Bounded Relative Error Importance Sampling and Rare Event simulation

Don McLeish
University of Waterloo
(Visiting ETH, Zurich)

Importance sampling (IS)

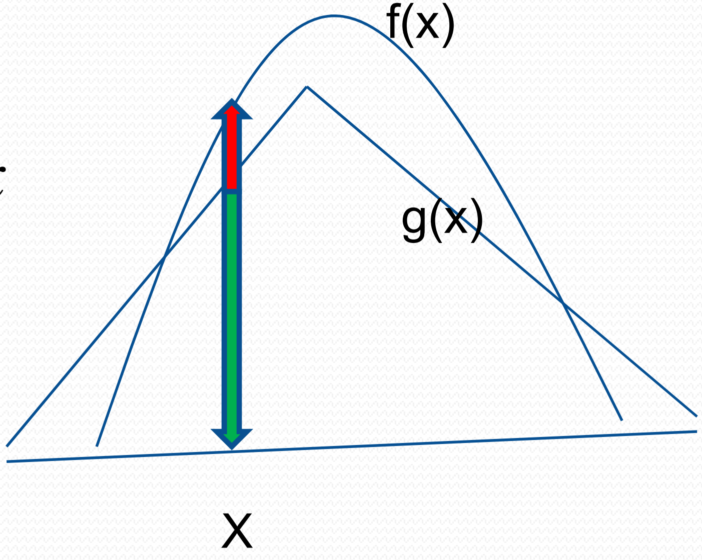
- Estimate an expected value of a function *h(x)* with respect to p.d.f. *f(x)* (hard to simulate).
- Simulate under an alternative probability density function *g(x)*. Then

$$\int h(x) f(x) dx = \int h(x) \frac{f(x)}{g(x)} g(x) dx$$

- Estimate the right hand side using

$$\frac{1}{N} \sum_{i=1}^{N} h(X_i) \frac{f(X_i)}{g(X_i)}, \text{ for } X_i \sim g$$

The IS weights



SIMULATION OF RARE EVENTS

- Basic problem (queueing, networks, risk management): simulate events of very small probability e.g. $P(S_n > t) = p$
- S_n partial sum of (i.i.d.) random variables, p is very small, for example 10^{-6}
- ***Crude:*** Simulate N values of S_n and then estimate p *using:* $\hat{p} = \dfrac{\text{Number of } S_n > t}{N}$
- Relative Error: $\dfrac{\sqrt{\mathrm{var}(\hat{p})}}{E(\hat{p})} = \dfrac{\sqrt{\dfrac{p(1-p)}{N}}}{p}$

Relative error (RE) acceptable, say *1%*, if we do $N = 10^{10}$ simulations

Use Importance Sampling (IS)

- Generate S (drop subscript n) from an exponential family of densities (for some $T(s)$):

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

R. Rubinstein &
D. Kroese
S. Asmussen &
P. Glynn

where $m(\theta) = \int e^{\theta T(s)} f(s) ds$

$f(s)$ is the probability density function of S

- Estimate p with

$$\text{average}\{\frac{f(S)}{f_\theta(S)} I(S > t)\}$$

$$= m(\theta)\ \text{average}\ \{e^{-\theta T(S)} I(S > t)\}$$

Questions.

- What is a “good” choice of IS distribution.
- When does the exponential tilt $T(s)=s$ deliver bounded relative error (as opposed to logarithmic efficiency) for estimating very small probabilities?
- Do we need to use cross-entropy to select optimal parameter values? What choice of entropy measure?
- Is there connection between Extreme Value Theory and efficient simulation of rare events?

Relative Error of Importance Sampling estimator

- RE of IS distribution is

$$N^{-1/2}\sqrt{\exp\{D_2(h; f_\theta)\}-1}$$

where $$h(s) = \frac{1}{p} f(s) I(s > t)$$

is the "target" density function, D_2 is Rényi generalized divergence.

