.16	10.72	12.82	14.41

Table 7.2: Significant figures of the Temme-9 approximation. Rows: σ, columns: α.

Compare the previous table to the 9 term Temme approximation displayed in table 7.2. The accuracy here is very high for large α, but fails entirely for small α. For the large α a new pattern also emerges which was absent in Temme-1, namely that the accuracy is higher for small and large deviations than it is for moderate deviations.

Each method will also be evaluated using two line graphs where the average accuracy is plotted against either α or deviation size. This allows for comparing

how the accuracy is affected by number of terms, each term count will be represented by its own colour.

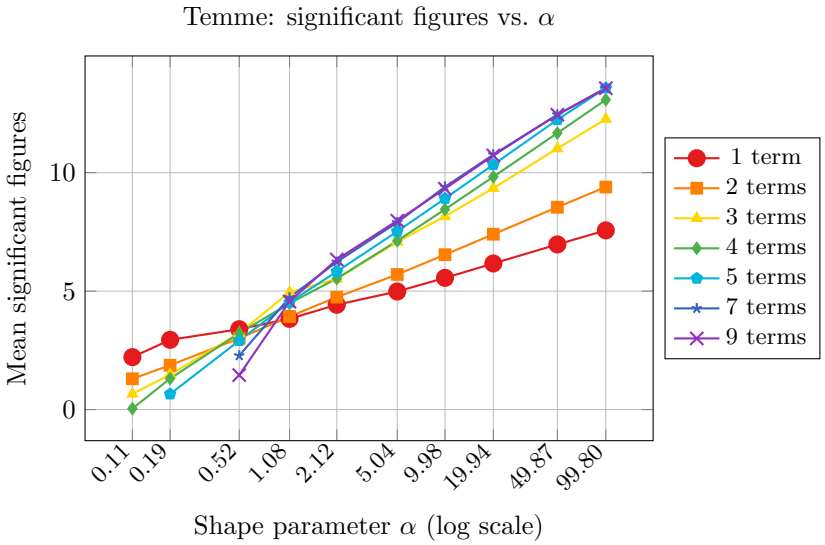

Figure 7.1: Mean significant figures of the Temme approximation as a function of α, averaged over all tested values of σ. Each curve is a fixed number of correction terms.

The pattern in figure 7.1 shows that adding extra terms does not improve the accuracy for small α, in general more terms actually lowers the accuracy here. In general it seems like each additional term up to three terms gives a notable increase in accuracy as long as α is moderately large, but higher order terms beyond this point has sharply diminishing rates of return. The accuracy strictly increases for larger α in every case.

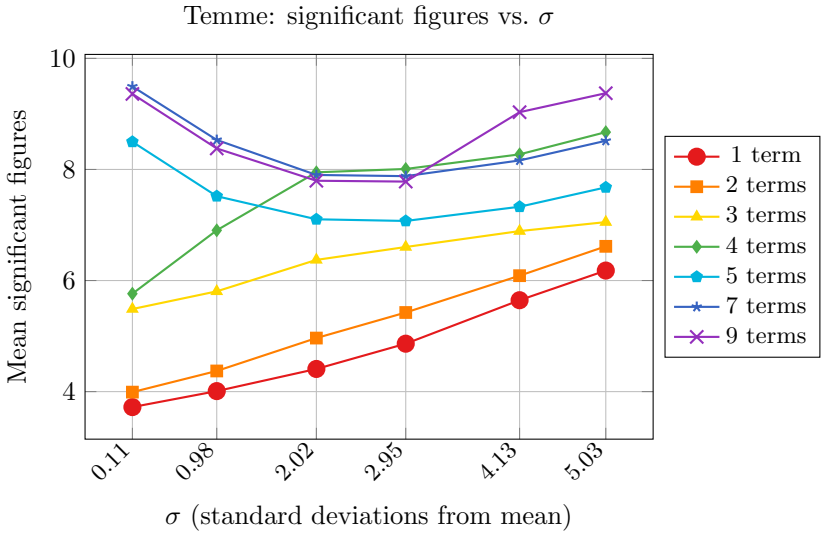

Figure 7.2: Mean significant figures of the Temme approximation as a function of σ, averaged over all tested values of α. Each curve is a fixed number of correction terms.

