Bounded Relative Error Importance Sampling and Rare Event Simulation

Don L. McLeish
Department of Statistics, U. Waterloo and ETH Zürich.

July 13, 2008

Abstract

We consider estimating tail events using exponential families of importance sampling distributions and show that under mild conditions on the exponential family we can achieve bounded relative error.

Keywords: rare event simulation, relative error, g&h distribution, Monte Carlo methods, Importance sampling, Cross-entropy, Rényi Divergence.

1 Introduction

Suppose $X$ is a random variable with cumulative distribution function $F$ and probability density function $f$ with respect to Lebesgue measure. Suppose we wish to estimate an expected value such as

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

where $g$ is an arbitrary integrable function. We wish to use importance sampling (IS): generate $X$ from an alternative distribution in an exponential family having probability density

$$f_{\theta}(x) = \frac{1}{m(\theta)} e^{\theta T(x)} f(x), \quad (1)$$

where $m(\theta) = \int e^{\theta T(x)} f(x)dx$. The modification of the original density by the multiplication of a term like $e^{\theta T(x)}$ when $T(x) = x$ is variously referred to in the literature as an exponential twist or tilt of the density $f$ but we will adopt this language to include a modification of $f$ to include a density of the form (1) and use the phrase standard exponential twist when $T(x) = x$. There are many potential Monte Carlo estimators of expected value which exploit specific features of the functions $f$ and $g$ to achieve variance reduction (see for example McLeish (2005), Chapter 4) but we will concentrate here on a specific technique,
importance sampling. Having generated independent $X_i$, $i = 1, ..., n$ from $f_\theta$, we estimate $E(g(X))$ with an importance sampling estimator

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) \frac{f(X_i)}{f_\theta(X_i)}.$$  

This is an unbiased estimator of $E(g(X))$.

Our primary concern in this paper is the efficiency or the variance of such estimators in the special case that we are interested in tail events $g(X) = I(X > t)$ and the probability

$$p_t = P(X > t) = \int_t^\infty f(x)dx$$

is small. Then when $n = 1$, the importance sample estimator of $p$ is

$$\hat{p}_t = \frac{f(X)}{f_\theta(X)} I(X > t), \text{ where } X \sim f_\theta(x). \quad (2)$$

Again $\hat{p}_t$ is an unbiased estimator of $p$, i.e. $E(\hat{p}_t) = p_t$.

It is natural to choose the value of $\theta$ which minimizes some criterion, one such being the variance of the IS estimator,

$$\text{var}_\theta(\hat{p}_t) = m(\theta) \int_t^\infty e^{-\theta T(x)} f(x)dx - p_t^2.$$  

Because of its relationship to the very substantial literature on risk measurement (see McNeil, Frey and Embrechts (2005)), there is a considerable interest in estimating probabilities of tail events, using either simulation or asymptotic approximations to the survivor function (see for example Embrechts, Degen and Lambrigger (2007)). For a detailed discussion of Monte Carlo estimation techniques for rare event probabilities, see two recent books on the topic, Kroese and Rubinstein (2008) and Asmussen and Glynn (2007) as well as Asmussen, Kroese and Rubinstein (2005), and Homem-de-Mello and Rubinstein (2002). In Kroese and Rubinstein (2008), cross-entropy is used to motivate iterative methods of choosing an appropriate parameter value $\theta$ for the importance sampling distribution (1). In Asmussen and Glynn (2007) a number of estimators similar to (2) are discussed, including a very efficient estimator obtained by conditioning. In this paper we develop some useful and simple rules for optimal or near optimal values of the parameter $\theta$, but more importantly discuss which statistics $T(x)$, i.e. which exponential families of distributions, can lead to importance sampling estimators with bounded relative error.

2 Minimizing Divergence, Importance Sampling and Bounded Relative Error

For tail events, the variance or standard error is less suitable than a version scaled by the mean, such as the relative error.
Definition 1. The relative error of the importance sample estimator is the ratio of the estimator’s standard deviation to its mean.

For \( n \) independent simulations \( X_i \) from the p.d.f. \( f \), the IS estimator is

\[
\hat{p}_t = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)}{f_\theta(X_i)} I(X_i > t),
\]

and the relative error of this estimator is

\[
\frac{n^{-1/2}}{\sqrt{D_2(ch; f_\theta) - 1}}.
\]

We will use \( c \) to denote a generic normalizing real constant (which may change on each use) so that for a non-negative integrable function \( h \), \( ch \) denotes a corresponding probability density function. We may minimize the relative error of the IS estimator by minimizing over \( \theta \) the Rényi generalized divergence \( D_\alpha(ch; f_\theta) \) of order \( \alpha = 2 \), where for two probability density functions \( ch(x) = cf(x)I(x > t) \) and \( f_\theta(x) \) (see Rényi, (1961)) we define

\[
D_\alpha(ch; f_\theta) = \begin{cases} 
\int \ln \left( \frac{ch(x)}{f_\theta(x)} \right) h(x) dx & \alpha = 1 \\
\frac{1}{\alpha - 1} \ln \left( \int \left( \frac{ch(x)}{f_\theta(x)} \right)^{\alpha - 1} h(x) dx \right) & \alpha > 0, \alpha \neq 1 
\end{cases}
\]

We assume, of course, the integrals in (4) exist. Since the relative error (3) is \( n^{-1/2}\sqrt{D_2(ch; f_\theta) - 1} \) we have:

Proposition 1. The variance of the importance sample estimator (2) is minimized if the importance distribution \( f_\theta \) is chosen to minimize \( D_2(ch; f_\theta) \) where \( h(x) = f(x)I(x > t) \).

If the parametric family \( f_\theta \) contains a density proportional to \( h \), then this obviously minimizes (4) because then the divergence equals zero, its minimum possible value, whatever the choice of \( \alpha > 0 \). Unfortunately, sampling from a density like \( ch(x) = cf(x)I(x > t) \) is often not possible. In the rare case when it is possible, it may focus too specifically on estimating a single probability \( P(X > t) \) when we are interested in the whole tail behaviour of the function, a point to be returned to shortly. The most important special case of Rényi generalized divergence (4) is the Kullbach-Leibler divergence from \( f_\theta \) to \( h \) when \( \alpha = 1 \). Other functions have also been used in the literature replacing \( D_\alpha \) (see for example Ridder and Rubinstein, (2007)).

The following metaprinciple is often invoked to generate IS estimators of a non-negative integrable function \( h(x) \). It is based on the idea that the closer we are to the “perfect” IS distribution \( ch \), the more efficient our estimator.

Minimum Divergence Principle. If we wish to estimate an integral \( \int h(x) dx \) using importance sampling, choose an IS distribution \( f_\theta \) which minimizes the
Rényi generalized divergence $D_\alpha(ch; f_\theta)$ between the family $f_\theta$ and the target $ch$.

Typically $\alpha$ is chosen to be 1 or 2 for the application of the above, and in many cases the minimization problem suggested by this principle is quite tractable. The first order condition (assuming differentiability under the integral) for $\alpha \neq 1$, is $\frac{\partial}{\partial \theta} D_{\alpha}(ch; f_\theta) = 0$. If $h(x) = f(x)I(x > t)$, the condition is

$$\frac{\partial}{\partial \theta} E_{ch} \left[ \left( \frac{ch}{f_\theta} \right)^{\alpha-1} \right] = \frac{1}{\theta^\alpha} \frac{\partial}{\partial \theta} \int t^\infty \left[ m(\theta) e^{-\theta T(x)} \right]^{\alpha-1} f(x)dx = 0 \text{ or }$$

$$\int t^\infty \left[ \frac{m'(\theta)}{m(\theta)} - T(x) \right] f(x)dx = 0 \text{ when } \alpha = 1, \text{ and in general }$$

$$\int t^\infty e^{-\theta T(x)(\alpha-2)} \left[ m'(\theta)e^{-\theta T(x)} - m(\theta)T(x)e^{-\theta T(x)} \right] f(x)dx = 0 \text{ or, }$$

$$\int t^\infty e^{-\theta T(x)(\alpha-1)} \left[ \frac{m'(\theta)}{m(\theta)} - T(x) \right] f(x)dx = 0.$$  

(5)

In general (5) says that $T$ is an unbiased estimator of $\frac{m'(\theta)}{m(\theta)}$ under a density proportional to $e^{-\theta T(x)(\alpha-1)} f(x)I(x > t)$. For $\alpha > 1$, we need to solve (5) for $\theta$ iteratively. The special case $\alpha = 1$ is slightly simpler in that the “weights” $e^{-\theta T(x)(\alpha-1)}$ are 1 and do not require the value of the parameter $\theta$.