Rényi generalized divergence: extends cross-entropy

$$D_\alpha(h; f_\theta) = \begin{cases} E_h \ln\left[\dfrac{h(x)}{f_\theta(S)}\right] & \alpha = 1 \\ \dfrac{1}{\alpha - 1} \ln E_h\left[\left(\dfrac{h(x)}{f_\theta(S)}\right)^{\alpha-1}\right] & \alpha \neq 1, \alpha > 0 \end{cases}$$

is Rényi generalized divergence of order α for pdfs *h and* f_θ.

Minimizing $D_\alpha(h; f_\theta)$

- Choose θ to minimize the cross entropy or the variance of the IS estimator.
- To minimize $D_\alpha(h; f_\theta)$ where $h(s) = \frac{1}{p} f(s) I(s > t)$

for α=1,2 solve respectively:

$$E_f[(T(S) - E_\theta T) I(S > t)] = 0$$

Rubinstein's minXent when α=1.

$$E_f[(T(S) - E_\theta T) e^{-\theta T(S)} I(S > t)] = 0$$

$E\,[T(S)]$ under f_θ

Minimum Divergence Principle

To estimate an integral $\int h(x)dx$ where $h \geq 0$ (*not necessarily a pdf*) using IS, 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. (c is normalizing constant)

α=1: minimum cross-entropy

α=2: minimum variance

Bounded Relative Error

- I wish to use a **common exponential family** as importance distribution for estimating rare event probabilities $p_t = P[S>t]$, t *large*.

Suppose G is a class of integrable functions g.

Definition: We will say the family f_θ has bounded RE for the class G if the orbit of the exponential family passes close enough to every function in G that its RE is bounded, that is if $\sup_{g \in G} \inf_\theta D_2(c|g|, f_\theta) < \infty$

When is RE bounded?

- Estimate the probability *p=P(S>t)*, *p* is very small, *S* ~ *f(x)* *and* *G* *the class of functions*

$$g(s) = f(s)I(s > t), \quad t \text{ large}$$

- Consider an IS distribution obtained from *standard exponential tilt.*

$$f_\theta(s) = \frac{1}{m(\theta)} e^{\theta s} f(s)$$

Bounded relative error as t $\to \infty$?

f(s)	Bounded RE?
Normal	NO
Exponential	NO
Pareto	NA
Uniform	YES

Why not use the “perfect” IS distribution?

- Try the family $f_\theta(s) = \dfrac{1}{P(s > \theta)} f(s) I(s > \theta)$
- **This family does have** RE error for rare event probabilities: (there is a member of this family which provides IS estimator with variance 0)
- f_θ is neither exponential family nor easy to simulate from.
- **Example:** let $S \sim f(x)$ be $N(0,1)$. Suppose we sample N times from the distribution for $S|S>t$.