Comparing the average accuracy to the deviation size in figure 7.2 shows that lower order approximations strictly increase in accuracy as the deviation grows larger, but for higher order terms the accuracy around the mean grows to be substantial, while the accuracy for medium deviations falls behind. More terms does increase the accuracy for all deviation sizes in general, but once again the rate of return seems to diminish after the first three terms.

7.2 Bahadur-Rao

std-alpha	0.11	0.19	0.52	1.08	2.12	5.04	9.98	19.94	49.87	99.8
0.11										
0.98	0.54	0.58	0.67	0.73	0.78	0.84	0.87	0.9	0.93	0.94
2.02	1.05	1.14	1.35	1.51	1.66	1.84	1.95	2.06	2.16	2.22
2.95	1.37	1.51	1.82	2.08	2.32	2.63	2.85	3.05	3.27	3.4
4.13	1.71	1.9	2.35	2.73	3.11	3.62	4.01	4.38	4.82	5.1
5.03	1.94	2.17	2.71	3.19	3.69	4.36	4.91	5.45	6.1	6.52

Table 7.3: Significant figures of the Bahadur-Rao approximation. Rows: σ, columns: α.

Bahadur-Rao does not have any version with several terms so only the heatmap will be presented. Table 7.3 is similar to the Temme-1/LR estimate but with lower accuracy everywhere. It is essentially a simplified version of LR which requires much larger deviations and α in order to achieve accuracy.

7.3 Taylor Series

std-alpha	0.11	0.19	0.52	1.08	2.12	5.04	9.98	19.94	49.87	99.8
0.11	15	14.85	15	13.69	10.22	6.51	4.29	2.68	1.4	0.85
0.98	15	15	12.89	9.94	7.37	4.57	2.89	1.71	0.81	0.46
2.02	14.9	13.39	9.83	7.36	5.25	3.02	1.76	0.92	0.35	0.17
2.95	13.19	11.33	8.02	5.79	3.94	2.07	1.07	0.47	0.13	0.05
4.13	11.21	9.44	6.37	4.37	2.76	1.25	0.53	0.17	0.03	0.01
5.03	10.05	8.34	5.42	3.56	2.11	0.83	0.29	0.07	0.01	0

Table 7.4: Significant figures of the Taylor series approximation with 15 terms. Rows: σ, columns: α.

The Taylor heatmap in table 7.4 shows the exact opposite pattern as was observed for Temme-1/LR. Small deviations and small α leads to highly accurate approximations while the accuracy shrinks rapidly as these parameters grow.

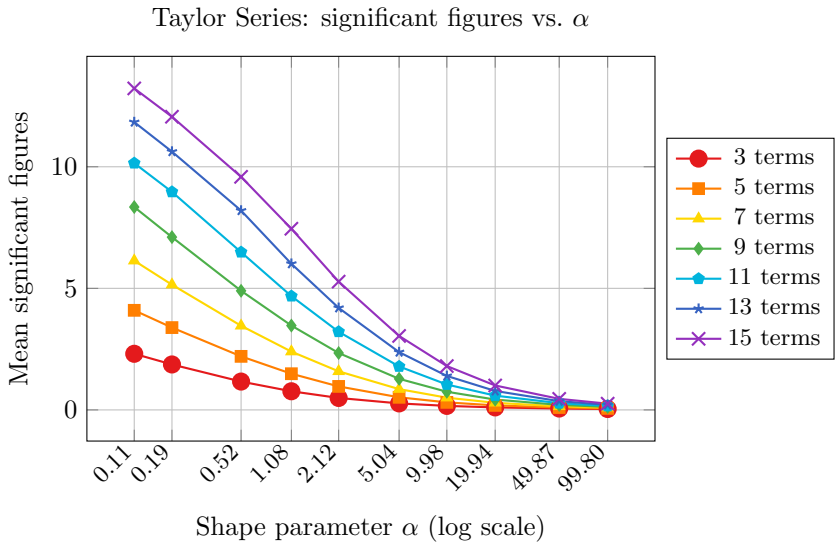

Figure 7.3: Mean significant figures of the Taylor series approximation as a function of α, averaged over all tested values of σ. Each curve is a fixed iteration count.