**Definition 2.** Suppose $\mathcal{H}$ is a class of integrable functions $h$. We say the family $f_\theta$ has bounded relative error for the class $\mathcal{H}$ if

$$\sup_{h \in \mathcal{H}} \inf_{\theta} D_2(ch; f_\theta) < \infty.$$  

(6)

This condition says that the orbit of the exponential family passes close enough to every function in $\mathcal{H}$ that the divergence is finite. For rare event simulation, it is easy to find a parametric class of distributions which provides bounded relative error for estimating the probability of events in the tail. In particular, if $\mathcal{H}$ is the class of functions $I(x > t)f(x)$ as $t \to \infty$, we may define the family of densities

$$f_\theta(x) = c_\theta f(x)I(x > \theta), \text{ for } \theta > 0.$$  

(7)

for normalizing constants $c_\theta$. Then for any $h \in \mathcal{H}$, this family includes a member which, when used as an IS distribution, has zero variance for estimating $\int h(x)dx$. However the class (7) of IS distributions is not generally an exponential family of distributions (in fact the distributions are not mutually absolutely continuous) and it is often very difficult to generate from members of this family.
Our preference is for a single exponential family (1) which passes close enough to every function \( h \in \mathcal{H} \) that its relative error is bounded in the sense of (6). With an exponential family of distributions generated by a single canonical sufficient statistic \( T \), we can easily aggregate information collected at different parameter values and estimate parameters.

We mentioned earlier that IS distributions such as (7), designed for a specific problem, may be highly inefficient for a closely related problem. Consider the following example: we wish to estimate \( P(X > t) \), for \( t > 0 \) large, where \( X \) follows a \( N(0, 1) \) distribution. The zero-variance importance sample distribution (7) is \( \frac{1}{1 - \Phi(t)} \varphi(x) I(x > t) \) where \( \varphi \) and \( \Phi \) are the standard normal p.d.f. and c.d.f. respectively. Suppose we are interested in the tail behaviour, including conditional probabilities of excess such as

\[
P(X > t + s | X > t) = \frac{1 - \Phi(t + s)}{1 - \Phi(t)} \quad \text{(8)}
\]

Sampling from the zero-variance importance sample distribution is highly inefficient for estimating \( P(X > t + s | X > t) \) when \( s > 0 \) since if \( X \) has p.d.f. \( \varphi(x)/(1 - \Phi(t)) \) for \( x > t \), the relative error for estimating \( P(X > t + s | X > t) \) is approximately

\[
\frac{\sqrt{e^{-st + s^2} (1 - e^{-st + s^2})}}{e^{-st + s^2}} = \sqrt{e^{st} - 1} - 1
\]

and this grows extremely rapidly in both \( t \) and \( s \). We would need a sample size of around \( 10^4e^{st}e^{s^2} \) from this IS density to achieve a relative error of 1%, about 60 trillion if \( s = 3 \) and \( t = 6 \). Suppose we apply the criterion (6) with \( \mathcal{H} \) the set of functions of the form \( I(x > t)\varphi(x) \) for \( t > 0 \) and \( f_\theta(x) \) the family of Gumbel distributions. We will see that being less greedy and settling for an IS estimator with positive variance for the immediate problem of estimating \( P(X > t) \) allows us a bounded relative error for the estimation of all tail probabilities.

There are several optimization problems related to the design of IS algorithms with bounded relative error, since in addition to possible values of \( \theta \) and a class \( \mathcal{H} \) of functions whose integrals are to be estimated, we may have a whole class \( T \) of potential functions \( T(x) \) for generating the exponential family (1). Choosing both \( T \) and \( \theta \) optimally for a specific \( h \) is equivalent to solving:

\[
\sup_{h \in \mathcal{H}} \inf_{T \in T} \inf_{\theta} D_2(ch(\cdot); \frac{1}{m(\theta)} e^{\theta T(\cdot)} f(\cdot)).
\]

More practically, requiring that the same exponential family (i.e. \( T(x) \)) be used for all \( h \in \mathcal{H} \) is equivalent to

\[
\inf_{T \in T} \sup_{h \in \mathcal{H}} \inf_{\theta} D_2(ch(\cdot); \frac{1}{m(\theta)} e^{\theta T(\cdot)} f(\cdot)). \quad \text{(9)}
\]
We are unable to solve (9) generally but we do provide conditions on the exponential family and \(f\) which guarantee bounded relative error for rare event simulation. We begin with some basic examples.

**Example 1. (Uniform distribution)** Suppose that \(f(x) = 1\), for \(0 < x < 1\) and we wish to design an IS estimator of the probability \(p = P(X > 1 - p)\). Consider using the standard exponential tilt with \(T(X) = X\), resulting in the probability density:

\[
f_\theta(x) = \frac{\theta}{e^\theta - 1} e^{\theta x}, \text{ for } 0 < x < 1, \text{ where } \theta > 0.
\]

Then with \(h(x) = I(x > 1 - p)\), and \(c = 1/p\), for \(\alpha > 1\),

\[
\exp\{D_\alpha(ch; f_\theta)\} = \int \left( \frac{ch(x)}{f_\theta(x)} \right)^{\alpha - 1} ch(x)dx = \frac{(1 - e^{-\theta})^{\alpha - 1}(e^{(\alpha - 1)\theta p} - 1)}{\theta^\alpha (\alpha - 1) p^\alpha}.
\]

To minimize this, we must solve

\[
\frac{\partial}{\partial \theta} \left\{ -\alpha \ln(\theta) + (\alpha - 1) \ln(1 - e^{-\theta}) + \ln(e^{(\alpha - 1)\theta p} - 1) \right\} = 0,
\]

or \((1 - \frac{1}{\alpha}) \left\{ \frac{1}{e^\theta - 1} + \frac{p}{1 - e^{-(\alpha - 1)p}} \right\} = \frac{1}{\theta}\). \hspace{1cm} (11)

Denote the value of \(\theta\) satisfying (11) by \(k(\alpha, p)/p\). Unless \(\theta \to \infty\) as \(p \to 0\), the relative error is unbounded. In fact if \(f_{\theta_p}\) is any sequence of probability density functions on \([0, 1]\) that are bounded above (say by the constant \(\zeta\)) as \(p \to 0\), and \(h_p(x) = \frac{1}{p} I(x > 1 - p)\), then

\[
\exp\{D_2(h_p; f_{\theta_p})\} = \frac{1}{p^2} \int_{1-p}^{1} \frac{1}{f_{\theta_p}(x)}dx \geq \frac{1}{p^2} \frac{p}{\zeta} \to \infty.
\]

From (11), the limiting value \(k(\alpha) = \lim_{p \to 0} k(\alpha, p)\) satisfies

\[
\frac{\alpha}{\alpha - 1} (1 - e^{-(\alpha - 1)k}) = k
\]

and gives a value of \(k\) between 2 (at \(\alpha = 1\)) and 1 (as \(\alpha \to \infty\). See Figure 1. (FIGURE ONE ABOUT HERE) (When \(\alpha = 2\), the solution of \(2(1 - e^{-k}) = k\) is \(k(2) \approx 1.5936\). Substituting \(\theta = k(\alpha)/p\) in the expression (10),

\[
\exp\{D_2(ch; f_\theta)\} = \frac{(1 - e^{-\theta})(e^{\theta p} - 1)}{\theta^2 p^2} \sim \frac{e^{k(\alpha)} - 1}{k^2(\alpha)}.
\]
This function is bounded, and achieves a minimum value for the value $k(2)$. The relative error is graphed in the case $n = 1$ in Figure 2 for $p = 0.01, 0.001$, and $0.0001$. (FIGURE 2 ABOUT HERE) The minimum relative error is obtained when $\alpha = 2$ because this corresponds to the minimum variance IS estimator and the value there is approximately $\sqrt{\frac{e^{k(2)} - 1}{k(2)^2}} - 1 \approx 0.738$ for all three values of $p$. This should be compared with the corresponding relative error of the crude Monte Carlo estimator $\sqrt{(1 - p)/p} \approx 10, 32, 100$ respectively so the gain in efficiency over crude Monte Carlo in the last case $p = 0.00001$ is (as a ratio of variances) around $(100/0.738)^2 \approx 18400$. A crude Monte Carlo estimator with sample size $1.84 \times 10^{10}$ has about the same precision as this IS estimator having sample size $10^6$.