N(0,1) example

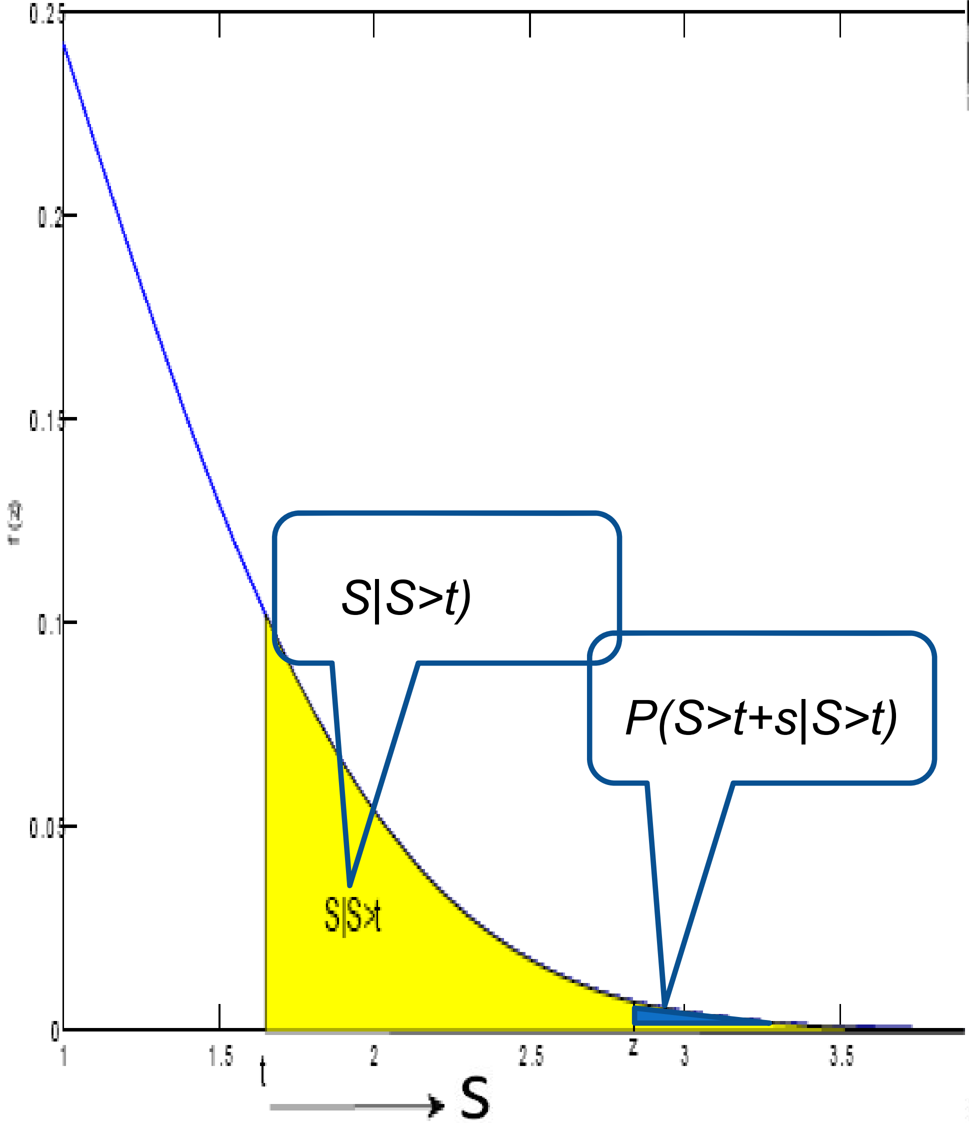
- let $S \sim f(x)$ is $N(0,1)$. Sample N from the distribution of $S|S>t$.
- Relative error for the estimation of rare event scenarios such as $P(S>t+s|S>t)$ for s fixed is, in the limit as $t \to \infty$,

$$N^{-1/2}\sqrt{e^{st+s^2/2}-1}$$

This grows rapidly in t and s. If $t=6$ and $s=3$ we need about 60 trillion simulations for a RE of 1%.

Why so poor??????

GREED!



We do have pills to combat GREED.
They cost $149 each.

A Less Greedy Alternative

- Try being less greedy about estimating a particular value of $p_t = P(S>t)$.
- Always (theoretically) possible to achieve bounded RE error for family of probabilities $p=P(S>t)$ as $t \to \infty$.
- For example transform $X=\Phi(S)$, where Φ is the standard normal c.d.f., converts problem to uniform[0,1] random variables X.
- Find a strategy that works for uniform.

U[0,1] Example.

- When S is $U[0,1]$, the IS distribution obtained from an exponential tilt:

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

$$\theta \approx \frac{1.5936}{p}$$

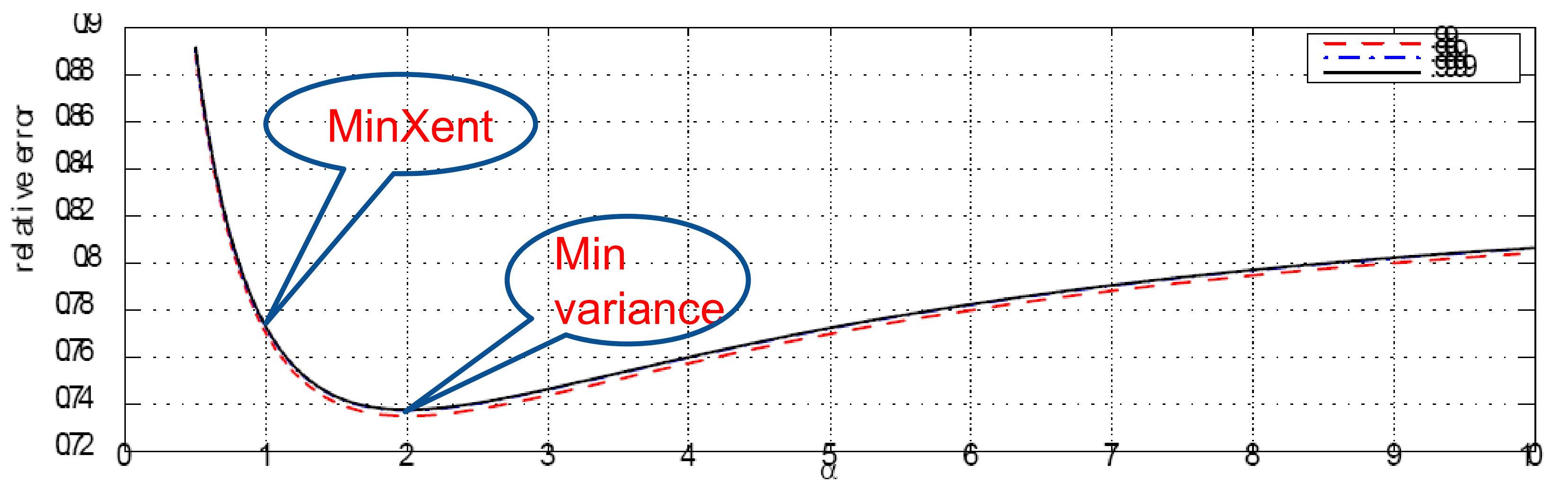
(this value of θ minimizes the limiting RE)

- Provides bounded RE with limit ($p \to 0$) around $0.738N^{-1/2}$.

Estimate p from preliminary simulation and use IS with $\theta = 1.5936/p$.

Uniform Example:

$N^{1/2}\times$ asymptotic relative efficiency (as $p \to 0$) is given for $p=0.01, 0.001$, and 0.0001. For $\alpha=2$ and $\theta=1.5936/p$, RE is about $0.738\ N^{-1/2}$. $N=5500$ provides RE less than about 1% for **all small *p*** (need p to estimate p?)



Value of α used for minimizing divergence

Figure 1: $N^{1/2}\times$Relative error$=\sqrt{\exp\{D_2(cfg, f_{\theta_{\alpha\gamma}})\}-1}$ as a function of α

for $p=0.01, 0.001, 0.0001$

Example: Exponential Distribution

- Suppose *S* has an exponential(1) distribution. Use the standard exponential tilt with $T(S)=S$ to estimate

$p=P[S>-\ln p]$. Tilted density is again an exponential distribution.

- Solve for optimal parameter θ and put in expression for RE:

$$RE \sim \sqrt{\frac{e}{2N}(-\ln p)}$$

Not bounded as $p \to 0$. **Normal**, same story.......

- However if we use a Gumbel IS distribution

$$f_\theta(x) = \theta \exp\{-\theta e^{-x} - x\}, \quad \theta = \frac{1.5936}{p}$$

for exponential, RE **is** bounded with limit around $0.738N^{-1/2}$

Proposition

Ideal Tilt uses T which is tail equivalent to survivor function

- Suppose we wish to estimate $p_t = P(X>t)$ using an importance sampling p.d.f. of the form

$$f_\theta(s) = \frac{1}{m(\theta)} e^{\theta T(s)} f(s) \text{ where } m(\theta) = \int e^{\theta T(s)} f(s) ds$$

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 this family of distributions provides IS estimators with asymptotically bounded relative error as $p_t \to 0$.