In figure 7.3 the relationship between accuracy and α is displayed for Taylor series with various number of terms. More terms seem to always give better results but even with 15 terms the approximation is very poor when α is moderately large.

Comparing the accuracy to deviations in figure 7.4 shows a similar relationship as in the previous figure. Accuracy always seems to grow with additional terms and approximations for small deviations are much more accurate than for large ones.

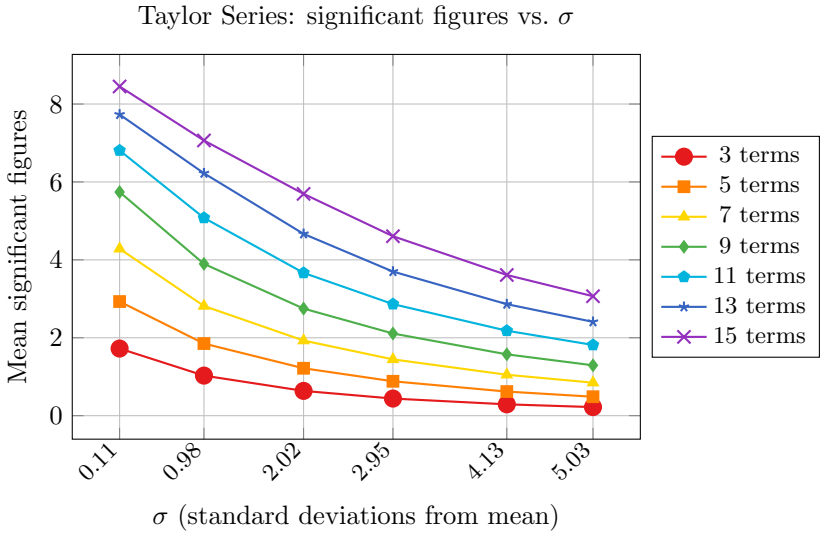

Figure 7.4: Mean significant figures of the Taylor series approximation as a function of σ, averaged over all tested values of α. Each curve is a fixed iteration count.

7.4 Continued Fractions

std-alpha	0.11	0.19	0.52	1.08	2.12	5.04	9.98	19.94	49.87	99.8
0.11	1.85	2.05	2.93	4.73	5.88	7.39	3.62	2.31	1.39	0.98
0.98	3.11	3.42	4.51	6.43	7.68	9.3	5.58	4.32	3.44	3.04
2.02	4.11	4.55	5.87	7.97	9.38	11.18	7.59	6.44	5.67	5.33
2.95	4.84	5.37	6.89	9.14	10.7	12.69	9.24	8.22	7.6	7.35
4.13	5.63	6.28	8.02	10.46	12.22	14.46	11.24	10.43	10.06	9.97
5.03	6.17	6.89	8.79	11.38	13.28	15	12.7	12.08	11.95	12.02

Table 7.5: Significant figures of the continued fractions approximation with 5 terms. Rows: σ, columns: α.

In table 7.5 the accuracy increases notably for larger deviations, the relationships between accuracy and α is complicated but seems to be based primarily on how close the value of α is to 5.