The minimum relative error is achieved when we design our importance sampling distribution using $D_2(ch; f_0)$. What if we minimize $D_\alpha(ch; f_0)$ with $1 < \alpha < 10$ for $\alpha \neq 2$ and choose $\theta = k(\alpha)/p$? The deterioration in the value of the relative error is small (less than 10% over its minimum value) and the variance of the importance sampling estimator is quite insensitive to the value of $\alpha$ used in choosing $\theta$. Furthermore, for the different values of $p$ in Figure 2, the relative errors are very close. For small $p$, the $D_\alpha$-optimal $\theta$ is $\hat{\theta}_{\alpha p} \sim \frac{k(\alpha)}{p}$, from which we obtain an approximation to the relative error of the IS estimator valid when $p \to 0$,

$$\hat{RE} \sim \frac{1}{k(\alpha)} \sqrt{e^{k(\alpha)} - 1 - k^2(\alpha)}.$$  

When this is plotted, the graph is essentially coincident with the curve for $p = 0.0001$ in Figure 2, indicating that the approximation is very good over the considered range of values of $\alpha$. To confirm that these asymptotics provide a very close approximation to the optimal parameter $\theta_{\alpha p}$ we compare below a few of the optimal values of $\theta_{\alpha p}$ to $\frac{k(\alpha)}{p}$ when $p = 0.0001$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{\alpha p}$</td>
<td>15938</td>
<td>14109</td>
<td>13068</td>
<td>12412</td>
<td>11969</td>
</tr>
<tr>
<td>$\frac{k(\alpha)}{p}$</td>
<td>$\frac{15935}{0.0001}$</td>
<td>14106</td>
<td>13068</td>
<td>12412</td>
<td>11969</td>
</tr>
</tbody>
</table>

This is not the only exponential family of distributions providing a bounded relative error for this problem. If we choose $T(X) = \ln(X)$, then the IS distribution $f_\theta$ is a Beta$(1 + \theta, 1)$ density with

$$E_\theta(T(X)) = (1 + \theta) \int_0^1 x^\theta \ln(x) dx = -\frac{1}{1 + \theta}.$$  

To estimate $p = P(X > 1 - p)$, from (5), the optimal parameter $\theta$ satisfies

$$\int_{1-p}^1 x^{-(\alpha-1)}(\ln(x) + \frac{1}{1 + \theta}) dx = 0 \quad (12)$$
giving the solution
\[ \theta_{1p} = \left( 1 + \frac{(1-p) \ln(1-p)}{p} \right)^{-1} - 1 \sim \frac{2}{p} \text{ as } p \to 0 \text{ in case } \alpha = 1, \text{ and} \]
\[ \theta_{\alpha p} \text{ solves } \frac{\alpha \theta}{((\alpha - 1)\theta - 1)(1 + \theta)} (1 - (1-p)^{(\alpha-1)\theta-1}) = -\ln(1-p) \text{ when } \alpha > 1. \]

A solution to (13) of the form \( \theta = \frac{k(\alpha, p)}{p} \) must solve, for \( k = k(\alpha, p) \),
\[ \frac{\alpha k}{((\alpha - 1)k - p)(p + k)} (1 - (1-p)^{(\alpha-1)(k/p-1)}) = \frac{-\ln(1-p)}{p}. \]
The limiting value of the solution \( k(\alpha) = \lim_{p \to 0} k(\alpha, p) \) satisfies \( \frac{\alpha}{\alpha - 1} (1 - \exp^{-(\alpha-1)k(\alpha)}) = k(\alpha) \) which is exactly the same equation we obtained earlier using the standard exponential tilt with \( T(x) = x \). The use of the beta distribution with parameter \( \theta = k(\alpha, p) \) is equivalent to the use of the standard exponential tilt, with the same limiting (bounded) relative error and the same parameter value \( \theta \). We will show in Corollary 1 that this is generally true for any function \( T(x) \) which, after a possible linear transformation, is tail-equivalent to the cumulative distribution function \( F \) in the sense that \( T(x) \sim F(x) \).

Example 2. (exponential distribution) Suppose \( X \) has an exponential(1) distribution and we use the standard exponential tilt with \( T(X) = X \) to estimate \( p = P[X > \ln p] \). Then the IS distribution
\[ f_\theta(x) = \frac{1}{m(\theta)} e^{\theta x} f(x) \]  
(14)
is also exponential with rate parameter \( 1 - \theta \), so \( E_\theta(T) = \mu(\theta) = \frac{1}{1-\theta} \). In this case the condition (5) for the optimal parameter \( \theta \) is
\[ \int_{-\ln p}^{\infty} \left( x - \frac{1}{1-\theta} \right) e^{-(\alpha-1)\theta x - x} \, dx = 0. \]
The optimal value of \( \theta \) solves the quadratic equation
\[ (\alpha - 1)\theta^2 + (2 - \alpha)\theta + \frac{\alpha \theta}{t} - 1 = 0, \]
where \( t = -\ln p \). The solution
\[ \theta_{\alpha p} = \begin{cases} \frac{4 t + \sqrt{(a-t\alpha+2\alpha)^2 - 4a(t\alpha+\alpha t^2)}}{2t(\alpha-1)} & \text{when } \alpha = 1 \\ \frac{4 t + \sqrt{(a-t\alpha+2\alpha)^2 - 4a(t\alpha+\alpha t^2)}}{2t(\alpha-1)} & \text{when } \alpha > 1 \end{cases} \]
is plotted in Figure 3 for various values of \( p \). For \( p \) small, \( \theta_{\alpha p} = 1 + \frac{1}{\infty} + o\left(\frac{1}{\infty}\right) \) for any value of \( \alpha \geq 1 \). This explains the insensitivity in the figure to the value
of \( \alpha \). (FIGURE 3 ABOUT HERE). The relative errors of the IS estimators with optimized tilting parameters \( \theta_{opt} \) are, for \( n = 1 \),

\[
\sqrt{\frac{1}{p^2} \int_{-\ln p}^{\infty} \frac{f(x)}{f_{\theta_{opt}}(x)} f(x) dx} - 1 = \sqrt{\frac{1}{1 - \theta_{opt}^p} p^{\theta_{opt} - 1} - 1} \\
\sim \left( \frac{e}{2} \right)^{1/2} \sqrt{-\ln p} \text{ for all } \alpha \geq 1.
\]

This example shows that even in the simple exponential example, an exponential tilt does not result in a bounded relative error as \( p \to 0 \). This is not because importance sampling is ill-suited to estimating the tails for the exponential distribution but because of a suboptimal choice of importance sampling distribution. If we use as IS distribution a Gumbel or Type I extreme value p.d.f.

\[ f_{\theta}(x) = \theta \exp\{-\theta e^{-x} - x\}, \text{ where } \theta = \frac{k(2)}{p}, \]

we do get bounded relative error (see Corollary 1)

\[
\sqrt{\frac{1}{p^2} \int_{-\ln p}^{\infty} \frac{f(x)}{f_{\theta}(x)} f(x) dx} - 1 = \sqrt{\frac{e^{k(2)} - 1}{k^2(2)}} - 1 \simeq 0.738.
\]

The standard exponential tilt (14), on the surface ideally suited to the exponential distribution, is surprisingly suboptimal as an importance sampling distribution for the exponential distribution. What about the normal distribution, a natural candidate for an exponential tilt?

**Example 3. (Normal distribution)** Suppose that \( f(x) = \varphi(x) \) is the standard normal probability density function and we use a standard exponential tilt with \( T(X) = X \). Then \( f_{\theta} \) is \( N(\theta, 1) \) probability density function and \( E_{\theta} [T(X)] = \theta \). To estimate \( P(X > z_{1-p}) = p \), the IS estimator is an average of terms like

\[
I[X > z_{1-p}] \left( \frac{\varphi(X)}{f_\theta(X)} \right),
\]

where \( X \) is \( N(\theta, 1) \). Then the expected squared estimator is

\[
\int_{z_{1-p}}^{\infty} \frac{\varphi(x)}{f_\theta(x)} \varphi(x) dx = \int_{z_{1-p}}^{\infty} e^{-\theta x + \theta^2/2} \varphi(x) dx = e^{\theta^2} [1 - \Phi(z_{1-p} + \theta)],
\]

\[
\sim e^{\theta^2} \frac{\varphi(z_{1-p} + \theta)}{z_{1-p} + \theta} \text{ as } p \to 0,
\]