$a_x \asymp b_x$ if $\limsup a_x/b_x < \infty$ and $\liminf a_x/b_x > 0$ as $x \to x_F$

Proposition (Fréchet MDA)

- Suppose that f is regularly varying at ∞ with index ρ-1, with $\rho = -1/\xi < 0$. Consider $\overline{T}(x) = (1+x)^{-\varsigma}$ for $0 < \varsigma < 2/\xi$
- Define the IS probability density function

$$f_\theta(s) = ce^{-\theta \overline{T}(s)} |\overline{T}'(s)|, 0 < s < \infty \text{ where } c(\theta) = 1 / \int e^{-\theta \overline{T}(s)} |\overline{T}'(s)| ds$$

Suppose $\theta = \theta_t$ is chosen so that $\theta_t \asymp 1/\overline{T}(t)$ as $t \to \infty$

Then the sequence of distributions f_θ provides importance sample estimators with bounded relative error as $p_t \to 0$.

It's OK to be out in the tail index by up to a factor of 2.

Proposition (Weibull MDA)

- Suppose that f is regularly varying at $0<x_F<\infty$ index ρ-1, with $\rho = -1/\xi < 0$. Consider $\overline{T}(x) = (x_F - x)^{\varsigma}$ for $0 < \varsigma < 2\rho$
- Define the IS probability density function

$$f_\theta(s) = ce^{-\theta\overline{T}(s)}\,|\overline{T}'(s)|, 0 < s < \infty \text{ where } c(\theta) = 1/\int e^{-\theta\overline{T}(s)}\,|\overline{T}'(s)|\,ds$$

Suppose $\theta = \theta_t$ is chosen so that $\theta_t \asymp 1/\overline{T}(t)$ as $t \to x_F$.

Then the sequence of distributions f_θ provides importance sample estimators with bounded relative error as $p_t \to 0$.

again OK to be out in index of RV by up to a factor of 2.

Example: Tukey's g&h distribution

- Used to model extreme events (e.g. Windspeed:Field & Genton, Insurance:Embrechts et al.)
- **Definition** $X = \mu + \sigma \frac{e^{gZ} - 1}{g} e^{hZ^2/2}$

where μ, σ are location and scale, g,h skewness and elongation parameters.

- Suppose (X_1, X_2) are i.i.d. g&h (g=0.1,h=0.2, σ=1, μ=0) random variables. We want to estimate $P(X_1+X_2>t)$ (if t=50 , answer about 4×10^{-6})

min(X_1,X_2)

and the distribution of $X_{(1)}|X_1+X_2>t$,.

Relative error of crude estimator, $N=10^6$, is about ½, IS estimator 10^{-3}

Importance distribution for g&h tail

- For such wide-tailed (sub-exponential) distributions, the probability in the tails is driven by the largest value $X_{(2)}$: $P[X_1+X_2>t] \sim P[X_{(2)}>t]$ as $t \to \infty$.
- Tilted on the distribution of the maximum by altering the beta distribution applied to uniform inputs U_2 where

