BOUNDED RELATIVE ERROR IMPORTANCE SAMPLING AND RARE EVENT SIMULATION

BY

DON L. MCLEISH

ABSTRACT

We consider estimating tail events using exponential families of importance sampling distributions. When the cannonical sufficient statistic for the exponential family mimics the tail behaviour of the underlying cumulative distribution function, we can achieve bounded relative error for estimating tail probabilities. Examples of rare event simulation from various distributions including Tukey's g&h distribution are provided.

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 (cdf) F and probability density function (pdf) 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 and the notation $E(*)$ denotes expected value with respect to the probability density function f. We wish to use *importance sampling* (IS) (see for example McLeish (2005) p. 183): generate X from an alternative distribution in an exponential family having probability density

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

where $m(\theta) = \int e^{\theta T(x)} f(x) dx < \infty$ and $\Theta \subset \{\theta \in \mathcal{R}; m(\theta) < \infty\}$. The modification of the original density by the multiplication of a term like $e^{\theta T(x)}$ when

Astin Bulletin 40(1), 377-398. doi: 10.2143/AST.40.1.2049235

$T(x) = x$ is variously referred to in the literature as an exponential *twist* or *tilt* of the density f. We will adopt this language to include a general family of densities of the form (1) and use the phrase *standard exponential twist* in the special case $T(x) = x$. Expected values with respect to the density f_θ are denoted $E_\theta(*)$. Having generated independent X_i, $i = 1, \ldots, n$ from f_θ, the IS estimator of $E(g(X))$ is the unbiased estimator of $E_\theta([g(X)f(X)/f_\theta(X)])$,

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

Our primary concern in this paper is the efficiency or the variance of such estimators in the special case of tail event $g(X) = I(X > t)$ where the probability $P(X > t)$ is small. We define

$$p_t = \bar{F}(t) = \int_t^\infty f(x)dx. \tag{2}$$

For n independent simulations X_i from the pdf f_θ, the IS estimator is

$$\hat{p}_t = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{f_\theta(X_i)}I(X_i > t). \tag{3}$$

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

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

$$var_\theta(\hat{p}_t) = n^{-1}\left[m(\theta)\int_t^\infty e^{-\theta T(x)}f(x)dx - p_t^2\right].$$

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 Degen, Embrechts, 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 θ for the importance sampling distribution (1). In Asmussen and Glynn (2007) a number of estimators similar to (3) 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 θ, and 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 because in estimating very small probabilities such as 0.0001, it is not the absolute size of the error that matters but the size of the error relative to the true value. This motivates the notion of relative error.

Definition 1. *The relative error (RE) 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 pdf f_θ, the RE of the IS estimator (3) is

$$n^{-1/2}\sqrt{\frac{m(\theta)}{p_t^2}\int_t^\infty e^{-\theta T(x)} f(x)\,dx - 1}\,. \tag{4}$$

Define the Rényi generalized divergence $D_\alpha(g;f_\theta)$ of order α, for two probability density functions $g(x)$ and $f_\theta(x)$ (see Rényi, (1961))

$$D_\alpha(g;f_\theta) = \begin{cases} \int \ln\left(\dfrac{g(x)}{f_\theta(x)}\right) g(x)\,dx & \text{if } \alpha = 1, \\ \dfrac{1}{\alpha-1}\ln\left(\int\left[\left(\dfrac{g(x)}{f_\theta}\right)^{\alpha-1} g(x)\,dx\right]\right) & \text{if } \alpha > 0, \alpha \neq 1. \end{cases} \tag{5}$$

We assume, of course, the integrals in (5) exist.

For a non-negative integrable function h, $c = c(h)$ denotes a normalizing constant so that the function ch is a pdf. For $h(x) = f(x)I(x > t)$ and $c = 1/p_t$, the RE (4) can be written $n^{-1/2}\sqrt{e^{D_2(ch;f_\theta)} - 1}$. We conclude:

Proposition 1. *The variance of the importance sample estimator (3) is minimized if f_θ is chosen to minimize $D_\alpha(ch;f_\theta)$ where $h(x) = f(x)I(x > t)$, and $\alpha = 2$.*

If the parametric family f_θ contains a density proportional to $f(x)I(x > t)$, then this obviously minimizes (5) because in this case the divergence equals zero, its minimum possible value, for all $\alpha > 0$. 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 $f(x)$, a point to be returned to shortly. The most important special case of the Rényi generalized divergence (5) is the Kullbach-Leibler divergence from f_θ to h corresponding to $\alpha = 1$. Other functions have also been used in the literature replacing D_α (see for example Ridder and Rubinstein (2007)).

The following metaprinciple is often invoked to generate IS estimators. It is based on the idea that the closer we are to the "perfect" IS distribution ch, the more efficient is our estimator.

Minimum Divergence Principle. *To obtain an IS estimator of* $\int h(x)\,dx$*, choose an IS distribution* f_θ *which minimizes the Rényi generalized divergence* $D_\alpha(ch; f_\theta)$ *between the family* f_θ *and the target* ch*.*

Typically α is chosen to be 1 (minimum cross-entropy) or $\alpha = 2$ (minimum variance) for the application of the above, and in many cases the minimization problem suggested by this principle is quite tractable. Now suppose we do not have a single integral $\int h(x)\,dx$ in mind that we wish to estimate, but a whole class of such integrals for $h \in \mathcal{H}$. We would like a single exponential family that performs well for estimating all functions in the class. This is the notion of bounded RE: for each function in $\mathcal{H}$, there is a member of the exponential family that provides finite RE.

Definition 2. *Suppose* H *is a class of non-negative integrable functions* h*. We say the family* $\{f_\theta;\, \theta \in \Theta\}$ *has bounded relative error for the class* H *if*

$$\sup_{h \in \mathcal{H}} \inf_{\theta \in \Theta} D_2(ch; f_\theta) < \infty \tag{6}$$

for $c = c(h)$ *such that* ch *is a pdf.*

Condition (6) says that the orbit of the exponential family passes close enough to every function $h \in \mathcal{H}$ that it has bounded RE. For rare event simulation, it is easy to find a parametric class of distributions which provides bounded RE for estimating the probability of events in the tail. In particular, if $\mathcal{H}$ is the class of functions $I(x > t)f(x)$ for all $t > 0$, we may define the family of densities

$$f_\theta(x) = c_\theta f(x) I(x > \theta), \quad \text{for all } \theta > 0 \tag{7}$$

for normalizing constants c_θ. Then for any $h \in \mathcal{H}$, this family includes a member (choose $\theta = t$) 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. Moreover, as we will see in Section 3, Example 1, the ideal IS distribution for $I(x > t)f(x)$ may perform very badly for $I(x > s)f(x)$ when $s > t$. Our preference is for a single exponential family (1) which passes close enough to every function $h \in \mathcal{H}$ that it has bounded RE. With an exponential family of distributions generated by a single canonical sufficient statistic T, we can easily aggregate information collected at different parameter values as would be required if the parameter θ were to be sequentially updated.

Our objective is therefore to answer the question *Under what circumstances does an exponential family provide bounded relative error for rare event simulations?* The simple answer to this question is quite informative about attempts to provide IS distributions using the standard exponential tilt. Bounded relative error is to be expected when $T(x)$ behaves in the tail like that of the target distribution.