\[
\sim \frac{1}{\sqrt{2\pi}} \exp\left( \frac{1}{2} \theta^2 - \theta z_{1-p} - \frac{1}{2} z_{1-p}^2 \right) \frac{1}{z_{1-p} + \theta}.
\]
Minimizing this over $\theta$ yields $\theta_{p} = z_{1} - p + \Delta \sim z_{1} - p$ where $\Delta = \theta_{p} - z_{1} - p = \frac{1}{z_{1} - p + \theta_{p}} \to 0$. Therefore

$$
\int_{z_{1} - p}^{\infty} \frac{\varphi(x)}{f_{\theta_{p}}(x)} \varphi(x) dx \sim \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}(z_{1} - p + \Delta)^{2} - (z_{1} - p + \Delta)z_{1} - p - \frac{1}{2} z_{1}^{2} - p\right) x_{1} - p \to 0
$$

$$
\sim \frac{1}{\sqrt{2\pi}} \exp(-z_{1}^{2} - p) \sim \frac{1}{\sqrt{2\pi}} \frac{2\pi z_{1}^{2} - p}{2z_{1} - p} = \sqrt{\frac{\pi}{2}} z_{1} - p \to 0,
$$

since $p = 1 - \Phi(z_{1} - p) \sim \frac{\varphi(z_{1} - x)}{z_{1} - p}$ implies $\exp(-z_{1}^{2} - p) \sim 2\pi z_{1}^{2} - p$. Therefore the relative error is

$$
\sqrt{\frac{1}{p^{2}} \int_{z_{1} - p}^{\infty} \frac{f_{0}(x)}{f_{\theta_{p}}(x)} f_{0}(x) dx - 1} \sim \sqrt{\frac{1}{p^{2}} \int \frac{2\pi z_{1}^{2} - p}{2z_{1} - p} - 1}
$$

$$
\sim \left( \frac{\pi}{2} \right)^{1/4} z_{1}^{1/2} \to \infty \text{ as } p \to 0.
$$

Once again the exponential tilt results in a sub-optimal importance sampling estimator, one with unbounded relative error. On the other hand, if $T(x) \sim x^{-1}e^{-x^{2}/2}$ as $x \to \infty$ in (1), we do obtain a bounded relative error IS estimator.

Both questions and conclusions emerge from these examples. Somewhat surprisingly, the standard exponential tilt produces suboptimal IS distributions for a large class of distributions, although bounded relative error IS estimators are possible. As Proposition 2 below shows, bounded relative error is obtained if the statistic $T(x)$ imitates the tail behaviour of the negative survivor function $-F = F - 1$ up to a linear transformation. To see why this might be plausible, suppose we wish to use an exponential family of importance distributions for simulating the rare event probability $p = P(X > t)$. Assuming $F$ is continuous so that $U = 1 - F(X)$ is $U[0, 1]$. Both the standard exponential tilt (i.e. an exponential IS distribution) and the beta$(1 + \theta, 1)$ IS distribution lead to bounded relative errors with the same asymptotic parameter value. Generating $U$ using the beta$(1 + \theta, 1)$ importance distribution with $\theta = k(2)/p$ and then obtain $X$ by inverse transform $X = F^{-1}(1 - U)$ is equivalent to using, as IS distribution, the c.d.f.

$$
(1 - F(X))^{\theta} = (1 - F(X))^{k(2)/p}.
$$

This is the cumulative distribution function of the maximum of a sample of size $\theta$ from the original distribution. For values of $X$ in the tail, $F(X)$ is small and

$$
(1 - F(X))^{k(2)/p} \approx \exp\left\{ -\frac{k(2)}{p} F(X) \right\}
$$

having probability density function

$$
c \exp\left\{ \frac{k(2)}{p} (F(x) - 1) \right\} f(x),
$$

10
an exponential family that is obtained from the original density \( f \) using \( T = F - 1 \). Since \( \frac{k(2)}{p} \to \infty \), we expect any function \( T \) with the same tail behaviour as \( F \) to result in an importance distribution with bounded relative error. This provides some intuitive support for the following results. For simplicity in the following we assume that \( x_F \leq \infty \) denotes a point such that \( P(X \leq x_F) = 1 \) and \( P(X \leq t) < 1 \) for all \( t < x_F \). Our asymptotics will apply as \( t \to x_F \) (approaching from the left) or equivalently as \( p_t = P(X > t) \to 0 \) and we will assume that \( F(x) \) is continuous at \( x_F \).

**Notation.** The notation \( a_x \lesssim b_x \) (as \( x \to x_F \)) means that \( a_x = O(b_x) \) and \( a_x \asymp b_x \) means there exist constants \( c_1, c_0 \) and \( x_0 < x_F \) such that

\[
0 < c_1 \leq \frac{a_x}{b_x} \leq c_0 < \infty \text{ for all } x_F > x > x_0.
\]

**Proposition 2.** Suppose we wish to estimate \( p_t = P(X > t) \) using an importance sampling p.d.f. of the form (1). Moreover suppose \( T(x) \) is non-decreasing in \( x \) and for some real number \( a \), \( T(x) = a \cdot F(x) - 1 \) as \( x \to x_F \). Then the family of distributions (1) provides IS estimators with bounded relative error as \( p_t \to 0 \).

The proof of this proposition is in the appendix. The conditions assert that the function \( T = F - 1 \), when translated, behaves like the survivor function \( 1 - F(x) \) because it is bounded above and below by positive multiples of the survivor function. Using importance sampling and guessing the correct tail behaviour pays very large dividends in reduced variance. In fact the proof shows that when \( \theta = k/p \), the limit superior as \( p \to 0 \) of the relative error is bounded by

\[
n^{-1/2} \sqrt{\frac{c_0}{k c_1} (\exp \{ k c_0 / c_1 \} - 1) - 1}.
\]

The following corollary is immediate on letting \( c_0 / c_1 \) decrease to one. With \( k = k(2) \approx 1.594 \) and \( \theta = k(2)/p \), we obtain the value we observed in the uniform case \( \frac{0.738}{\sqrt{n}} \) as \( p \to 0 \).

**Corollary 1.** Suppose \( X \) has a continuous distribution with cumulative distribution function \( F \). Suppose that \( T \) is non-decreasing and that for some real numbers \( a \) and \( c > 0 \) we have \( T(x) = a \sim c(F(x) - 1) \) as \( x \to \infty \). Then the IS estimator for sample size \( n \) obtained from density (1) with \( \theta = k(2)/(cp) \) has bounded relative error asymptotic to approximately \( 0.738n^{-1/2} \) as \( p_t \to 0 \).

A condition like \( T(x) = a \sim c(F(x) - 1) \) as \( x \to \infty \) in Corollary 1 allows us to use in place of \( T \) any linear function of a c.d.f. which is tail-equivalent to \( F \). For example suppose we wish to simulate probability of the tail event \( p_t = P(S_t > t) \) for partial sums \( S_t = \sum_{i=1}^d X_i \) of independent random variables \( X_i \) having
a subexponential or wide-tailed distributions distribution such as the Pareto distribution with c.d.f. \( F(x) = 1 - (1 + x)^{-\beta} \) on \( x > 0 \). For such subexponential distributions, it is well-known that \( P(\sum_{i=1}^{d} X_i > t) \sim P(X_{(d)} > t) \sim d P(X_1 > t) \) as \( t \to \infty \) where \( X_{(d)} \) denotes the largest \( X_i \). Therefore by Corollary 1, the IS distribution which tilts the joint p.d.f. \( \prod_{i=1}^{d} f(x_i) \) of the Pareto random variables using the maximum to produce an IS distribution

\[
c \exp\{-\theta d x_{(d)}^{-\beta}\} \prod_{i=1}^{d} f(x_i),
\]

with \( \theta \simeq \frac{1.594}{p} \) results in bounded relative error with asymptotic value (when \( n = 1 \)) as \( t \to \infty \) around 0.738. It is easy to generate from a probability density function like (15). We generate the maximum value \( x_{(d)} \) from the appropriate tilted distribution and then generate the remaining values independently conditional on being less than the maximum.

Corollary 1 appears to suggest that we should choose \( T = -T \) to be a linear function of a survivor function in the same extreme value domain of attraction as \( F \): If two survivor functions \( F_1 \) and \( F_2 \) are in the Fréchet maximal domain of attraction (see McNeil, Frey and Embrechts (2005), Theorem 7.8), then \( F_1(x) = \mathcal{F}_2(x)L(x) \) where \( L(x) \) is a slowly varying function. This is a weaker condition than requiring tail equivalence. In pursuit of this greater generality, we review some standard results concerning regularly varying functions. For more detail the reader is referred to Bingham et al. (1987).