$$X_{(2)} \text{ generated as } Z_{(2)} = \Phi^{-1}(U_2)$$

$$\mu + \sigma \frac{\exp\{gZ_{(2)}\} - 1}{g} \exp\{hZ_{(2)}{}^2\}$$

then generate input for $X_{(1)}$, U_1 as U$[0, U_2]$

Joint conditional distribution of g&h order statistics

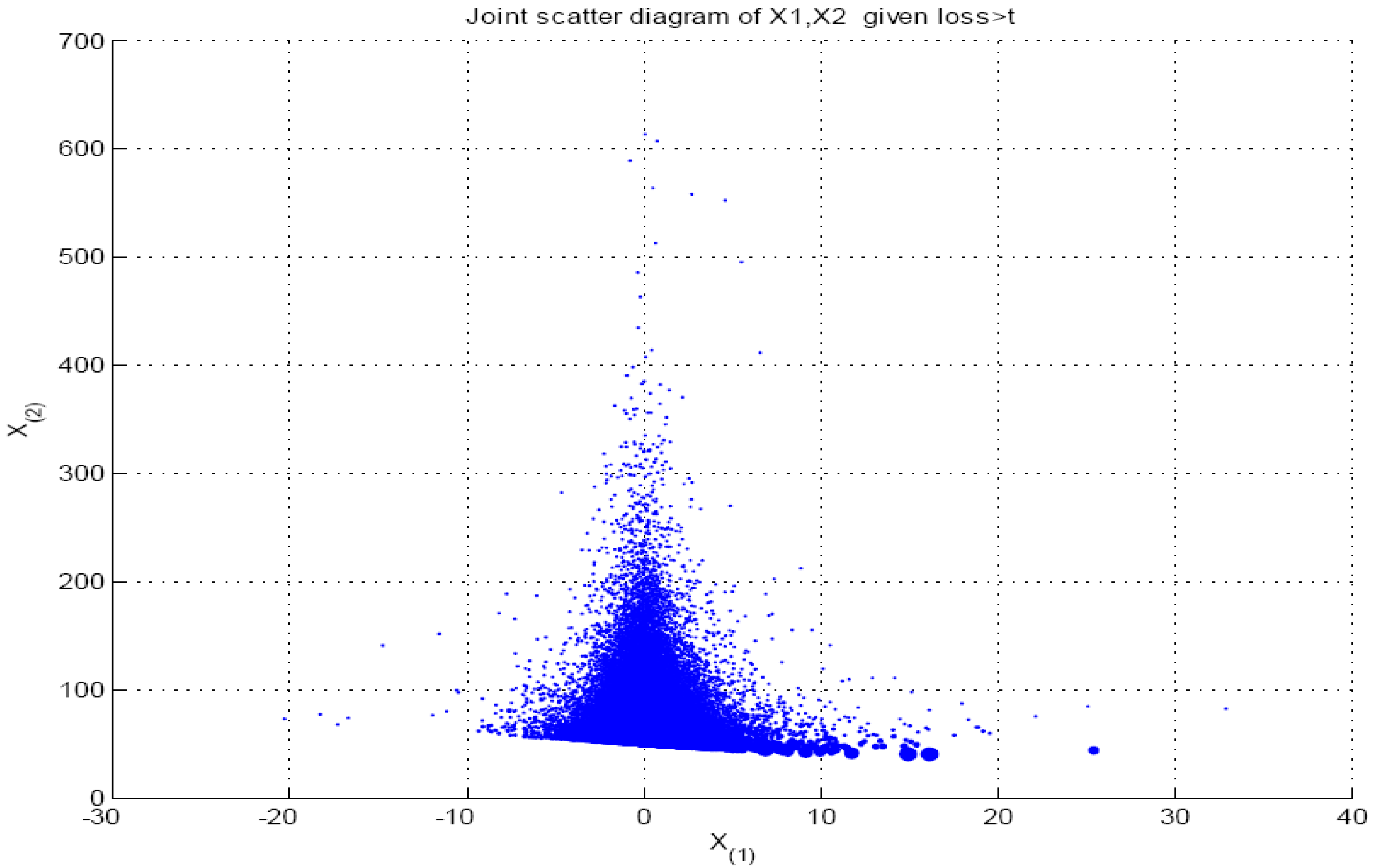


Figure 7: Simulated distribution of $(X_{(1)}, X_{(2)})$ given $X_1 + X_2 > 50$ for the g&h distribution.