Notation:

- Denote by $x_F = \sup\{x; F(x) < 1\} \leq \infty$. Our asymptotics will apply as $t \to x_F^-$ (approaching from the left) or equivalently as $p_t = P(X > t) \to 0$.
- The notation $a_x \lesssim b_x$ as $x \to x_F^-$ means that $a_x = O(b_x)$ as $x \to x_F$ and $a_x \asymp b_x$ means there exist constants c_1, c_0 and $x_0 < x_F$ such that

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

For simplicity we will always assume the existence of a pdf, so that $F(x) = \int_{-\infty}^{x} f(z)dz$ is continuous at x_F^-.

Proposition 2. *Suppose we wish to estimate $p_t = P(X > t)$ using an importance sampling pdf of the form (1). Moreover suppose $T(x)$ is non-decreasing in x and for some real number a, $T(x) + a \asymp F(x) - 1$ 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 all of the results, including this proposition, is in the appendix. The conditions assert that the function $-T(x)$, when translated by a constant, behaves like the survivor function $\bar{F}(x) = 1 - F(x)$ 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 terms of reduced variance. The following corollary specifies the optimal value of the parameter and the asymptotic value of the RE and obtains from a small modification to the proof of Proposition 2.

Corollary 1. *Suppose X has a continuous distribution with cdf F. Suppose that $T(x)$ is non-decreasing and for some real numbers a and $c > 0$ we have $a + T(x) \sim -c\bar{F}(x)$ as $x \to x_F^-$. Then the IS estimator for sample size n obtained from density (1) with $\theta = \theta_t = \frac{k(2)}{cp_t}$ has bounded RE asymptotic to*

$$\frac{1}{k(2)\sqrt{n}}\sqrt{e^{k(2)} - 1 - (k(2))^2} \simeq \frac{0.738}{\sqrt{n}}$$

as $p_t \to 0$ where, for $\alpha \geq 1$, $k(\alpha)$ is the unique positive solution to the equation

$$e^{-(\alpha-1)k} = 1 - k\left(1 - \frac{1}{\alpha}\right),$$

and $k(2) \simeq 1.5936$.

Corollary 1 appears to suggest that we should choose $T(x)$ to be a linear function of a survivor function $\bar{F}(x)$ but can we weaken this condition substantially? In pursuit of greater generality, we review some standard results concerning regularly varying functions. For more detail the reader is referred to Bingham, Goldie and Teugels (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 ς if*

$$\frac{f(x\lambda)}{f(x)} \to \lambda^{\varsigma} \text{ as } x \uparrow 0 \text{ for all } \lambda > 0. \tag{8}$$

The definitions of regularly and slowly varying functions at ∞ are similar, and $f(x)$ is regularly varying at a point $x_0^- < \infty$ if $f(x + x_0)$ is regularly varying at 0^-. Note that $f(x)$ is regularly varying at 0^- if and only if $f(\frac{1}{-y})$ is a regularly varying function at ∞. The function $f(x)$ is said to be *slowly varying* if (8) holds with $\varsigma = 0$. A simple example of a function that is defined on $(-\infty, 0)$ and is regularly varying at 0^- is $f(x) = (-x)^{\varsigma} \ln(-1/x)$.

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.

Lemma 1.*(a) (**Karamata's Theorem**) For an arbitrary function g which is regularly varying at ∞ with index $\gamma < -1$,*

$$\int_t^{\infty} g(y)\,dy \sim \frac{t}{|\gamma + 1|}\, g(t) \text{ as } t \to \infty.$$

(b) For an arbitrary function g which is regularly varying at 0^- with index $\gamma > -1$,

$$\int_{-\varepsilon}^{0} g(y)\,dy \sim \frac{\varepsilon}{\gamma + 1}\, g(-\varepsilon) \text{ as } \varepsilon \downarrow 0.$$

Part (a) of Lemma 1 is proved in Mikosch (1999), Theorem 1.2.6, and part (b) in the appendix. For the following result, see, for example Bingham, Goldie and Teugels (1987) pp. 37-38, and Mikosch (1999), Theorem 1.2.10.

Lemma 2 (Karamata's Tauberian Theorem). *Let U be a right-continuous non-decreasing function on $\mathcal{R}$ with $U(x) = 0$ for $x < 0$, with Laplace transform $\hat{U}(s) = \int_{[0,\infty)} e^{-su} dU(u)$. Let L be a slowly varying function at ∞ and let c and ρ 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 $\hat{U}(s) \sim cs^{-\rho} L(s) = \Gamma(1+\rho)\ U(\frac{1}{s})$ as $s \to \infty$.*
2. *$U(x) \sim cx^{\rho} L(x)/\Gamma(1+\rho)$ as $x \to \infty$ if and only if $\hat{U}(s) = \int_0^{\infty} e^{-su} dU(u) \sim cs^{-\rho} L(1/s) = \Gamma(1+\rho)\ U(\frac{1}{s})$ as $s \to 0^+$.*

If $c = 0$, then 1 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 $s \to \infty$.

Regularly varying functions are closely tied to the maximum domain of attraction of distributions. If there are sequences of real constants (c_n) and (d_n) where $c_n > 0$ for all n, such that

$$F^{n}(d_n + c_n x) \to H(x) \text{ as } n \to \infty,$$

for some non-degenerate cdf $H(x)$, then we say that F is in the maximum domain of attraction (MDA) of the cdf H and write $F \in MDA(H)$. The Fisher-Tippett Theorem (see Theorem 7.3 of McNeil, Frey and Embrechts (2005)) characterizes the possible limiting distributions H as members of the generalized extreme value distribution (GEV) which have a cdf given by

$$H_{\xi}(x) = \begin{cases} \exp(-e^{-x}) & \text{if } \xi = 0, \\ \exp(-(1+\xi x)^{-1/\xi}) & \text{if } \xi \neq 0 \text{ and } \xi x > -1. \end{cases}$$

Special cases of this distribution are the Fréchet ($\xi > 0$), the Gumbel ($\xi = 0$) and the Weibull ($\xi < 0$) distributions. If two survivor functions $\bar{F}_1$ and $\bar{F}_2$ are in the Fréchet maximal domain of attraction (see McNeil, Frey and Embrechts (2005) Theorem 7.8), then $\bar{F}_1(x) = \bar{F}_2(x)L(x)$ where $L(x)$ is a slowly varying function. Is it possible to weaken the conditions of Corollary 1 to the more natural requirement, from the point of view of extreme-value theory, that $-T(x) = \bar{F}(x)L(x)$ where $L(x)$ is a slowly varying function? The following standard result shows that the maximum domain of attraction is based on the tail behaviour.

Proposition 3. (see McNeil, Frey and Embrechts (2005), Section 7.1 and 7.3.2).

(a) *$F \in MDA(H_{\xi})$ for $\xi > 0$ (Fréchet) if and only if $\bar{F}$ is regularly varying at $x_F = \infty$ with index $-1/\xi$.*

(b) *$F \in MDA(H_{\xi})$ for $\xi < 0$ (Weibull) if and only if $x_F < \infty$ and $\bar{F}$ is regularly varying at x_F^{-} with index $-1/\xi$.*

(c) *Suppose there exists $z < x_F$ such that F is twice differentiable on (z, x_F) with pdf $f = F'$ and $F'' < 0$ in (z, x_F). Then $F \in MDA(H_{\xi})$ for $\xi = 0$ (Gumbel) if*

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

The parameter ξ is not the index of regular variation of the survivor function, but its negative reciprocal and $1/\xi$ is sometimes referred to as the *tail index* of the distribution.