**Definition 3.** We say the positive measurable function \( f \) is regularly varying at \( 0^- \) (i.e. regularly varying from the left at 0) with index \( \varsigma \) if

\[
\frac{f(x \lambda)}{f(x)} \to \lambda^\varsigma \text{ as } x \uparrow 0 \text{ for all } \lambda > 0.
\]

Then \( f(x) \) is regularly varying at \( 0^- \) if and only if \( f(\frac{1}{y}) \) is a regularly varying function at \( \infty \). The function is said to be slowly varying if (16) holds with \( \varsigma = 0 \). The definitions of regular and slow varying functions at \( \infty \) are similar.

We will assume that all functions here are locally bounded (every point \( x < x_F \) has a neighborhood in which the function is bounded), a consequence, for example, of continuity. A simple example of a function that is regularly varying at \( 0^- \) is the function \( f(x) = (-x)^\varsigma \ln(-1/x) \). For finite \( b \), the function \( f \) is regularly varying at \( b^- \) if \( g(y) = f(b + y) \) is regularly varying at \( 0^- \).

**Lemma 1.** (Karamata’s Theorem: See Mikosch, Theorem 1.2.6): (a) For an arbitrary function \( g \) which is regularly varying at \( \infty \) with index \( \gamma < -1 \),

\[
\int_{t}^{\infty} g(y)dy \sim \frac{t}{|\gamma + 1|} g(t) \text{ as } t \to \infty.
\]
such that for some non-degenerate c.d.f.

It follows from Lemma 2 that

(b) For an arbitrary function \( g \) which is regularly varying at \( 0^- \) with index \( \gamma > -1 \),
\[
\int_{-\varepsilon}^{0} g(y) dy \sim \frac{\varepsilon}{\gamma+1} g(-\varepsilon) \quad \text{as} \quad \varepsilon \downarrow 0.
\]

For the following result, see, for example Bingham et al. pp. 37-38, Athreya and Fidkowski, Lemma 1 and Mikosch, Theorem 1.2.10.

**Lemma 2 (Karamata’s Tauberian Theorem)** Let \( U \) be a right-continuous non-decreasing function on \( \mathbb{R} \) with \( U(x) = 0 \) for \( x < 0 \), with Laplace transform \( \hat{U}(s) = \int_{\mathbb{R}} e^{-su} dU \). Let \( L \) be a slowly varying function at \( \infty \) and let \( c \) and \( \rho \) be non-negative constants. Then:

1. \( U(x) \sim cx^\rho L(1/x)/\Gamma(1+\rho) \) as \( x \to 0^+ \) if and only if \( \hat{U}(s) \sim cs^{-\rho} L(s) \) as \( s \to \infty \).

2. \( U(x) \sim cx^\rho L(x)/\Gamma(1+\rho) \) as \( x \to \infty \) if and only if \( \hat{U}(s) = \int_0^{\infty} e^{-su} dU \sim cs^{-\rho} L(1/s) \) as \( s \downarrow 0 \).

If \( c = 0 \), then above is interpreted to mean that \( U(x) = o(x^\rho L(1/x)/\Gamma(1+\rho)) \) as \( x \to 0^+ \) if and only if \( \hat{U}(s) = o(s^{-\rho} L(s)) \) as \( s \to \infty \).

To use the Tauberian theorem in our context, suppose that \( T(X) \) is a non-positive random variable with \( \hat{F}_T(x) = P(T(X) > x) \), and \( \hat{F}_T(x) = 0 \) for \( x > 0 \). Define \( m(\theta) = E[e^{\theta T(X)}] \). If \( T = -T(X) \), then \( T \) is non-negative, the c.d.f. of \( T \) is \( U(y) = \hat{F}_T(-y) \) and the Laplace transform of \( T \) is
\[
\hat{U}(s) = E[e^{-sT}] = E[e^{\theta T(X)}] = m(s).
\]

It follows from Lemma 2 that \( m(\theta) \sim c \theta^{-\rho} L(s) \) as \( \theta \to \infty \) if and only if \( U(x) = \hat{F}_T(-x) \sim cx^\rho L(1/x)/\Gamma(1+\rho) \) as \( x \to 0^+ \). Therefore, if \( \hat{F}_T(-x) = O(x) \) as \( x \to 0^+ \), then for some \( C < \infty, \hat{F}_T(-x) \leq Cx \) and
\[
m(\theta) = E[e^{\theta T(X)}] = \theta \int_{-\infty}^{0} \hat{F}_T(z)e^{\theta z} dz \leq C \theta \int_{-\infty}^{0} |z|e^{\theta z} dz = O(\frac{1}{\theta}). \tag{17}
\]

Regularly varying functions are closely tied to the maximum domain of attraction of distributions. If there are normalizing constants \( c_n, d_n \) such that such that
\[
F^n(d_n + c_n x) \to H(x)
\]
for some non-degenerate c.d.f. \( H(x) \), then we say that \( F \) is in the maximum domain of attraction (MDA) of the c.d.f. \( H \) and write \( F \in MDA(H) \). The Fisher-Tippet Theorem (see Theorem 7.3 of McNeil et al.) characterizes the possible limiting distributions \( H \) as members of the generalized extreme value distribution (GEV)
\[
H_\xi(x) = \begin{cases} 
\exp(-e^{-x}) & \text{if } \xi = 0 \\
\exp(-(1 + \xi x)^{-1/\xi}) & \text{if } \xi \neq 0, \text{ and } \xi x > -1 
\end{cases}
\]
Special cases of this distribution are the Fréchet ($\xi > 0$), the Gumbel ($\xi = 0$) and the Weibull ($\xi < 0$) distributions. The following standard result shows that the maximum domain of attraction is based on the behaviour of the tail.

**Proposition 3** (see McNeil et al. Section 7.1 and 7.3.2).

(a) $F \in MDA(H_\xi)$ for $\xi > 0$ if and only if $F$ is regularly varying at $\infty$ with index $-1/\xi$.
(b) $F \in MDA(H_\xi)$ for $\xi < 0$ if and only if $x_F < \infty$ and $F$ is regularly varying at $x_F^-$ with index $-1/\xi$.
(c) $F \in MDA(H_\xi)$ for $\xi = 0$ with density $f$ if

$$F(x) \sim -\frac{f^2(x)}{f'(x)} \text{ as } x \to x_F$$

Note that the parameter $\xi$ is not the index of regular variation of the survivor function, but its negative reciprocal. Additionally confusing, $1/\xi$ rather than $-1/\xi$ is sometimes referred to as the tail index of the distribution.

We now return to the problem at hand: we wish to estimate $p_t = P(X > t) = \int_t^\infty f(x)dx$ using IS estimates for large values of $t$ or equivalently small values of $p_t$. Assume $0 < x_F$ since if $x_F \leq 0$ we could simply shift the random variable to accommodate the constraint. We use importance sampling from a p.d.f. $f_\theta$ which is positive on $[0, x_F)$ possibly a proper subset of the support of $f$. The estimator is

$$\hat{p}_t = \frac{f(X)}{f_\theta(X)}I(X > t) \text{ where } X \sim f_\theta(x).$$

For $0 < t < x_F$,

$$E[\hat{p}_t] = \int \frac{f(x)}{f_\theta(x)}I(x > t)f_\theta(x)dx = \int f(x)I(x > t)dx = p_t$$

so $\hat{p}_t$ is an unbiased estimator of $p_t$. Recall that the IS estimator appears to have bounded relative error if $T$ has tails like $F - 1$ or equivalently $T = -T$ has tails like the survivor function $F$. Motivated by this, we might approximate or guess the tail behaviour and adopt an IS distribution based on this guess. Remarkably, as the following result shows, bounded relative error obtains as long as our IS distribution has tails that are sufficiently heavy. The conditions of Proposition 4 imply, for $\xi < 0$, that $\overline{F}$ is regularly varying with index $1/|\xi|$, and $F \in MDA(H_\xi)$ (the Weibull MDA).

**Proposition 4 (Weibull MDA).** Suppose $\xi < 0$, $0 < x_F < \infty$ and $f$ is regularly varying at $x_F^-$ with index $\rho - 1$, with $\rho = -1/\xi$. Consider $\overline{F}(x) = \ldots$
\((x_F - x)^\zeta\), for \(0 < \zeta < 2\rho\). Define the IS p.d.f.

\[
f_\theta(x) = c\zeta \exp\{-\theta(x_F - x)^\zeta\}(x_F - x)^{\zeta - 1}, \text{ for } 0 \leq x < x_F.
\]  

(18)

Suppose \(\theta = \theta_1\) is chosen so that \(\theta_1 \asymp 1/T(t)\) as \(t \to x_F^-\). Then the sequence of distributions \(f_{\theta_t}\) provides importance sample estimators with bounded relative error as \(p_t \to 0\).