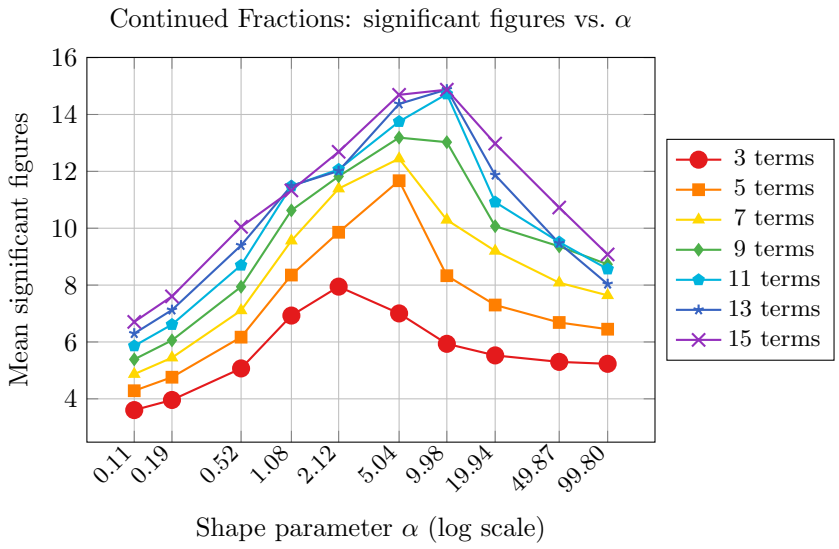

Figure 7.5: Mean significant figures of the continued fractions approximation as a function of α, averaged over all tested values of σ.

Comparing accuracy to α in figure 7.5 shows that the accuracy peak around $\alpha = 5$ displayed in the heatmap is not independent of term count, each version of the continued fraction achieves its maximum accuracy when term count is roughly equal to α. With possibly a few minor exceptions, the accuracy however always seems to be higher when more terms are used, though seemingly with somewhat diminishing returns.

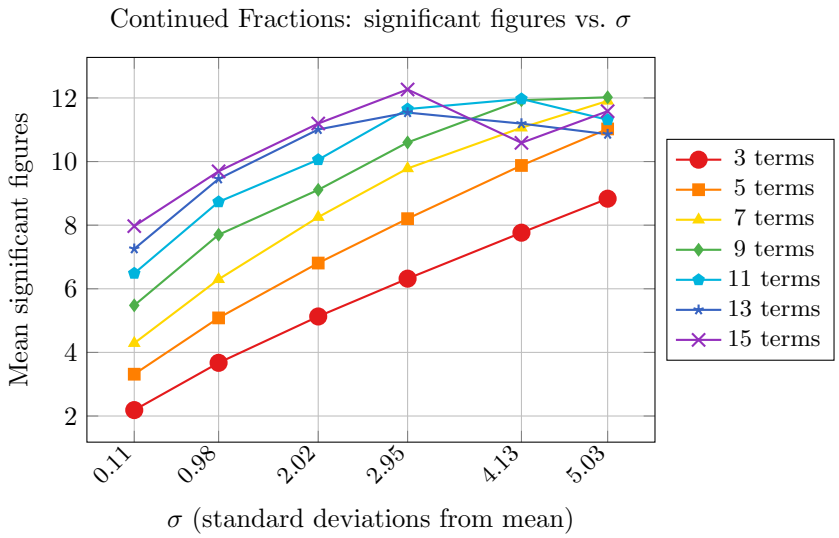

Figure 7.6: Mean significant figures of the continued fractions approximation as a function of σ, averaged over all tested values of α.

The pattern in figure 7.6 indicates that larger deviations generally increases the accuracy, but that this pattern begins to break for large deviations with many terms. For these large deviations the approximations with 5 or more terms begin to converge.

7.5 Integration by Parts

std-alpha	0.11	0.19	0.52	1.08	2.12	5.04	9.98	19.94	49.87	99.8
0.11						5.63	9.73	4.08	1.65	0.99
0.98					1.59	7.92	11.58	5.59	2.85	2.01
2.02				0.46	4.16	10.12	13.53	7.3	4.31	3.34
2.95				2.41	6.02	11.86	15	8.84	5.74	4.7
4.13				4.44	8.04	13.86	15	10.82	7.72	6.67
5.03				5.76	9.4	15	15	12.36	9.32	8.33

Table 7.6: Significant figures of the integration by parts method at 13 iterations. Rows: σ, columns: α.

Integration by parts with 13 terms displayed in table 7.6 shows a pattern somewhat similar to that observed for continued fractions. Large deviations increases the accuracy but the α parameter seems much more important in determining accuracy. In this case $\alpha \approx 10$ gives the most accurate estimate while small α fails entirely due to including too many terms.