Let us now return to the problem of finding IS estimators of rare event probabilities

$$p_t = \int_t^{x_F} f(x)\,dx$$

which have bounded relative error as $t \to x_F^-$. Assume $0 < x_F$ since otherwise if $x_F \leq 0$ we could simply shift the distribution. We enlarge slightly the family of possible IS distributions, using a more general dominating distribution. Specifically let $h(x)$ be an arbitrary pdf having the same support as f, let $\bar{H}(x) = \int_x^{\infty} h(z)\,dz$ and assume $x_H = \sup\{x; \bar{H}(x) > 0\} = x_F$. Consider the family of IS distributions

$$f_\theta(x) = \frac{1}{m(\theta)} \exp(\theta T(x))\, h(x) \text{ for } -\infty \leq x < x_F \tag{9}$$

for some statistic $T(x)$ which is non-decreasing and bounded above, where $m(\theta) = \int_{-\infty}^{x_F} e^{\theta T(x)} h(x)\,dx$. Without loss of generality we can assume $T(x)$ is negative and $T(x) \uparrow 0$ as $x \to x_F^-$. The following lemma links the tail behaviours of $h(x), f(x)$, $T(x)$ to the existence of bounded relative error IS estimators.

Lemma 3. *Suppose* $\theta_t \asymp -\frac{\rho}{T(t)}$ *where:*

1. $T(t)$ *is a negative strictly increasing continuous function on* $(-\infty, x_F)$ *such that* $\lim_{x \to x_F^-} T(x) = 0$.
2. *There exists a right continuous, non-decreasing function* $g(y)$ *of regular variation at* 0^+ *with index* $\rho > 0$ *such that*

$$\bar{H}(t) \lesssim g(-T(t)) \text{ as } t \to x_F^-. \tag{10}$$

3. $$E\left[\frac{f(X)}{h(X)}\middle| X > t\right] \lesssim \frac{\bar{F}(t)}{g(-T(t))} \text{ as } t \to x_F^-. \tag{11}$$

Then the IS distributions (9) with $\theta = \theta_t$ *provide a sequence of IS estimators with bounded relative error.*

Lemma 3 allows us to use a more convenient dominating measure than f to construct the IS distributions provided that the tail behaviour of $\bar{H}$ is dominated by a regularly varying function and, from (11), the tails of f and h are similar. The following result shows, in the case of the Weibull MDA, that bounded relative error obtains as long as the IS distribution has tails that are sufficiently heavy. Somewhat surprisingly, we do not need tail equivalence to obtain bounded relative error.

Proposition 4 (Weibull MDA). *Suppose* $0 < x_F < \infty$ *and* f *is supported on the interval* $(0, x_F)$ and is *regularly varying at* x_F^- *with index* $\rho - 1$, *with* $\rho > 0$. *Define the IS pdf*

$$f_\theta(x) = c\exp\{-\theta(x_F - x)^\zeta\}(x_F - x)^{\zeta-1}, \text{ for } 0 \le x < x_F \text{ and } 0 < \zeta < 2\rho \quad (12)$$

where c is a normalizing constant. Suppose $\theta_t \asymp (x_F - t)^{-\zeta}$. *Then the sequence of distributions* f_{θ_t} *provides importance sample estimators with bounded relative error as* $p_t \to 0$.

This shows that the standard exponential tilt ($\zeta = 1$) provides bounded relative error for distributions within the Weibull MDA provided that $\rho > \frac{1}{2}$. This includes the beta(α,β) family of distributions for all $\beta > \frac{1}{2}$, so in particular it includes the uniform distribution discussed in Example 2.

There is a parallel result that applies in the Fréchet MDA. The conditions of Proposition 5 imply that $\bar{F}$ is regularly varying at ∞ with index $\rho = -1/\xi$ for $\xi > 0$ and $F \in MDA(H_\xi)$. Once again the IS distributions take the form

$$ce^{-\theta\bar{H}(x)}h(x)$$

for some density $h(x)$ and the assumptions require that tails of h are sufficiently heavy.

Proposition 5 (Fréchet MDA) *Suppose that f is regularly varying at* ∞ *with index* $\rho - 1 < -1$. *Define the IS pdf*

$$f_\theta(x) = \frac{c\zeta}{(1+x)^{\zeta+1}}\exp\{-\theta(1+x)^{-\zeta}\}, \text{ for } 0 \le x < \infty, \text{ and } 0 < \zeta < -2\rho \quad (13)$$

where c is the normalizing constant. Suppose $\theta = \theta_t$ *is chosen so that* $\theta_t \asymp t^\zeta$ *as* $t \to \infty$. *Then the sequence of distributions* f_{θ_t} *provides importance sample estimators with bounded relative error as* $p_t \to 0$.

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. If we are able to approximate the tail behaviour so that Proposition 2 applies, for example for the $N(0,1)$ distribution, if we choose $T(x) \sim -x^{-1}e^{-x^2/2}$ *as* $x \to \infty$ in (1), we can obtain a bounded relative error IS estimator. In general, if we choose $T(t) \asymp -\bar{F}(t)$ as $t \to x_F^-$ and $\theta_t \asymp \frac{1}{\bar{F}(t)}$, the IS distributions (1) provide a sequence of IS estimators with bounded relative error. For the Gumbel class, it is more difficult to find bounded relative error IS distributions with tail behaviour different from that of the underlying distribution $f(x)$. Fortunately, distributions in the Gumbel class have smaller tails anyway, and so simulating tail behaviour is generally less difficult.

3. Examples

Example 1. (Normal distribution) We mentioned earlier that IS distributions such as (7), designed to estimate $P(X > t)$ for a specific t, may be highly inefficient when t is replaced by $s > t$. Suppose for example X follows a $N(0, 1)$ distribution. The zero-variance importance sample distribution (7) is

$$\frac{1}{1-\Phi(t)}\,\varphi(x)I(x>t) \tag{14}$$

where φ and Φ are the standard normal pdf and cdf respectively. However suppose we wish to estimate as well $P(X > t+s\,|\,X > t) \sim e^{-st-\frac{s^2}{2}}$ as $t \to \infty$, for $s > 0$ fixed. Sampling from (14) is highly inefficient since for n simulations from pdf, the RE for estimating $P(X > t+s\,|\,X > t)$ is approximately $n^{-1/2}\sqrt{e^{st+\frac{s^2}{2}}-1}$ and this grows extremely rapidly in both t and s. We would need a sample size of around $n = 10^4 e^{st+\frac{s^2}{2}}$ (or about 60 trillion if $s = 3$ and $t = 6$) from this IS density to achieve a RE of 1%.

Is it possible to achieve bounded relative error using the usual standard exponential tilt with $T(x) = x$? Then $f_\theta(x) = \frac{1}{m(\theta)}e^{\theta} \times f(x)$ is the $N(\theta,1)$ probability density function. To estimate $P(X > t) = p$, the expected squared value of the IS estimator for sample size $n = 1$ is, after minimizing over the parameter value θ, asymptotic to $n^{-1}\sqrt{\frac{\pi}{2}}\,tp^2$ and the relative error is $\sim (\frac{\pi}{2})^{1/4}\sqrt{t/n} \to \infty$ as $p \to 0$. In other words, the standard exponential tilt **does not** provide IS estimators with bounded relative error, although Proposition 2 verifies that such exponential families do exist.