This shows that the standard exponential tilt \(T(x) = x_F - x\) or the equivalent, \(T(x) = x\) provides bounded relative error for distributions within the Weibull MDA provided that \(\rho > \frac{1}{2}\) or \(\xi > -2\). Neither the IS family of distributions nor the parameter \(\theta\) need otherwise depend on the underlying density \(f\), provided it is regularly varying at \(x_F^-\) with index \(\rho - 1 > -\frac{3}{2}\).

There is a parallel result that applies in the Fréchet MDA. The conditions of Proposition 5 imply that \(T\) is regularly varying at \(\infty\) with index \(\rho = -1/\xi\) for \(\xi > 0\) and \(F \in MDA(H_\xi)\).

**Proposition 5 (Fréchet MDA)** Suppose that \(f\) is regularly varying at \(\infty\) with index \(\rho - 1\), with \(\rho = -1/\xi < 0\). Consider \(T(x) = (1+x)^{-\xi}\), for \(0 < \zeta < 2/\xi\). Define the IS

\[
f_\theta(x) = \frac{c\zeta}{(1+x)^{\zeta+1}} \exp\{-\theta(1 + x)^{-\zeta}\}, \text{ for } 0 \leq x < \infty,
\]  

(19)

Suppose \(\theta = \theta_1\) is chosen so that \(\theta_1 \asymp 1/T(t)\) as \(t \to x_F^-\). Then the sequence of distributions \(f_{\theta_t}\) provides importance sample estimators with bounded relative error as \(p_t \to 0\).

This shows that a Fréchet tilt \(T(x) = (1+x)^{-1}\) or the equivalent, \(T(x) = 1/x\) provides bounded relative error for distributions within the Fréchet MDA provided that \(0 < \xi < 2\). Again neither IS distribution nor \(\theta\) need otherwise depend on the \(f\), provided it is regularly varying at \(\infty\) with index \(\rho - 1 < -\frac{3}{2}\). Both propositions 4 and 5 are a consequence of the following lemma.

**Lemma 3:** Suppose \(T(x)\) is positive, ultimately non-increasing to 0 as \(x \to x_F^-\), and both \(\frac{1}{T(x)}\) and \(|T'(x)|\) are locally bounded for \(x < x_F\). If \(\int_t^{x_F} \frac{f_{\theta_t}(x)}{|T(x)|} \, dx = O(\frac{p_t^2}{T(t)})\) as \(t \to x_F^-\), then the IS distribution with p.d.f.

\[
f_{\theta_t}(x) = c \exp(-\theta_t T(x))|T'(x)| \text{ for } 0 \leq x < x_F, \text{ for } \theta_t \asymp \frac{1}{T(t)}
\]  

(20)
provides an IS estimator with bounded relative error. The limit is less than or equal to \( e \) if \( \theta_t \sim \frac{1}{T(t)} \).

There is a similar lemma which allows us to handle IS distributions that are tilts of the original density \( f \).

**Lemma 4:** Suppose \( T(x) \) is positive, ultimately non-increasing and \( T(t) = O(p_t) \) as \( t \to x_F \). Suppose that \( P(T(X) > \varepsilon) = O(\varepsilon) \) as \( \varepsilon \downarrow 0 \). Then the IS distribution with p.d.f.

\[
f_\theta(x) = c \exp\{-\theta T(x)\} f(x) \quad 0 \leq x < x_F, \quad \text{and} \quad \theta_t \leq \frac{1}{T(t)} \quad (21)
\]

provides an IS estimator with bounded relative error.

For the Gumbel MDA, it is more difficult to characterize IS distributions with bounded relative error because this class has a greater variety of tail-behaviour. Certainly if we capture this tail behaviour sufficiently closely that Proposition 2 applies, bounded relative error obtains, but these are not the only IS distributions with bounded relative error. In view of the Gumbel limit, it is natural to choose \( T(x) = e^{-x/\beta} \) for some scale parameter \( \beta \), but it seems very likely that we need \( \beta \) to depend on \( p_t \), so our earlier objective to find a single exponential family of distributions with common value of \( T(x) \) is not met. We will leave further discussion on this point to a future paper.

### 3 Examples

We close with two examples of importance sample estimates both with remarkable accuracy. The first is a simulation of the tail behaviour for Tukey’s g&h distribution, used both in insurance operational risk applications (see McNeil et al. (2005) and Embrechts, Degen and Lambrigger (2007)) and in the simulation of wind speed. The g&h distribution is defined as the distribution of

\[
X = \mu + \sigma \frac{e^{gZ} - 1}{g} e^{hZ^2/2}, \quad (22)
\]

where \( Z \) is standard normal and \( g \) and \( h \) govern skewness and elongation of \( X \). The probability density function is inconvenient but it is obviously very easy to simulate values for this distribution. In Dupuis and Field (2004) and Field and Genton (2006), maxima wind speed data from 30 tropical and eight mid-latitude buoys are fit both with the generalized extreme value distribution (GEVD) and the g&h distribution and it is found that the latter gives a better fit.

**Example 4. (Tukey’s g&h distribution)** Suppose \( X = (X_1, X_2) \) has i.i.d. g&h distributed components (22) with \( \mu = 0, g = 0.1, h = 0.2, \) and \( \sigma = 1 \). We wish to simulate the probability \( p = P(X_1 + X_2 > t) \) for \( t \) large, and the
joint distribution of the order statistics \((X_1, X_2)\) given that \(X_1 + X_2 > t\). If \(t = 50\), then \(p \approx \frac{4}{1000000}\), so a crude simulation of 1,000,000 has relative error around 50%. We cannot do importance sampling in the space of \(g\&h\) random variables because the IS weights require the \(g\&h\) density but we can employ importance sampling to modify the distribution of the uniform inputs to the \(g\&h\). For such subexponential random variables, large values of the sum are dominated by the maximum so that tilting the distribution of the maximum \(X_2\) alone provides bounded relative error. Thus we generate \(U_2\) from a beta distribution (or equivalently a standard exponential tilt) and then generate \(U_1\) conditionally uniform on \((0, U_2)\). We generate \(U_2\) under the importance c.d.f.

\[
g_1(u_1 | u_2) g_2(u_2) = 2(1 + \theta)u_2^{\theta+1} I(0 < u_1 < u_2).
\]

Algorithm: For an initial guess at the value of \(p = P[X_1 + X_2 > t]\), put \(\theta = \frac{k(2)}{p}\).

1. Simulate values \((U_1, U_2)\), for \(j = 1, 2, ..., N\) from (23) and corresponding (ordered) standard normal random variables \(Z_j = \Phi^{-1}(U_j)\) and ordered \(g\&h\) random variables:

\[
X_j(1) = e^{gZ_j(1)} - 1 \quad \frac{1}{g} e^{hZ_j(1)/2}, \quad X_j(2) = e^{gZ_j(2)} - 1 \quad \frac{1}{g} e^{hZ_j(2)/2}.
\]

Define the loss on the \(j\)'th simulation \(L_j = X_j(1) + X_j(2)\).

2. Attach IS weights to the \(j\)'th simulation using the ratio of beta density functions

\[
W_j = \frac{2I(U_1 < U_2)}{(2(1 + \theta)U_2^{\theta+1} I(U_1 < U_2))} = \frac{U_2^{-\theta}}{(1 + \theta)}
\]

3. Estimate the probability \(P(X_1 + X_2 > t)\) and any other features of the excess distribution using a weighted average of the values, e.g.

\[
\hat{p} = \frac{1}{n} \sum_{j=1}^{n} W_j I(X_j(1) + X_j(2) > t)
\]
If necessary, we update the parameters $p$, $\theta$ and repeat steps 1-3 above until the relative error is approximately minimized.

With the value of $p = 4 \times 10^{-6}$ determined from a crude simulation of $10^6$ values, the corresponding $\theta \simeq 398.400$, the estimator was $3.785 \times 10^{-6}$ with relative error around 0.8 per simulation (i.e. $8 \times 10^{-4}$ for $10^6$ simulations) reasonably close to the theoretical value of $7.38 \times 10^{-4}$. This also provides a very accurate estimate of the tail behaviour of the loss function and relative errors are close to those experienced in the one-dimensional uniform example.

