Optimal Importance Sampling Change of Measure for Large Sums of Random Variables

Sassan Mokhtar

June 8, 2020

Outline

Table of Contents

Rare Events

- The objective is to estimate,

$$\alpha(\gamma, N) = P\Big(\sum_{i=1}^{N} X_i \leq \gamma\Big)$$

  where $X_1, X_2, ..., X_N$ are i.i.d. non-negative random variables with probability density function $f_X(.)$ and cumulative distribution function $F_X(.)$.
- If N is large and γ is small $\rightarrow$ the regime is called rare event.
- *Rare events*: Infrequent events that might have widespread effect on the system and even destabilize it.
- *Real world applications*:
  - In Control Systems: Probability of collision of two aircrafts.
  - In Insurance: Probability of ruin a company.
  - ...

Crude Monte Carlo

We have,

$$\alpha(\gamma, N) = P\left(\sum_{i=1}^{N} X_i \leq \gamma\right) = \mathbb{E}\left[\mathbb{1}_{(\sum_{i=1}^{N} X_i \leq \gamma)}\right]$$

Thus, the crude Monte Carlo estimator is given by,

$$\hat{\alpha} = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}^{(j)}_{(\sum_{i=1}^{N} X_i \leq \gamma)}$$

where X are i.i.d. and $X \sim f_X$.

$\rightarrow$ To compare different estimators, we use the squared coefficient of variation (SCV), which is defined to be the ratio between the variance of the estimator and its squared mean.

Crude Monte Carlo

Since the variance of crude Monte Carlo estimator is,

$$Var(\mathbb{1}_{(\sum_{i=1}^{N} X_i \leq \gamma)}) = \mathbb{E}\left[\left(\mathbb{1}_{(\sum_{i=1}^{N} X_i \leq \gamma)}\right)^2\right] - \mathbb{E}\left[\mathbb{1}_{(\sum_{i=1}^{N} X_i \leq \gamma)}\right]^2$$
$$= \mathbb{E}\left[\mathbb{1}_{(\sum_{i=1}^{N} X_i \leq \gamma)}\right] - \mathbb{E}\left[\mathbb{1}_{(\sum_{i=1}^{N} X_i \leq \gamma)}\right]^2 = \alpha(\gamma, N) - \alpha^2(\gamma, N)$$

The squared coefficient of variation is given by,

$$SCV_{crude} = \frac{\alpha(\gamma, N) - \alpha^2(\gamma, N)}{\alpha^2(\gamma, N)} \approx \frac{1}{\alpha(\gamma, N)} \tag{1}$$

The estimator is called to have a bounded relative error, if it holds $\limsup_{\alpha(\gamma,N)\to 0} SCV < \infty$.
Since, $SCV_{crude} \to \infty$ as $\alpha \to 0$, the crude Monte Carlo estimator does not have a bounded relative error.

Crude Monte Carlo

For example, if α is of order 10^{-6}, then according to CLT, the relative error is estimated as $\frac{c\sqrt{SCV_{crude}}}{\sqrt{M}}$.

Then, to meet 5% relative error we need $M \approx \frac{1.96^2.SCV_{crude}}{(0.05)^2} \approx 10^9$ samples.

$\rightarrow$ We need to apply some variance reduction techniques.

Table of Contents

Approach Based on Sample Rejection

This approach, which is the most simple Importance Sampling estimator, is based on sample rejection.

Let $A = \bigcap_{i=1}^{N} \{X_i \leq \gamma\}$, then according to the law of total probability,

$$P\big(\sum_{i=1}^{N} X_i \leq \gamma\big) = P\Big(\sum_{i=1}^{N} X_i \leq \gamma | A\Big).P(A) + P\Big(\sum_{i=1}^{N} X_i \leq \gamma | A^c\Big).P(A^c)$$

Since the second term in the right hand side is zero and the fact that,

$$P(A) = P(X_1 \leq \gamma, X_2 \leq \gamma, ..., X_N \leq \gamma) = \big(F_X(\gamma)\big)^N$$

We obtain,

$$\alpha(\gamma, N) = P\Big(\sum_{i=1}^{N} X_i \leq \gamma\Big) = P\Big(\sum_{i=1}^{N} \omega_i \leq 1\Big)\big(F_X(\gamma)\big)^N \tag{2}$$

where $\omega_i = \{\frac{X_i}{\gamma} | \frac{X_i}{\gamma} \leq 1\}$

Approach Based on Sample Rejection

Then, the estimator is obtained by estimating the term $P\Big(\sum_{i=1}^{N}\omega_i \leq 1\Big)$ by crude Monte Carlo method, where it is not rare event anymore, as long as N is not large.

$\rightarrow$ Note that this estimator can be interpreted as applying importance sampling with biased PDF being the truncation of the underlying PDF over the hypercube $[0,\gamma]^N$.

The next step is to compute the SCV for the approach based on sample rejection.