Example 2. (Uniform distribution) Suppose that $f(x) = 1$, for $0 < x < 1$ and we wish an IS estimator of $p_t = 1 - t$ for $t \to 1^-$. For simplicity assume $n = 1$. The standard exponential tilt, with $T(x) = x$, results in the pdf;

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

Then with $ch(x) = \frac{1}{p}I(x > 1-p)$, it is easy to check that

$$\exp\{D_\alpha(ch;f_\theta)\} = \frac{(1-e^{-\theta})^{\alpha-1}(e^{(\alpha-1)\theta p}-1)}{(\theta p)^\alpha(\alpha-1)}. \tag{15}$$

(15) is minimized when θ satisfies

$$\frac{\alpha}{\alpha-1}\left(\frac{\theta}{e^\theta-1} + \frac{\theta p}{1-e^{-(\alpha-1)\theta p}}\right) = 1. \tag{16}$$

Unless $\theta \to \infty$ as $p \to 0$, the RE is unbounded. Denote the value of θp satisfying (16) by $k(\alpha, p)$. The limit $k(\alpha) = \lim_{p\to 0} k(\alpha, p)$ satisfies

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

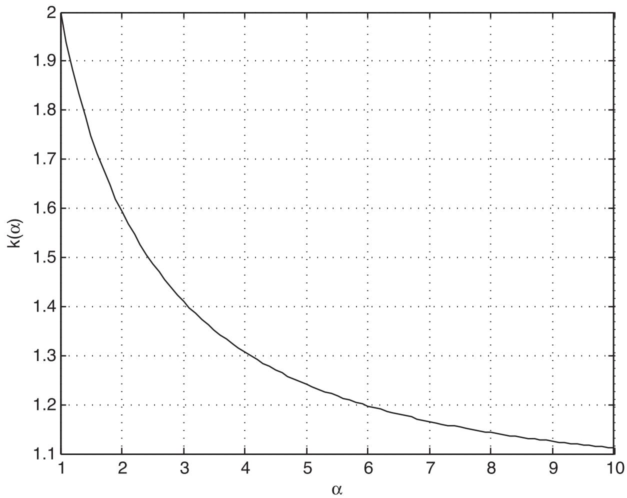

FIGURE 1: The value of $k(\alpha)$ as a function of α.

and gives a value of $k = k(\alpha)$ between 2 (at $\alpha = 1$) and 1 (as $\alpha \to \infty$) with $k(2) \simeq 1.5936$. See Figure1. Substituting $\theta = k(\alpha)/p$ in (15) as $p \to 0$,

$$\exp\{D_\alpha(ch; f_\theta)\} \sim \frac{e^{k(\alpha)} - 1}{k^2(\alpha)},$$

and this leads to the RE

$$RE \sim \frac{1}{k(\alpha)}\sqrt{e^{k(\alpha)} - 1 - k^2(\alpha)} \text{ as } p \to 0,$$

a bounded function with its minimum value at $\alpha = 2$. The asymptotic (as $p \to 0$) RE is graphed in Figure 2. The minimum RE is approximately $\sqrt{\frac{e^{k(2)}-1}{k^2(2)} - 1} \simeq 0.738$ for $p \le 0.01$. This compares with the corresponding RE $\sqrt{\frac{1-p}{p}} \simeq 100$ respectively of the crude Monte Carlo estimator when $p = 0.0001$, so the gain in efficiency over crude Monte Carlo (the ratio of variances) in this case is around $(100/.738)^2 \simeq 18400$. An IS estimator having sample size 10^6 is equivalent to a crude Monte Carlo estimator of sample size 1.84×10^{10}.

Figure 2 shows that the minimum relative error is achieved when we choose the IS parameter which minimizes $D_2(ch; f_\theta)$ ($\theta \simeq k(2)/p$). However if we instead minimize $D_\alpha(ch; f_\theta)$ with $\alpha \neq 2$, we continue to have bounded RE, and the resulting increase in the value of the RE is relatively small. The variance of the IS estimator is quite insensitive to the value of α used in the Rényi generalized divergence.

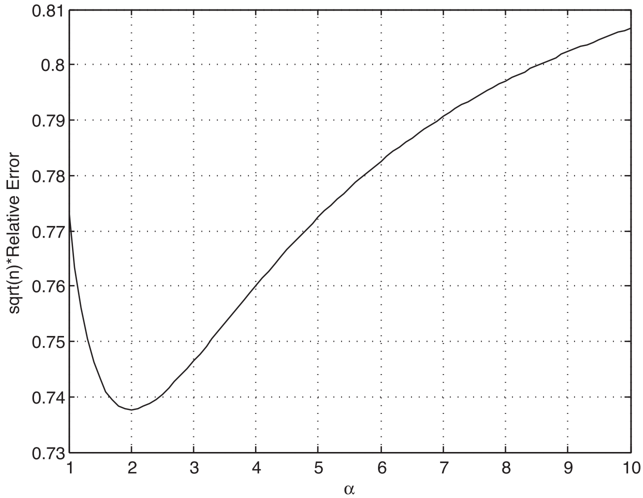

FIGURE 2: $n^{1/2} \times$ Asymptotic Relative error: $\sqrt{\exp\{D_2(ch, f_{\theta_\alpha})\} - 1}$ when $\theta_\alpha = k(\alpha)/p$.

The standard exponential tilt is not the only exponential family of distributions providing a bounded relative error for this problem. By Proposition 2 any family with the same tail behaviour has the same property. For example $T(x) = 1 + \ln(x)$ has similar behaviour to $T(x) = x$ since in both cases $T(x) \sim x$ as $x \to 1$. When $T(x) = 1 + \ln(x)$, the IS distribution f_θ is a Beta$(1 + \theta, 1)$ density. Once again the value of θ which minimizes $D_\alpha(ch; f_\theta)$ is of the form $\theta = \frac{k(\alpha,p)}{p}$ where the limiting value of the solution $k(\alpha) = \lim_{p \to 0} k(\alpha,p)$ satisfies $\frac{\alpha}{\alpha-1}(1 - e^{-(\alpha-1)k(\alpha)}) = k(\alpha)$, exactly the same equation we obtained earlier using the standard exponential tilt. Thus, the beta IS distribution is equivalent to the using standard exponential tilt and also results in bounded RE.