We plot in Figure 4 the joint distribution of the two components, $(X_{(1)}, X_{(2)})$ given that the loss is greater than 50 with the marker area roughly proportional to the weight on the point. (FIGURE 4 ABOUT HERE)

The conditional survivor function of the sum $X_1 + X_2$ given that $X_1 + X_2 > 50$ can be quite accurately determined by the same simulation and is plotted on a log scale in Figure 5. (FIGURE 5 ABOUT HERE)

The considerable regularity in this graph out to the region where the conditional survivor function is of the order of $10^{-4}$, and so the unconditional survivor function is of the order of $10^{-10}$, is remarkable, and renders a relative error of approximately $8 \times 10^{-4}$ for $10^6$ simulations credible. The estimated probability $3.785 \times 10^{-6}$ was compared to that obtained by performing $2 \times 10^8$ crude simulations of $p$, and the values were close.

In order to test whether importance sampling is indeed providing the “right” answer with the indicated level of precision in spite of possible numerical problems (for example with very large exponents), it is necessary to compare simulated values with the true value in special cases where the probability $p$ is known. This is the case under some parameter values with the generalized skewed normal distribution. For properties of the skewed normal distribution as well as a basic simulation algorithm, see Genton (2004).

**Example 6. (Skewed Normal)** The generalized skewed normal distribution has probability density function

$$2 \text{mvn}(x; \xi, \Omega) \pi(x - \xi)$$

where $\text{mvn}(x; \xi, \Omega)$ is the multivariate normal probability density function with mean $\xi$ and covariance matrix $\Omega$ and where $\pi(x)$ is a skew function with values in $[0, 1]$ such that $\pi(x) = 1 - \pi(-x)$ (which means that $\pi(0) = \frac{1}{2}$). For the simulations we used

$$\Omega = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{and} \quad \xi = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and considered estimating the conditional distribution of the excess $X_1 - t$ and $X_2$ given $X_1 > t$ for various values of the correlation. Note first that in the special case $\rho = 0$ and $\pi(x) = \Phi(x_1)$, the probability density function becomes

$$2 \varphi(x_1) \varphi(x_2) \Phi(x_1)$$
and so
\[ P(X_1 > t) = 2 \int_t^{\infty} \varphi(x_1) \Phi(x_1) dx_1 = 1 - \Phi^2(t). \]

When \( t = 4 \) this gives \( 6.3341 \times 10^{-5} \). We used as an IS distribution on the unit square of uniform inputs a mixture between two beta distributions for \( x_1 \) only, since this is the primary determinant of the event \( X_1 > t \). The corresponding p.d.f. was
\[ \frac{1}{2}(1 + \theta)[x^\theta + (1 - x)^\theta], \text{with } \theta = \frac{1.5936}{\rho/2}. \]

Two simulations when \( \rho = 0 \) of 2 million each gave estimated probabilities \( 6.3325 \times 10^{-5} \) and \( 6.335 \times 10^{-5} \) with an estimated relative error around 0.74, and both of these values are well within 2 standard errors (around 0.1\%) of the true value. Moreover in this case the distribution of \( X_2 - t \) given \( X_2 > t \) is exponential with rate parameter \( t = 4 \), so with mean \( 1/t \). We ran the simulation in this case with \( t = 4 \) and obtained a nearly perfect fit to the exponential and a virtually perfect fit to the (normal) distribution of \( X_2 | X_1 > t \) (see Figure 6). (FIGURE 6 ABOUT HERE). If we modify the skew function to \( \pi(x) = \Phi(x_1 + x_2) \) and \( \rho = 0.8 \), there is now no simple form for the probability \( p \), and we obtain on two simulations \( 6.337 \times 10^{-5} \) and \( 6.336 \times 10^{-5} \) and a comparable relative error around \( 0.737 n^{-1/2} \).

4 Conclusion

Under some conditions on the tail behaviour of the function \( T \) used to generate the importance sampling distribution, we have shown that we can achieve bounded relative error for estimating the probability of tail events. If the tails decrease according to the power law (Fréchet MDA) so that \( T(x) \sim x^\rho, \rho < -\frac{1}{2} \) as \( x \to \infty \) we may use \( T(x) \sim \frac{1}{x} \) and obtain a bounded relative error. If \( f \) is in the Weibull domain of attraction so that \( T(x) \) is slowly varying at \( x_{\infty} < \infty \) with index \( \rho > \frac{1}{4} \), we may use a standard exponential tilt for bounded relative error. If the tails decrease exponentially fast so that \( T(x) \sim \exp(-px) \) then we may choose \( T(x) = e^{-x} \) resulting in a Gumbel IS distribution and bounded relative error. For normally distributed tails, it suffices to use a function \( T(x) \) which is asymptotic to \( |F(x)|^{1/\rho} \), a power of the survivor function. The efficiency of IS with appropriately chosen importance distribution is verified in examples where sample size of less than 5500 often provides a relative error of less than 1\%.

Acknowledgement 1 I am indebted to Guus Balkema, Paul Embrechts and Daniel Alai for extremely insightful comments from on drafts of this paper, and thank Parthanil Roy and Søren Asmussen for conversations related to this work. I am particularly grateful to Paul Embrechts and the FIM in the department of Mathematics of ETH Zürich for providing facilities, support and an extraordinary research environment during my stay at ETH. This work was supported by an NSERC (Canada) grant.
References


5 Appendix

Proof: Proposition 2. By assumption, $T$ is non-decreasing and there exist $a \in \mathbb{R}$, $x_0 < x_F$ with $F(x_0) < 1$ and $c_1, c_0 > 0$ such that

$$a + c_0(F(x) - 1) \leq T(x) \leq a + c_1(F(x) - 1) \text{ for all } x_F > x \geq x_0.$$ 

We wish to show that the relative error (3) is bounded, i.e. that

$$\lim_{p \to 0} \sup \inf \frac{1}{p^2} \int_{-\infty}^{x_F} e^{\theta T(x)} f(x) dx \int_{F^{-1}(1-p)}^{x_F} e^{-\theta T(x)} f(x) dx < \infty.$$ 

Note that under a linear transformation (replacing $T$ by $a + bT$), with $b \neq 0$, the infimum is unchanged since

$$\inf_{\theta} \int_{-\infty}^{x_F} e^{\theta(a+bT(x))} f(x) dx \int_{F^{-1}(1-p)}^{x_F} e^{-\theta(a+bT(x))} f(x) dx = \inf_{\theta} \int_{-\infty}^{x_F} e^{\theta T(x)} f(x) dx \int_{F^{-1}(1-p)}^{x_F} e^{-\theta T(x)} f(x) dx.$$ 

Therefore by replacing $T$ by $\frac{T-a}{c_1}$ and $c_0$ by $c_0/c_1$, we can assume without loss of generality that

$$c_0(F(x) - 1) \leq T(x) \leq F(x) - 1 \text{ for all } x_F > x \geq x_0.$$ 

For some fixed positive constant $k$, consider $\theta = k/p \to \infty$ as $p \to 0$. Then provided $p \leq 1 - F(x_0)$,

$$\int_{F^{-1}(1-p)}^{x_F} e^{-\theta T(x)} f(x) dx \leq \int_{F^{-1}(1-p)}^{x_F} e^{-\theta c_0(F(x) - 1)} f(x) dx$$

$$= \int_{1-p}^{1} e^{-\theta c_0(u-1)} du$$

$$= \frac{1}{\theta c_0} (e^{\theta c_0} - 1).$$

Similarly, since $T(x) \leq F(x) - 1$ for all $x \geq x_0$ and $T(x) \leq T(x_0) \leq F(x_0) - 1$ for $x < x_0$,

$$\int_{-\infty}^{x_F} e^{\theta T(x)} f(x) dx = \int_{-\infty}^{x_0} e^{\theta T(x)} f(x) dx + \int_{x_0}^{x_F} e^{\theta T(x)} f(x) dx$$

$$\leq e^{\theta(F(x_0) - 1)} \int_{-\infty}^{x_0} f(x) dx + \int_{-\infty}^{x_F} e^{\theta(F(x) - 1)} f(x) dx$$

$$= o\left(\frac{1}{\theta}\right) + \int_{0}^{1} e^{\theta(u-1)} du \text{ since } F(x_0) < 1$$

$$= o\left(\frac{1}{\theta}\right) + \frac{1}{\theta} \text{ as } \theta = k/p \to \infty.$$ 

21
Therefore, for \( p \) sufficiently small,
\[
\frac{1}{p^2} \int_{-\infty}^{x_F} e^{\theta T(x)} f(x) dx \int_{F^{-1}(1-p)}^{x_F} e^{-\theta T(x)} f(x) dx \leq \frac{1}{p^2} \left( \frac{1}{\theta} + o\left(\frac{1}{\theta}\right) \right) \frac{1}{\theta c_0} (e^{\theta c_0} - 1) = \frac{1}{k^2 c_0} (e^{k c_0} - 1) + o(1) = O(1).
\]