Conditional survivor function

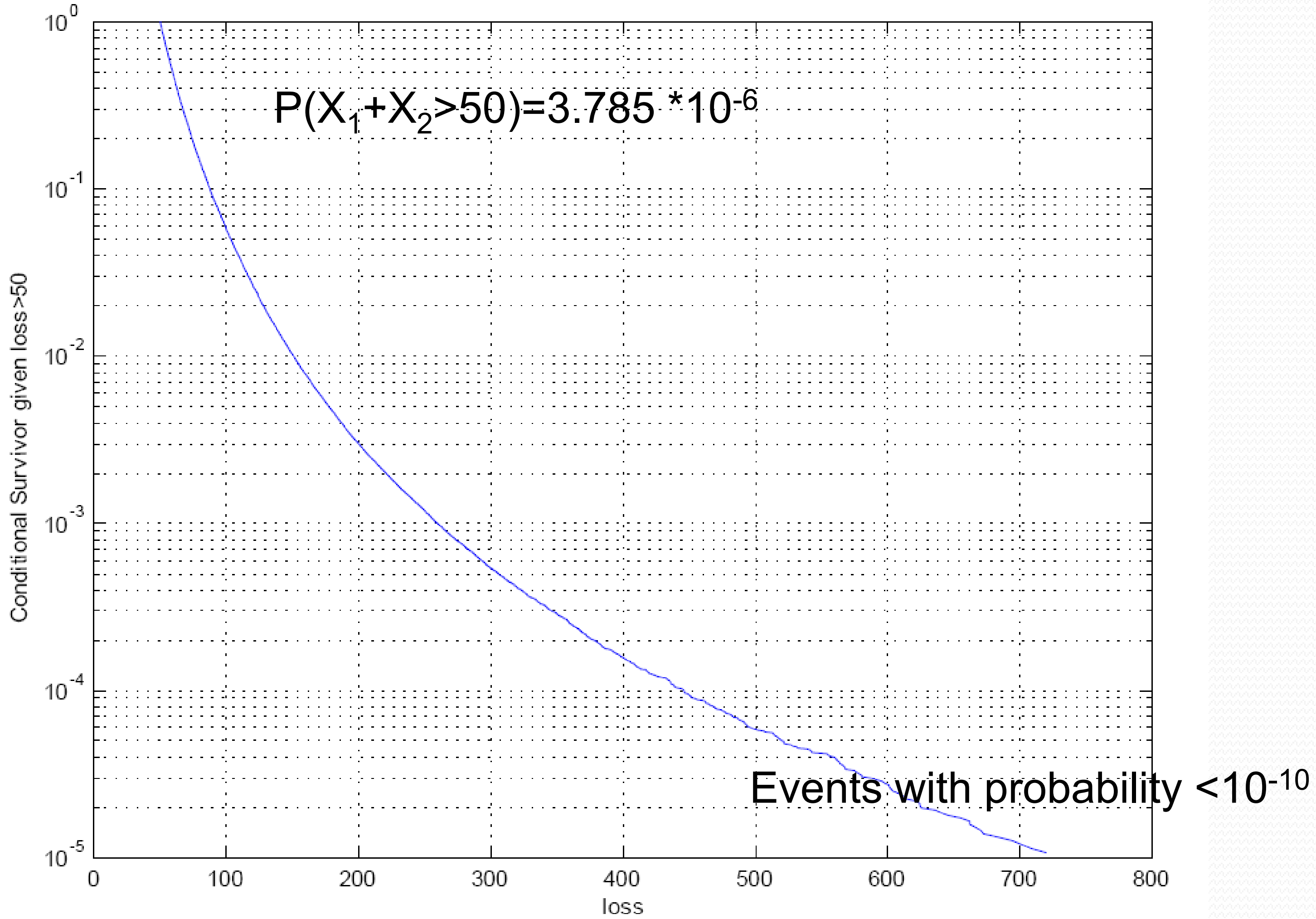


Figure 8: $P[X_1 + X_2 > x | X_1 + X_2 > 50]$ for g&h distributed random variables

Applications:

- **Further Example: Skewed Normal**
- **IS Permits bounded relative error estimation of tail probabilities when tails are:**
  - Sums or linear combinations of independent random variables with regularly varying tails
  - Sums of random variables where one tail dominates
  - Scale mixtures of regularly varying random variables
- **We can determine the optimal parameter value without minimizing divergence if we know the asymptotic tail behaviour.**

Conclusions

1. Simulating rare event probabilities with bounded relative error **is** possible using importance sampling.
2. Thanks.

Happy 70th Birthday Reuven and many ideas & productive years ahead!