Example 3. (Exponential distribution) Suppose X has an exp(1) distribution, $f(x) = e^{-x}$, for $x > 0$ and we use the standard exponential tilt with $T(x) = x$ to estimate $p = p_t = P(X > t)$, where $t = -\ln p$. Then the IS distribution (1) is exponential with rate parameter $1 - \theta$, for $0 < \theta < 1$. The parameter θ minimizing $D_\alpha(ch; f_\theta)$ solves the quadratic equation

$$(\alpha - 1)\theta^2 + (2 - \alpha)\theta + \frac{\alpha\theta}{t} - 1 = 0,$$

so

$$\theta_{p,\alpha} = \begin{cases} \dfrac{t}{1+t} & \text{if } \alpha = 1, \\ \dfrac{-(\alpha - t\alpha + 2t) + \sqrt{\alpha(\alpha + 4t - 2\alpha t + \alpha t^2)}}{2t(\alpha - 1)} & \text{if } \alpha > 1, \end{cases} \tag{17}$$

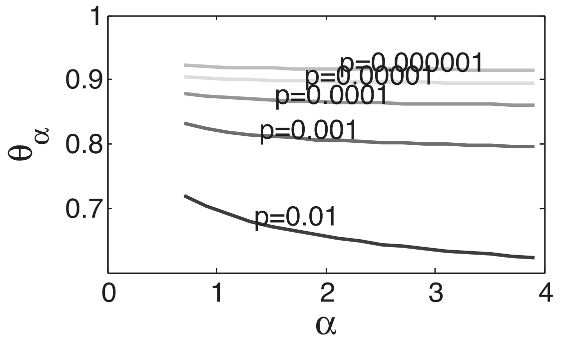

FIGURE 3: The dependence on α of the min D_α value of parameter for the exponential distribution with $e^{-t} = p$.

is plotted in Figure 3 for various values of p. From (17), $\theta_{p,\alpha} = 1 - \frac{1}{t} + o(\frac{1}{t})$ as $p \to 0$ for all $\alpha \geq 1$, explaining the insensitivity in Figure 3 to the value of α. The RE of the IS estimators with parameters $\theta_{p,\alpha}$ are, for n simulations,

$$n^{-1/2}\sqrt{\frac{1}{p^2}\int_{-\ln p}^{\infty}\frac{f(x)}{f_{\theta_{p,\alpha}}(x)}f(x)\,dx - 1} \sim n^{-1/2}\left(\frac{e}{2}t\right)^{1/2} \text{for all } \alpha \geq 1 \text{ as } p \to 0. \quad (18)$$

This implies that the criteria minimizing $D_\alpha(ch; f_\theta)$ for different choices of α all behave equivalently in the tail. However, since $t = -\ln p \to \infty$ as $p \to 0$, it follows from (18) that RE $\to \infty$. A standard exponential tilt does not result in a bounded RE, not because importance sampling is ill-suited to estimating the tails for the exponential distribution but because of a sub-optimal choice of family of IS distributions. If we use instead a Gumbel or Type I extreme value pdf

$$f_\theta(x) = \theta \exp\{-\theta e^{-x} - x\}, \text{ where } \theta = \frac{k(\alpha)}{p},$$

we do get bounded relative error by Corollary 1, and $RE \simeq 0.738n^{-1/2}$ when $\alpha = 2$.

Example 4. (Sums of subexponentially distributed random variables) The condition $T(x) + a \sim -c\bar{F}(x)$ *as* $x \to \infty$ of Corollary 1 allows us to use in place of $T(x)$ any function which is tail-equivalent to $\bar{F}$. Suppose we wish to simulate the probability of the tail event $p_t = P(S_d > t)$ for partial sums $S_d = \sum_{i=1}^{d} X_i$ of independent random variables X_i having a subexponential or wide-tailed distribution such as the Pareto. For such subexponential distributions, it is well-known that

$$P\left(\sum_{i=1}^{d} X_i > t\right) \sim P(X_{(d)} > t) \sim dP(X_1 > t) \text{ as } t \to \infty$$

where $X_{(d)}$ denotes the largest X_i. Therefore by Corollary 1, the IS distribution with pdf

$$c\exp\left\{-\theta d\bar{F}(x_1)\right\}\prod_{i=1}^{d} f(x_i), \tag{19}$$

with $\theta \simeq \frac{1.594}{p}$ results in bounded relative error with asymptotic value as $t \to \infty$ around $0.738n^{-1/2}$. The distribution (19) is equivalent to that obtained by generating the maximum value $x_{(d)}$ (assume using symmetry that $x_{(d)} = x_1$) from the appropriate tilted distribution with pdf $\exp\{-\theta\bar{F}(x_1)\}f(x_1)$ and then generating the remaining values independently from the original density $f(x)$.

For example suppose $f(x)$ is the density of Tukey's g&h distribution, used in insurance operational risk applications (see McNeil, Frey and Embrechts (2005) and Degen, Embrechts and Lambrigger (2007)). $f(x)$ is defined as the pdf of

$$X = \mu + \sigma\frac{e^{gZ}-1}{g}e^{hZ^2/2}, \text{ for } \mu, g, h \in \mathcal{R}, \sigma > 0 \tag{20}$$

where Z is standard normal and g and h are parameters governing the skewness and elongation of X. Typical values of the parameters used by Dutta and Perry (2006) to model operational risk are $g \in [1.79, 2.30]$ and $h \in [0.10, 0.35]$. The probability density function is inconvenient but it is obviously very easy to simulate values for this distribution: for a $U[0,1]$ random variable U, if Φ^{-1} is the inverse standard normal cdf,

$$X(U) = \mu + \sigma\frac{e^{g\Phi^{-1}(U)}-1}{g}e^{h(\Phi^{-1}(U))^2/2} \tag{21}$$

As well as insurance applications, the g&h distribution has been used to model maxima wind speed data (see for example Dupuis and Field (2004), Field and Genton (2006) where it is concluded that the g&h distribution provides a better fit than the generalized extreme value distributions).

Suppose $X = (X_1, X_2)$ has independent identically g&h distributed components as in (20). We chose $d = 2$ for simplicity and $\mu = 0$, $g = 0.1$, $h = 0.2$, $\sigma = 1$. For example if X_1, X_2 are losses from two independent time periods, we might wish to estimate the probability of a large loss, $L(X_1, X_2) = X_1 + X_2$, such as $p = P(L(X_1, X_2) > t)$ for t large. We are also interested in the individual losses given that a large loss has occurred, obtained from the joint distribution of the order statistics $(X_{(1)}, X_{(2)})$ given that $X_1 + X_2 > t$. If $t = 50$, then $p \simeq \frac{4}{1000000}$, so a crude simulation of 1,000,000, which has relative error around 50%, is of little value. As in (21), assume are able to transform uniform[0,1] random variables U to generate a value of X using an increasing transformation, say $X(U)$. For general dimension d, we begin with an initial guess at the tail probability $p = P(\sum_{i=1}^{d} X_i > t)$ and initialize $\theta = \frac{k(2)}{p}$. Since in subexponential distributions such as the g&h distribution, the partial sum S is essentially controlled by the maximum value, we need only apply importance sampling to this random variable. Since $X(U)$ is increasing, we can generate the maximum $U_{(d)}$ from a standard exponential tilt of its original distribution, and then generate $U_{(i)}$, $i = 1, 2, \ldots, d-1$ conditionally uniform on $(0, U_{(d)})$, transforming all to obtain the sample $X(U_i)$, $i = 1, 2, \ldots, d$. There are many asymptotically equivalent alternatives: for example we can generate $U_{(d)}$ either under the beta IS cdf $[F_{U_{(d)}}(u)]^\theta$ where $F_{U_{(d)}}(u) = u^d$, for $0 < u < 1$, and $\theta = \frac{k(2)}{p}$, or under an exponential tilt, with pdf proportional to $\exp(\theta u) u^{d-1}$, $0 < u < 1$. Note that for the former, the cdf of the random variable $Z = \theta d(1 - U_{(d)})$ is $P(\theta d(1 - U_{(d)}) \leq z)$ $= P(U_{(d)} \geq 1 - \frac{z}{\theta d}) = 1 - (1 - \frac{z}{\theta d})^{\theta d} \to 1 - e^{-z}$ for $z > 0$ as $\theta \to \infty$ so that rather than use a beta generator we can transform a standard exponential conditioning the generated value of $U_{(d)}$ to the unit interval. Our simulation, with a couple of other modest simplifications, was conducted as follows:

1. Simulate the random variable $1 - U_d = -\frac{1}{\theta d} \ln(V_1)$ where $V_1 \sim U[e^{-\theta d}, 1]$. Generate U_j, $j = 1, \ldots, d-1$ as independent $U[0,1]$ random variables.
2. Transform the U_i to obtain tilted g&h random variables
$$X_i = X(U_i),\ i = 1, 2, \ldots, d$$
3. Record the value of the sum $S = \sum_{i=1}^d X_i$ as well as the IS weight attached to this simulation,
$$W = \frac{1}{\theta} e^{\theta d(1 - U_d)}$$
4. After n simulations, re-estimate the probability p, the variance and relative error and any other features of the distribution using a weighted average of the values, e.g.
$$\widehat{p}_t = \frac{1}{n} \sum_{j=1}^{n} W I(S > t)$$

Revise the estimate of p and hence θ and repeat if necessary until a desired precision is obtained.