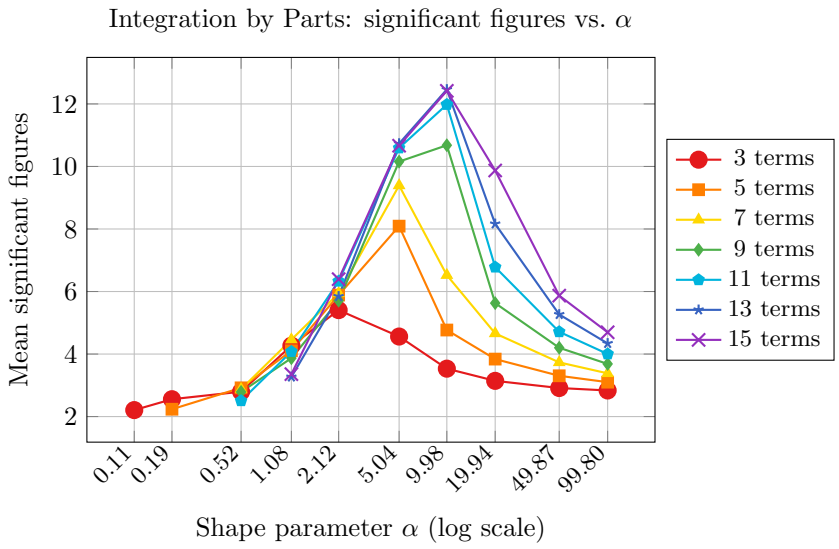

Figure 7.7: Mean significant figures of the integration by parts method as a function of α, averaged over all tested values of σ.

In figure 7.7 it becomes evident that no matter the number of terms used, this method is ill suited for small α. Just as with continued fractions the accuracy for integration by parts peaks when the term count matches the α value as closely as possible.

The comparison in figure 7.8 shows that the accuracy grows steadily for larger deviations and that more terms in general increases the accuracy, though with some exceptions where the 11, 13 and 15 term versions alternate as the most accurate for moderate deviations. In this case it is quite likely that this pattern has more to do with how higher numbers of terms interact with small α than any actual competitive advantage of using fewer terms for certain deviation ranges.

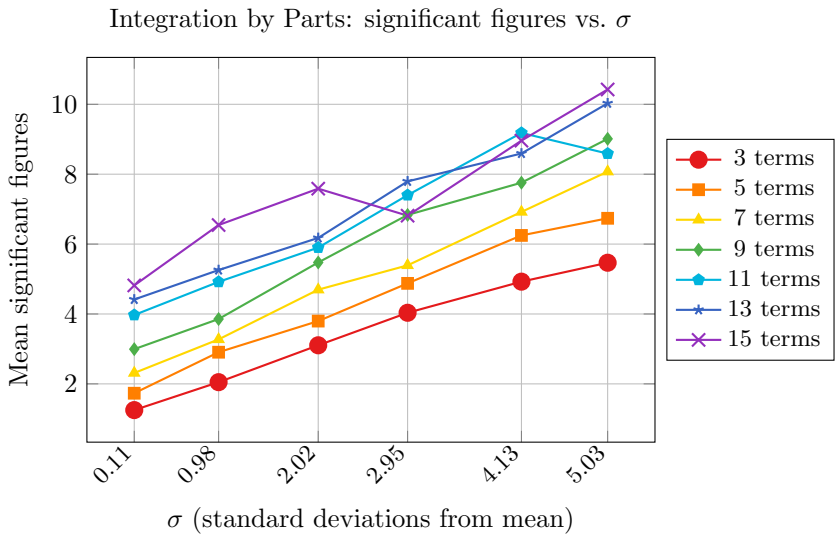

Figure 7.8: Mean significant figures of the integration by parts method as a function of σ, averaged over all tested values of α.

7.6 Method Comparison

Having looked at all methods individually some averaged metrics for all methods will be compared. The accuracy metric is much more nuanced when studied individually as this provides information about which methods are suitable in different situations. Averaged accuracy on the other hand heavily favours methods which are generalists rather than specialists and methods which happen to be accurate for the specific parameters being tested. That all being said, some genuine novel insights can also arise from such broad comparisons.