**Proof of Lemma 1, (b).** Note that \( g \) is regularly varying at \( 0^- \) with index \( \gamma > -1 \) if and only if \( g(\frac{1}{y}) = h(y) \) is a regularly varying function at \( \infty \) with index \(-\gamma\) so with \( x = \frac{1}{y} \), as \( \varepsilon \downarrow 0 \),
\[
\int_{-\varepsilon}^{0} g(y) dy = \int_{1/\varepsilon}^{\infty} h(x) \frac{1}{x^2} dx \\
= \frac{1}{\varepsilon} \int_{1/\varepsilon}^{\infty} h(x) \frac{1}{x^2} dx \\
= \frac{1}{\varepsilon} h(\varepsilon^{-1}) = \frac{1}{\varepsilon + 1} g(-\varepsilon).
\]

**Proof of Lemma 3.** We have assumed that \( T \) is ultimately non-increasing which implies \( \sup \{ T(x); x \geq t \} = O(T(t)) \) as \( t \to x_F^- \). Choose \( z \) large enough that \( T(x) \) is non-increasing for \( x > z \) and \( \inf \{ T(x); x \leq z \} = \delta > 0 \). Then
\[
c(\theta) = \int_0^{x_F} \exp(-\theta T(x)) |T'(x)| dx \leq e^{-\theta \delta} \int_0^{x_F} T'(x) dx + \int_0^{\delta} \exp(-\theta u) du \leq O\left(\frac{1}{\theta}\right) + O\left(\frac{1}{\theta}\right) since \ |T'(x)| \ locally\ bounded.
\]

We assumed that \( \frac{\varphi_T(\theta_t)}{p_t^2} = O\left(\frac{p_t^2}{T(t)}\right) \) so with \( \theta_t > \frac{1}{T(t)} \),
\[
1 + \frac{\varphi_T(\theta_t)}{p_t^2} = \frac{1}{p_t^2} \int_t^{x_F} \frac{f^2(x)}{f_{\theta_t}(x)} dx = \frac{1}{p_t^2} c(\theta_t) \int_t^{x_F} \frac{f^2(x)}{|T'(x)|} e^{\theta_t T(x)} dx \\
\leq \frac{1}{p_t^2} \frac{1}{\theta_t} e^{\theta_t T(t)} \frac{p_t^2}{T(t)} \text{ since } e^{\theta_t T(x)} \leq e^{\theta_t T(t)} \text{ if } t \text{ large, } x > t \\
= \frac{1}{\theta_t T(t)} e^{\theta_t T(t)}.
\]

This is bounded if \( \theta_t > \frac{1}{T(t)} \) and has a limit \( c \) if \( \theta_t \sim \frac{1}{T(t)} \).

**Proof of Lemma 4.** Suppose we choose a sequence \( \theta_t \) such that \( \lim_{t \to x_F^-} \theta_t p_t > 0 \). This implies, since \( p_t \to 0 \), that \( \theta_t \to \infty \). We have shown at (??)
since $\overline{T}(X)$ is non-negative with $P(\overline{T}(X) < \varepsilon) = O(\varepsilon)$ as $\varepsilon \downarrow 0$, that $m(\theta) = E(e^{-\theta T(X)}) = O(\frac{1}{\theta})$ as $\theta \to \infty$. For $t$ sufficiently large, since $T(t)$ is non-increasing, $\sup\{T(x); x \geq t\} \leq T(t)$. Thus

$$1 + \frac{\text{var}_{\theta}(\overline{p}_t)}{p_t^2} = \frac{1}{p_t^2} \int_t^{\infty} \frac{f^2(x)}{f_{\theta_t}(x)} dx = \frac{1}{p_t^2} m(\theta_t) \int_t^{\infty} f(x)e^{\theta_t T(x)} dx \leq \frac{1}{p_t^2} \frac{1}{\theta_t} e^{\theta_t T(t)} \int_t^{\infty} f(x)dx = \frac{1}{\theta_t p_t} e^{\theta_t T(t)} = O\left(\frac{T(t)}{p_t}\right) \text{ with } \theta_t = \frac{1}{T(t)}.$$  

This is bounded provided $\frac{T(t)}{p_t} = O(1)$.

**Proof of Proposition 4.** Assume without loss of generality that $x_F = 1$, so that $\overline{T}(0) = 1$ and $\overline{T}(1) = 0$. We verify the conditions of Lemma 3. Note that $\overline{T}$ is monotonically decreasing on $[0, x_F]$ so that $\sup\{\overline{T}(x); x \geq t\} = \overline{T}(t) = O(T(t))$. Also

$$c(\theta) = \int_0^1 e^{-\theta T(x)}|\overline{T}'(x)|dx = \int_0^1 e^{-\theta u}du = \frac{1}{\theta}(1 - e^{-\theta}) = O\left(\frac{1}{\theta}\right).$$

By assumption, the function $f(x)$ is regularly varying at 1 with index $\rho - 1 > -1$. Therefore, by Lemma 1 (b),

$$p_t = \int_t^1 f(x)dx \sim \frac{1 - t}{\rho} f(t) \to 0 \text{ as } t \to 1.$$  

For $1 > t > 0$, since $\frac{f^2(x)}{|\overline{T}'(x)|}$ is regularly varying at 1 with index $2\rho - 2 - \zeta + 1 = 2\rho - \zeta - 1 > -1$,

$$\int_t^1 \frac{f^2(x)}{|\overline{T}'(x)|} dx \sim \frac{1 - t}{2\rho - \zeta} \frac{f^2(t)}{|\overline{T}'(t)|} \text{ as } t \to 1^- \text{ by Lemma 1(b)}$$

$$= \frac{(1 - t)^2 f^2(t)}{2\rho - \zeta} \zeta T(t) \text{ since } (1 - t)|\overline{T}'(t)| = \zeta T(t)$$

$$= O\left(\frac{p_t^2}{T(t)}\right)$$

The result follows from Lemma 3.

**Proof of Proposition 5.** Note that $\overline{T}$ is positive, decreasing with $\overline{T}(0) = 1$ and $\overline{T}(\infty) = 0$. Again we verify the conditions of Lemma 3.

$$c(\theta) = \int_0^\infty e^{-\theta T(x)}|\overline{T}'(x)|dx = \int_0^1 e^{-\theta u}du = O\left(\frac{1}{\theta}\right).$$
By assumption, the function $f(x)$ is regularly varying at $\infty$ with index $\rho - 1 < -1$. Therefore, by Lemma 1,

$$p_t = \int_t^\infty f(x)dx \sim \frac{1}{|\rho|} t f(t) = \xi t f(t) \to 0 \text{ as } t \to \infty.$$ 

Since $\frac{f^2(x)}{|T'(x)|}$ is also regularly varying at $\infty$ with index $2\rho - 2 + \zeta + 1 = 2\rho + \zeta - 1 < -1$, we have

$$\int_t^\infty \frac{f^2(x)}{|T'(x)|} dx \sim \frac{t}{|2\rho + \zeta|} \frac{f^2(t)}{|T'(t)|} \text{ as } t \to \infty$$

$$= O\left(\frac{p_t^2}{t|T'(t)|}\right) = O\left(\frac{p_t^2}{T(t)}\right).$$

If $\theta = \theta_t$ is chosen so that $\theta_t \asymp 1/T(t)$ as $t \to \infty$ then by Lemma 3, the relative error is bounded.
Figure 1: The value of \( k(\alpha) \) as a function of \( \alpha \).

Figure 2: \( n^{1/2} \times \text{Relative error} = \sqrt{\exp\{D_2(cfg,f_{\theta,\gamma})\} - 1} \) as a function of \( \alpha \) for \( \gamma = 0.99, 0.999, 0.9999 \).
Figure 3: The dependence on $\alpha$ of the min $D_\alpha$ value of parameter for the exponential distribution with $e^{-\gamma} = p$.

Figure 4: Simulated distribution of $(X(1), X(2))$ given $X_1 + X_2 > 50$ for the g&h distribution.
Figure 5: Monte Carlo Estimate of $P[X_1 + X_2 > x | X_1 + X_2 > 50]$ for g&h distributed random variables

Figure 6: (a) Log conditional Survivor Distribution for Skewed normal (b) QQ plot for conditional Normal distribution of $X_2 | X_1 > t$