We began with an estimated value of $p = 4 \times 10^{-6}$ determined from a crude simulation of $n = 4 \times 10^6$ values, resulting in $\theta \simeq 398{,}400$. The IS estimate from $n = 4 \times 10^6$ simulations was 3.782×10^{-6} with relative error around 0.83 per simulation (i.e. $\frac{0.83}{2000} \simeq 0.0004$ for 4×10^6 simulations) reasonably close to the theoretical value of approximately $\frac{0.738}{\sqrt{n}}$ of the RE. This provides a very accurate estimate of the tail behaviour of the loss function and relative errors are close to those experienced in the uniform case (Example 2). The total time required for this simulation was about 6 seconds running Matlab on Windows with an Intel Core 4 CPU @2.5 GHZ with 4 GB RAM.

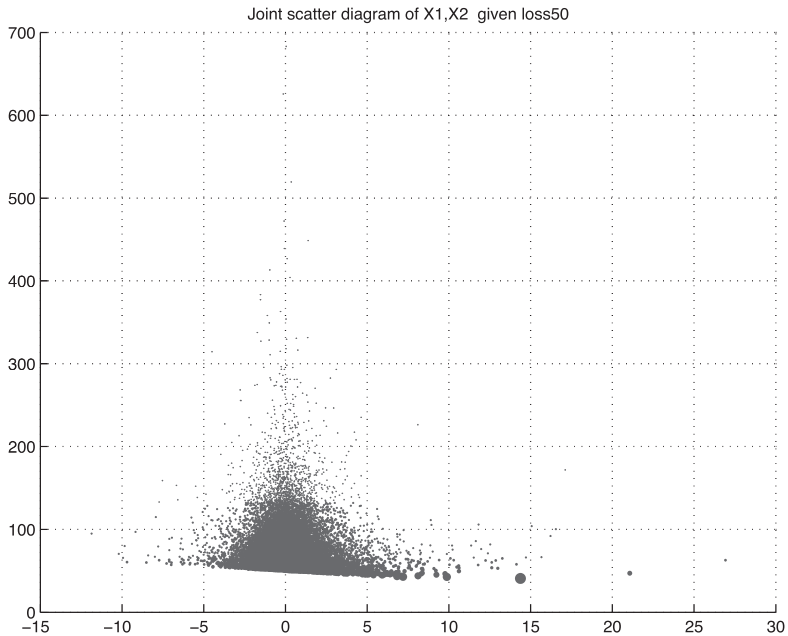

FIGURE 4: Simulated distribution of $(X_{(1)}, X_{(2)})$ given $X_1 + X_2 > 50$ for the g&h distribution.

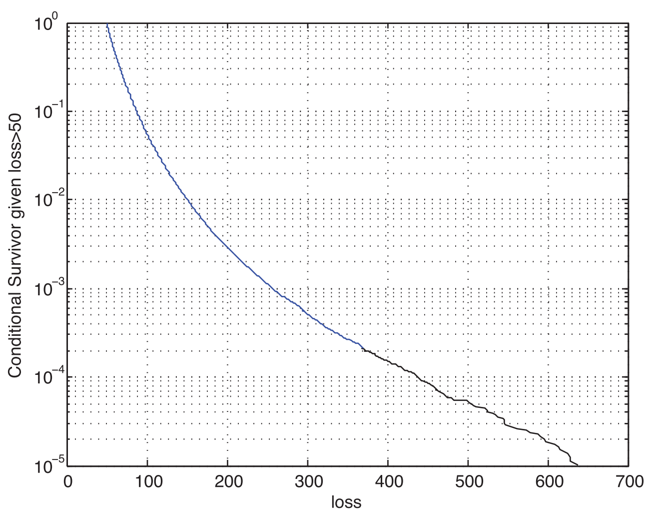

FIGURE 5: Monte Carlo Estimate of $P(X_1 + X_2 > x \mid X_1 + X + 2 > 50)$ for g&h distributed random variables

We plot in Figure 4 the joint distribution of the two components, $(X_{(1)}, X_{(2)})$ given that the loss $X_1 + X_2$ is greater than 50 with the marker area roughly proportional to the weight on the point.

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. 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 4×10^{-4} for 4×10^6 simulations credible. To attempt to validate the simulation, the estimated probability 3.782×10^{-6} was compared to that obtained by performing 2.4×10^9 crude simulations. The crude estimate was $\hat{p} = 3.812 \times 10^{-6}$ with standard error around 3.96×10^{-8}. The crude estimate was not significantly different from the IS estimate, (although this is only a rough verification of the estimate since the time required for the crude simulations was around 51 minutes, achieving a much more precise estimator by crude Monte Carlo is virtually infeasible).

4. Conclusion

We have shown that it is always possible to find an exponential family of distributions which provide bounded relative error for estimating the probability of rare events if $-T(x) \asymp \bar{F}(x)$ and when $-T(x) \sim \bar{F}(x)$, we specify the value of the parameter $\theta = k(2)/p$ that achieves the minimum asymptotic variance. Under regularly varying conditions, and when the target distribution lies in the Fréchet or the Weibull MDA, we are able to achieve bounded relative error when $-T(x)$ has tail behaviour within a substantially wider range than that of $\bar{F}(x)$. The efficiency of IS with appropriately chosen importance distribution is verified in several examples including the g&h distribution. Examples verify that the relative error is typically around its minimum asymptotic value of $0.738n^{-1/2}$ for n simulations.

Acknowledgement

I am very grateful to the referees for extremely detailed comments and corrections as well as Guus Balkema, Paul Embrechts and Daniel Alai for insightful comments on this paper, and Parthanil Roy and Søren Asmussen for conversations related to this work. I particularly thank Paul Embrechts and the FIM in the Department of Mathematics, 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

Asmussen, S. and Glynn, P.W. (2007) *Stochastic simulation: algorithms and analysis*, Springer, New York.

Asmussen, S., Kroese, D.P. and Rubinstein, R.Y. (2005) Heavy tails, importance sampling and cross-entropy. *Stochastic Models*, **21**, 57-76.