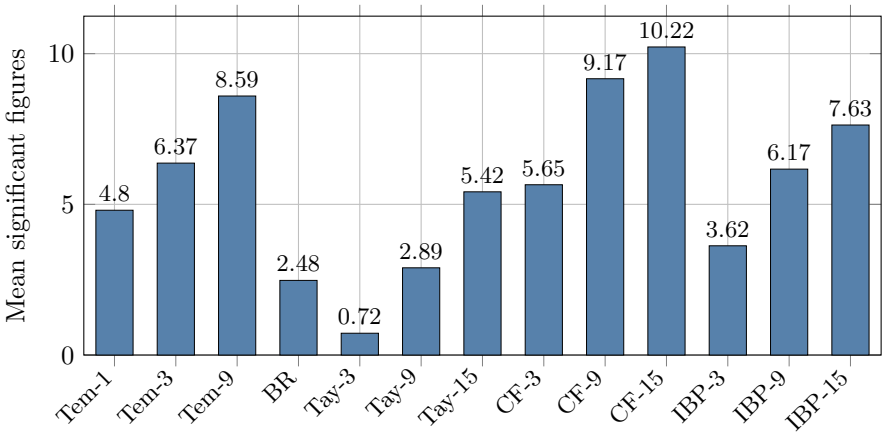

Figure 7.9: Mean significant figures for each method, averaged over all valid (k, α) combinations in the messy parameter set.

The average accuracy in figure 7.9 shows that continued fractions have the highest average accuracy for this parameter set, with a large accuracy increase between 3 and 9 terms but with a far smaller increased accuracy when 15 terms are used. Temme has a decent average, but one that does not grow as much as other methods when a larger number of terms are used. The Taylor expansion is the method which has the largest need for many terms as the method starts out with terrible accuracy if only 3 terms are used, but the accuracy grows substantially when more terms are included. Bahadur-Rao is unsurprisingly very poor on average. Integration by parts displays an average performance and displays significant diminishing return.

Next the runtime of each method will be compared. This is a far less common metric to analyze than accuracy and can be regarded as one of the novel contributions of this thesis.

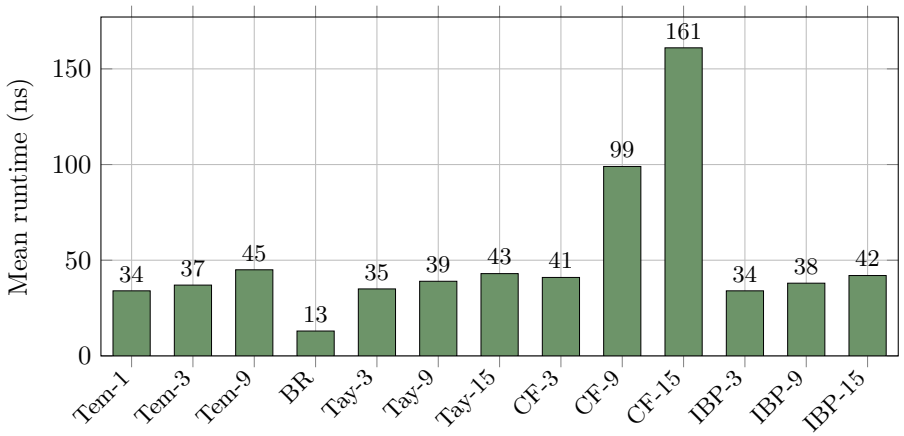

Figure 7.10: Mean runtime in nanoseconds for each method, averaged over all valid (k, α) combinations in the messy parameter set.

The runtime comparison in figure 7.10 displays several key results. First of all, the runtimes of the various methods are quite similar overall. There is nothing that guarantees this and it would have been perfectly plausible for the methods to display a far larger range of runtimes.

Secondly, only continued fractions grows rapidly in runtime as number of terms increase. This means that the baseline computational cost of initializing all required variables and evaluating the auxiliary functions is much more computationally expensive than adding additional terms to the approximation, with the exception of continued fractions. For other methods using many terms is cheap.

The Temme approximation displays some increase in runtime as the number of terms increase, but it relies on precomputed parameters of which only 9 exist in the studied sources. The runtime for Temme with 9 terms does not indicate a troubling potential trend, it is already the slowest possible version.