BINGHAM, N.H., GOLDIE, C.M. and TEUGELS, J.L. (1987) *Regular Variation*, Encyclopedia of Mathematics and its Applications, Cambridge University Press.
DUPUIS, D.J. and FIELD, C.A. (2004) Large wind speeds: modeling and outlier detection, *J. Agricultural Biol. Environ. Statist.*, **9**, 105-121.
DUTTA, K. and PERRY, J. (2006) A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital. *Federal Reserve Bank of Boston*, Working Paper N° 06-13.
DEGEN, M., EMBRECHTS, P. and LAMBRIGGER, D.D. (2007) The quantitative modeling of operational risk: between g-and-h and EVT. *ASTIN Bulletin* **37(2)**, 265-291.
FIELD, C. and GENTON, M.G. (2006) The multivariate g-and-h distribution *Technometrics*, **48**, 104-111.
HOMEM-DE-MELLO, T. and RUBINSTEIN, R.Y. (2002) Rare event simulation and combinatorial optimization using cross entropy: estimation of rare event probabilities using cross-entropy *Proceedings of the 34th Conference on Winter Simulation: exploring new frontiers*, 310-319.
KROESE, D. and RUBINSTEIN, R.Y. (2008) *Simulation and the Monte Carlo Method*. (Second Edition) Wiley, New York.
MCLEISH, D.L. (2005) *Monte Carlo Simulation and Finance*. Wiley, New York.
MCNEIL, A.J., FREY, R. and EMBRECHTS, P. (2005) *Quantitative Risk Management*, Princeton University Press, Princeton.
MIKOSCH, T. (1999) Regular Variation, Subexponentiality and Their Applications in Probability Theory. EURANDOM Report 99-013.
RÉNYI, A. (1961) On Measures of entropy and information. *Proceedings of the 4th Berkeley symposium on mathematical statistics and probability*. Editor, Neyman, University of California Press, Berkeley, Calif. 547-561.
RIDDER, A. and RUBINSTEIN, R. (2007) Minimum cross-entropy methods for rare-event simulation. *Simulation*, **83**, 769-784.

DON L. MCLEISH
Department of Statistics and Actuarial Science
University of Waterloo
Canada

5. APPENDIX

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

$$c_1 \leq \frac{-a - T(x)}{\bar{F}(x)} \leq c_0 \text{ for all } x_F > x > x_0, \text{ or equivalently}$$
$$a + c_1 \bar{F}(x) \leq -T(x) \leq a + c_0 \bar{F}(x).$$

We wish to show that the relative error (4) is bounded. Defining $\bar{F}^{-1}$ to be the generalized inverse of $\bar{F}$ and for $\varepsilon > 0$ with $t = t(\varepsilon) = \bar{F}^{-1}(\varepsilon)$, we wish to show

$$\lim_{\varepsilon \to 0} \sup \inf_{\theta > 0} \frac{1}{\varepsilon^2} \int_{-\infty}^{x_F} e^{\theta T(x)} f(x)\,dx \int_{t(\varepsilon)}^{x_F} e^{-\theta T(x)} f(x)\,dx < \infty.$$

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

$$\begin{aligned}&\inf_{\theta>0}\int_{-\infty}^{x_F} e^{\theta(a+bT(x))}f(x)\,dx\int_{t(\varepsilon)}^{x_F} e^{-\theta(a+bT(x))}f(x)\,dx\\&=\inf_{\theta>0}\int_{-\infty}^{x_F} e^{\theta T(x)}f(x)\,dx\int_{t(\varepsilon)}^{x_F} e^{-\theta T(x)}f(x)\,dx.\end{aligned}$$

Therefore by replacing T by $T + a$ we can assume without loss of generality that

$$c_1\bar{F}(x) \le -T(x) \le c_0\bar{F}(x) \text{ for all } x_F > x > x_0.$$

Then for $\theta > 0$, provided $\varepsilon \le \bar{F}(x_0)$,

$$\begin{aligned}\int_{t(\varepsilon)}^{x_F} e^{-\theta T(x)}f(x)\,dx &\le \int_{t(\varepsilon)}^{x_F} e^{\theta c_0\bar{F}(x)}f(x)\,dx\\&= \int_0^{\varepsilon} e^{\theta c_0 u}du \text{ since } \bar{F} \text{ is continuous}\\&= \frac{1}{\theta c_0}(e^{\theta c_0\varepsilon}-1).\end{aligned}$$

Consider $\theta \to \infty$ as $\varepsilon \to 0$. Then since $\bar{F}(x_0) > 0$,

$$\begin{aligned}\int_{-\infty}^{x_F} e^{\theta T(x)}f(x)\,dx &= \int_{-\infty}^{x_0} e^{\theta T(x)}f(x)\,dx + \int_{x_0}^{x_F} e^{\theta T(x)}f(x)\,dx\\&\le e^{-\theta c_1\bar{F}(x_0)}\int_{-\infty}^{x_0} f(x)\,dx + \int_{x_0}^{x_F} e^{-\theta c_1\bar{F}(x)}f(x)\,dx\\&= o\left(\frac{1}{\theta}\right) + \int_0^{\bar{F}(x_0)} e^{-\theta c_1 u}du\\&= o\left(\frac{1}{\theta}\right) + \frac{1}{\theta c_1}\left(1-e^{-\theta c_1\bar{F}(x_0)}\right)\\&\sim \frac{1}{c_1\theta}.\end{aligned}$$

Therefore, as $\varepsilon \to 0^+$,

$$\frac{1}{\varepsilon^2}\int_{-\infty}^{x_F} e^{\theta T(x)}f(x)\,dx\int_{\bar{F}^{-1}(\varepsilon)}^{x_F} e^{-\theta T(x)}f(x)\,dx \lesssim \frac{1}{\varepsilon^2}\frac{1}{\theta c_1}\frac{1}{\theta c_0}(e^{\theta c_0\varepsilon}-1).$$

The right side is minimized over θ when $\theta = \frac{k}{c_0\varepsilon}$ where $k = k(2)$, the positive root to the equation $e^{-k} = 1 - \frac{k}{2}$ and the minimum value is then $\frac{c_0}{c_1}\frac{1}{k^2}$ $(e^k - 1)$.

Proof of Corollary 1. We are given that $a + T(x) \sim -c\bar{F}(x)$ and so the conditions of Proposition 2 hold for any pair of values (c_1, c_0) for which $c_1 < c < c_0$. In the proof of Proposition 2 we obtained the asymptotic bound $\frac{c_0}{c_1} \frac{1}{k^2(2)}$ $(e^{k(2)} - 1)$ for the limit superior of the values $1 + \frac{var_\theta(\widehat{p_\varepsilon})}{p_\varepsilon^2}$ when $\theta = \frac{k}{c_0 \varepsilon}$. The result of Corollary 1 follows on letting $c_1 \to c^-$ and $c_0 \to c^+$. ■

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 ∞ with index $-\gamma$ so with $x = \frac{1}{-y}$,, as $\varepsilon \downarrow 0$,

$$\int_{-\varepsilon}^{0} g(y)dy = \int_{1/\varepsilon}^{\infty} h(x) \frac{1}{x^2}\, dx$$

$$\sim \frac{\varepsilon^{-1}}{|-\gamma-1|\,\varepsilon^{-2}} h(\varepsilon^{-1}) \text{ since } h(x)\frac{1}{x^2} \text{ is regularly varying at } \infty \text{ with index } -\gamma-2,$$

$$= \frac{\varepsilon}{\gamma+1} h(\varepsilon^{-1}) = \frac{\varepsilon}{\gamma+1} g(-\varepsilon).$$