None of the methods have a runtime which seems dependent on the deviation size or α which is why no graphs comparing such a dependency are displayed. This is a result in and of itself, the runtime is invariant for large ranges of parameter values. This pattern is not guaranteed to hold for integer parameters, as the difference in difficulty of evaluating small and large integers may be larger than the difference in difficulty of evaluating small and large decimal values.

8 Discussion

The data from the previous section will now be analyzed and some final conclusions and remarks will be drawn. In general the computer library implementations of gamma approximations are highly optimized and indeed do use every method in roughly the circumstances where test results indicate they are most accurate. The results can be seen as a validation of algorithms like DiDonato-Morris.

8.1 Advantages and Disadvantages of Each Method

Taylor Expansion
In the small deviation, small α regime, the Taylor method performs very well. It is not only accurate given enough terms, but also robust. Additional terms are cheap to compute and always increase the accuracy. There are no obvious problematic cases, the only issue with the method is that it requires more total terms than the other methods and especially for large deviations/large α a very large number of terms need to be computed.

Continued Fractions
Continued fractions is perhaps the most difficult method to evaluate, because it has several "preferences" which are somewhat conflicting. It works well when α is small, but not when it is smaller than 1. Ideally α should be an integer or at least close to an integer. In the tests continued fractions generally increase in accuracy for larger deviations, but the method is not asymptotically accurate at any given truncation point, so this pattern is unlikely to hold forever. Yet the tests do show that with the parameters tested, bigger deviations is strictly better. Continued fractions also has a runtime which grows much faster for each additional term than the other iterated methods, meaning that only a moderate amount of terms can be computed in the same time as other methods can compute a very large amount. This then reduces its ability to scale number of terms to grow larger than α. On the other hand continued fractions display a very high average level of accuracy even for a moderate amount of terms compared to other methods. In summary then continued fractions is most suited for small α and large deviations, though with many qualifiers.

Integration by Parts
Integration by parts at the zoomed out level has many similar properties as

continued fractions. It works best in cases where number of terms is similar to α and where α is an integer or almost an integer. It also works best for large deviations. There are however some very important differences as well. If too many terms are used the approximation actually fails entirely, this happens first for small deviations and especially small α. This is quite a serious failure mode and introduces an optimal truncation problem. The Taylor series can for example always add more terms in order to increase accuracy, while this is not true for integration by parts. One big advantage compared to continued fractions is how little extra time each additional term takes to compute however, so in regimes where additional terms does increase accuracy, this process is far cheaper. This then also means that integration by parts can adapt number of terms in order to surpass moderately large α in a way that continued fractions cannot. Integration by parts is known to be asymptotically accurate, meaning that larger deviations should always lead to more accurate approximations, though a few datapoints observed seem to conflict with this theory. In summary then integration by parts is most suitable for very large deviations and for small or moderate α, again with many qualifiers.

Temme and Lugannani-Rice

The LR/Temme-1 estimate displays then a very simple pattern. Bigger deviations and bigger α give more accurate estimates, the exact opposite to Taylor. Both of the previously mentioned methods also benefit from large deviations, but neither of them benefited in the same way from large α, this can then be regarded as the biggest relative advantage of LR/Temme. When Temme devised his gamma approximation, which matches the LR approximation, he spent great effort to numerically derive the next handful of terms in the series. The biggest downside of LR/Temme then is the inability to compute higher order terms needed to achieve a specified level of precision. Higher order Temme approximations break some of the patterns observed in the LR case, larger deviations are not always unambiguously better in these cases and some small α approximations fail entirely when too many terms are used. One potential explanation why larger deviations may not be strictly better for higher order Temme approximations is because the corrective terms have two separate components to them, one which corrects the normal CDF error and one which corrects the large deviation error. It is possible that the medium deviation zone falls behind somewhat as there are no corrective components targeting this range directly. As a matter of fact the higher order Temme terms are so efficient at correcting the normal error that the method becomes very good for small deviations. The higher order terms then change the initial picture from the LR case, only the low order approximations have accuracy with a clear large deviation dependency. For any number of terms however large α give by far the most accurate results.

Bahadur-Rao

Bahadur-Rao is as previously mentioned just a computationally cheaper version of LR with lower accuracy. An advantage is however that it does not rely on any normal distribution approximation and thus involves no error from such an approximation. Its accuracy scaling makes it only suitable for very large deviations and overall it displays very few advantages over LR.

The fact that higher order approximations of LR are difficult to derive somewhat limits the method's usefulness, but overall the moderate term count implementations of Temme perform well on average. None of the other methods benefit in the same way from larger α either, which establishes a clear niche for Temme.

8.2 Runtime Comparison

Runtime is far less studied than accuracy for each of these methods. The main findings are that most methods have very similar runtimes, that additional terms/iterations are in general computationally cheap compared to the rest of the calculations which each method performs, with the notable exception of continued fractions which does have an expensive iteration cost. Parameter size has a negligible impact on the total runtime in most cases. Since the total runtime is measured on the scale of tens of nanoseconds, in many cases a slightly longer runtime for higher accuracy is likely worthwhile.

8.3 A Proof of Concept

The fact that the LR approximation of the gamma distribution would yield a previously established approximation, Temme's, came as a surprise. Nowhere in the surveyed literature has this been stated, therefore it can arguably be considered a novel result, though it is unlikely that this would surprise subject matter experts. It is more likely that this is a non-standard way of phrasing it, and the fact that both these methods approximate a CDF using saddle point related methods is well known.

The fact that Temme's approximation has proved so useful however and the fact that LR provides a general method for making similar approximations for other distributions leads to the intriguing possibility that similar approximations for other distributions may be genuinely novel. A basic requirement for this to be plausible is that the CGF should be substantially simpler to evaluate than the CDF. It is also highly beneficial if, as with the gamma distribution, the saddle point can be expressed in explicit terms. Otherwise it will require a root-finding algorithm, which would slow down the approximation notably. Among the distributions surveyed the most promising case is the non-central χ^2 distribution, which meets both of these criteria. Multivariate and discrete distributions have not been investigated at all however and such cases may also exist. Note that this is speculative and based on cursory research, it may for example be the case that this would once again reproduce a previously known approximation.

References

[1] H. Cramér, "Sur un nouveau théorème-limite de la théorie des probabilités," *Actualités scientifiques et industrielles*, no. 736, pp. 2–23, Hermann

& Cie, Paris, 1938. English translation by H. Touchette, "On a new limit theorem in probability theory," arXiv:1802.05988, 2018. Available at: https://arxiv.org/pdf/1802.05988 (Accessed: 18 February 2026).

[2] H. E. Daniels, "Saddlepoint approximations in statistics," *The Annals of Mathematical Statistics*, vol. 25, no. 4, pp. 631–650, 1954.

[3] R. Lugannani and S. Rice, "Saddle point approximation for the distribution of the sum of independent random variables," *Advances in Applied Probability*, vol. 12, no. 2, pp. 475–490, 1980.

[4] R. R. Bahadur and R. Ranga Rao, "On deviations of the sample mean," *The Annals of Mathematical Statistics*, vol. 31, no. 4, pp. 1015–1027, December 1960. https://doi.org/10.1214/aoms/1177705674

[5] E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis with Applications to Engineering, Science, and Mathematics*, 3rd ed. Pearson Education Limited, 2013.

[6] R. W. Butler, *Saddlepoint Approximations with Applications*, Cambridge University Press, Cambridge, 2007.

[7] A. R. DiDonato and A. H. Morris Jr., "Computation of the incomplete gamma function ratios and their inverse," *ACM Transactions on Mathematical Software*, vol. 12, no. 4, pp. 377–393, 1986.

[8] N. M. Temme, "The asymptotic expansion of the incomplete gamma functions," *SIAM Journal on Mathematical Analysis*, vol. 10, no. 4, pp. 757–766, 1979. https://doi.org/10.1137/0510071

[9] C. Lanczos, "A precision approximation of the gamma function," *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, vol. 1, no. 1, pp. 86–96, 1964.

[10] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*, Springer, New York, 1999. https://doi.org/10.1007/978-1-4757-3069-2

[11] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed., Cambridge University Press, Cambridge, 2007, sec. 5.2 ("Evaluation of Continued Fractions").