Proof of Lemma 3. Assume for simplicity and without loss of generality that $n = 1$. Then as $t \to x_F^-$ so that $\theta_t \to \infty$,

$$\begin{aligned}
1 + \frac{var_\theta(\widehat{p_t})}{p_t^2} &= \frac{E(\widehat{p_t}^2)}{p_t^2} = \frac{1}{p_t^2} E_{\theta_t}\left[I(X > t) \frac{f^2(X)}{f_{\theta_t}^2(X)} \right] \\
&= \frac{m(\theta_t)}{p_t^2} \int_t^{x_F} e^{-\theta_t T(x)} \frac{f^2(x)}{h(x)}\, dx \\
&\lesssim \frac{m(\theta_t)}{p_t^2} e^{-\theta_t T(t)} E\left[\frac{f(X)}{h(X)} I(X > t) \right] \\
&\lesssim \frac{m(\theta_t)}{p_t^2} e^{-\theta_t T(t)} E\left[\frac{f(X)}{h(X)} \middle| X > t \right] \bar{F}(t) \\
&\lesssim \frac{m(\theta_t)}{p_t} e^{-\theta_t T(t)} \frac{\bar{F}(t)}{g(-T(t))} \text{ by (11)} \\
&\lesssim \frac{m(\theta_t)}{g(-T(t))} \text{ since } e^{-\theta_t T(t)} = O(1) \text{ and } p_t = \bar{F}(t).
\end{aligned}$$

We now approximate $m(\theta_t) = \int_{-\infty}^{x_F} e^{\theta_t T(x)} h(x)dx$. Note that this is the Laplace transform evaluated at θ_t of the function

$$\begin{aligned}
U_1(y) &= \int_{\{x; -T(x) \le y\}} h(x)dx = \int_{T^{-1}(-y)}^{\infty} h(x)dx \\
&= \bar{H}(T^{-1}(-y)) \text{ with } T^{-1} \text{the inverse function of } T \\
&\lesssim g(-T(T^{-1}(-y))) \text{ as } y \to x_F^- \text{ by (10)} \\
&= g(y).
\end{aligned}$$

From Karamata's Tauberian theorem, since $g(y) \sim cx^{\rho}L(1/x)/\Gamma(1+\rho)$ as $x \to 0^+$ is a right continuous non-decreasing regularly varying function at 0^+, its Laplace transform $\int e^{-\theta y} dg(y) \sim \Gamma(1+\rho)\, g(\frac{1}{\theta})$ as $\theta \to \infty$ and $U_1(y) \lesssim g(y)$ implies, since $\theta_t > 0$,

$$m(\theta_t) = \int e^{\theta_t T(x)} h(x)\, dx \lesssim g\left(\frac{1}{\theta_t}\right) \text{ as } \theta_t \to \infty.$$

Therefore

$$\frac{m(\theta_t)}{g(-T(t))} \lesssim \frac{g\left(\frac{1}{\theta_t}\right)}{g(-T(t))} = \frac{g\left(\frac{-T(t)}{\rho}\right)}{g(-T(t))} \sim \frac{\rho^{-\rho} g(-T(t))}{g(-T(t))} = O(1).$$

Proof of Proposition 4. We verify the conditions of Lemma 3. Here $h(x) = c_1(x_F - x)^{\zeta-1}$ for $0 < x < x_F$ where $c_1 = \zeta / x_F^{\zeta}$ is the normalizing constant, and $\bar{H}(x) = \frac{c_1}{\zeta}(x_F - x)^{\zeta} = -\frac{c_1}{\zeta} T(x)$ where $T(x) = -(x_F - x)^{\zeta}$ is continuous and strictly increasing to 0 as $x \to x_F^-$ verifying 1. With $g(y) = y$, we have $\bar{H}(x) \lesssim g(-T(x))$ verifying (10). Finally for (11) we need

$$E\left[\frac{f(X)}{h(X)}\middle| X > t\right] \lesssim \frac{\bar{F}(t)}{g(-T(t))} \text{ as } t \to x_F^-$$

or equivalently,

$$E\left[\frac{f(X)}{(x_F - X)^{\xi-1}} I(X > t)\right] \lesssim \frac{\bar{F}^2(t)}{(x_F - t)^{\xi}} \text{ as } t \to x_F^-. \tag{22}$$

Since $f(x)$ and $\frac{f^2(x)}{(x_F - x)^{\zeta-1}}$ are functions of regular variation at x_F^- with indexes $\rho - 1$ and $2\rho - \zeta - 1$ respectively, it follows that

$$\bar{F}(t) = \int_t^{x_F} f(x)\, dx \sim \frac{(x_F - t)}{\rho} f(t)$$

$$\int_t^{x_F} \frac{f^2(x)}{(x_F - x)^{\zeta-1}} dx \sim \frac{(x_F - t)}{(2\rho - \zeta)} \frac{f^2(t)}{(x_F - t)^{\zeta-1}} = \frac{1}{(2\rho - \zeta)} \frac{f^2(t)}{(x_F - t)^{\zeta-2}}$$

$$\text{since } 2\rho - \zeta > 0$$

and so (22) becomes

$$\frac{1}{2\rho - \zeta} \frac{f^2(t)}{(x_F - t)^{\zeta-2}} \lesssim \frac{\bar{F}^2(t)}{(x_F - t)^{\zeta}} \sim \frac{1}{\rho^2 (x_F - t)^{\zeta-2}} f^2(t)$$

and this holds since $2\rho - \zeta > 0$ and $\rho > 0$. ■

Proof of Proposition 5. Again, we verify the conditions of Lemma 3. Here $h(x) = \zeta(1+x)^{-\zeta-1}$ and $\bar{H}(x) = (1+x)^{-\zeta} = -T(x)$ where $T(x) = -(1+x)^{-\zeta}$, for $0 < x < \infty$. $T(x)$ is continuous, increasing to 0 as $x \to \infty$ verifying 1. With $g(y) = y$, we have $\bar{H}(x) \lesssim g(-T(x))$ verifying 2. Finally we need to show

$$E\left[\frac{f(X)}{h(X)}\middle| X > t\right] \lesssim \frac{\bar{F}(t)}{g(-T(t))} \text{ as } t \to \infty, \text{ or} \tag{23}$$

$$E\left[f(X)(1+X)^{\zeta+1} I(X > t)\right] \lesssim \bar{F}^2(t)(1+t)^{\zeta} \text{ as } t \to \infty \text{ or} \tag{24}$$

$$\int_t^\infty f^2(x)(1+x)^{\zeta+1} dx \lesssim \left(\int_t^\infty f(x)\,dx\right)^2 (1+t)^{\zeta} \text{ as } t \to \infty. \tag{25}$$

$f(x)$ and $f^2(x)(1+x)^{\zeta+1}$ are functions of regular variation with index $\rho - 1 < -1$ and $2\rho + \zeta - 1 < -1$ respectively, so using Karamata's Theorem, Lemma 1, the left side of (25) is $\int_t^\infty f^2(x)(1+x)^{\zeta+1} dx \sim \frac{t}{|2\rho+\zeta|} f^2(t)(1+t)^{\zeta+1}$ and $\int_t^\infty f(x)\,dx \sim \frac{t}{|\rho|} f(t)$. On substitution, (25) becomes

$$\frac{t}{|2\rho+\zeta|} f^2(t)(1+t)^{\zeta+1} \lesssim \frac{t^2(1+t)^{\zeta}}{\rho^2} f^2(t)$$

which holds under the conditions as $t \to \infty$. ■