First, notice that,

$$\alpha(\gamma, N) = \mathbb{E}\Big[\big(F_X(\gamma)\big)^N \mathbb{1}_{(\sum_{i=1}^{N}\omega_i \leq 1)}\Big]$$

Let $Z_\gamma = \big(F_X(\gamma)\big)^N \mathbb{1}_{(\sum_{i=1}^{N}\omega_i \leq 1)}$.

Approach Based on Sample Rejection

Then,

$$Var(Z_\gamma) = \mathbb{E}[(Z_\gamma)^2] - \mathbb{E}[Z_\gamma]^2 = \left(F_X(\gamma)\right)^{2N} \frac{\alpha(\gamma, N)}{\left(F_X(\gamma)\right)^N} - \alpha(\gamma, N)^2$$

Thus,

$$SCV_{SR} = \frac{\left(F_X(\gamma)\right)^N}{\alpha(\gamma, N)} - 1$$

→ Notice that for fixed N, this estimator achieves the bounded relative error with respect to the parameter γ for distributions that satisfy $F_X(x) \sim cx^d$ (where $d > 0$) as $x \to 0$. This is due to the fact that

$$\alpha(\gamma, N) = P\Big(\sum_{i=1}^{N} X_i \leq \gamma\Big) \geq \prod_{i=1}^{N} P(X_i \leq \frac{\gamma}{N}) = \Big(F_X(\frac{\gamma}{N})\Big)^N$$

Approach Based on Sample Rejection

Then, for the given CDF, it satisfies,

$$SCV_{SR} \leq \frac{\Big(F_X(\gamma)\Big)^N}{\Big(F_X(\gamma/N)\Big)^N} - 1 = \mathcal{O}(1)$$

$\rightarrow$ **Question:** What happens when N is large?
Assume that $X_i, \ i = 1, ..., N$ are uniform between $[0, 1]$, then it can easily proven by induction that,

$$P\Big(\sum_{i=1}^{N} \omega_i \leq 1\Big) = \frac{1}{N!}$$

Therefore, the estimator has a SCV approximately equal to $N!$ which is worse than any exponential increase.

Approach Based on Sample Rejection

In general, by applying the Chernoff inequality and using the fact that $X_1, ..., X_N$ are i.i.d., we obtain the following upper bound,

$$P\left(\sum_{i=1}^{N} \omega_i \leq 1\right) \leq \min_{\eta \geq 0} \exp\left(\eta + N \log(\mathbb{E}[e^{-\eta\omega}])\right)$$

This shows that the estimate based on sample rejection does not perform well when N is large.

Approach Based on Maximum Entropy

The idea of this approach is to propose an importance sampling estimator to estimate $P\Big(\sum_{i=1}^{N}\omega_i \leq 1\Big)$ for large values of N.
By doing this, the rarity parameter N is incorporated in variance reduction procedure, as well.
Therefore, the goal is to find the density f_ω^* that solves,

$$\begin{gathered}\max_f \; -\int_0^1 \log\big(f(x)\big) f(x)dx \\ \text{s.t. } \int_0^1 x f(x)dx = \frac{1}{N} \\ f(x) \geq 0,\; 0 \leq x \leq 1.\end{gathered}$$

when N is large the solution to this problem is exponential with rate $\lambda(N) \approx N$.

Approach Based on Maximum Entropy

Therefore, the marginal density of each ω_i is,

$$f^*_\omega(x) = \frac{Ne^{-Nx}}{1-e^{-N}}, \quad 0 \leq x \leq 1.$$

Then, by applying this change of measure to the previous approach we can write,

$$P\Big(\sum_{i=1}^{N} \omega_i \leq 1\Big) = \mathbb{E}_{f^*_\omega}\left[\mathbb{1}_{(\sum_{i=1}^{N} \omega_i \leq 1)} \prod_{i=1}^{N} \frac{(1-e^{-N}) f_\omega(\omega_i)}{Ne^{-N\omega_i}}\right]$$

The next step is to obtain the squared coefficient of variation for this maximum entropy estimator.

Approach Based on Maximum Entropy

Notice that,

$$\alpha(\gamma, N) = \mathbb{E}_{f_\omega^*}\left[\mathbb{1}_{\left(\sum_{i=1}^N \omega_i \le 1\right)} \prod_{i=1}^N \frac{(1-e^{-N}) f_\omega(\omega_i)}{N e^{-N\omega_i}}\right] \left(F_X(\gamma)\right)^N$$

Let $Z_\gamma = \left(F_X(\gamma)\right)^N \mathbb{1}_{\left(\sum_{i=1}^N \omega_i \le 1\right)} \prod_{i=1}^N \frac{(1-e^{-N}) f_\omega(\omega_i)}{N e^{-N\omega_i}}$. Then,

$$Var(Z_\gamma) = \mathbb{E}_{f_\omega^*}\left[\mathbb{1}_{\left(\sum_{i=1}^N \omega_i \le 1\right)} \prod_{i=1}^N \frac{(1-e^{-N})^2 f_\omega(\omega_i)^2}{N^2 e^{-2N\omega_i}}\right] \left(F_X(\gamma)\right)^{2N} - \alpha^2(\gamma, N)$$

And,

$$SCV_{ME} = \frac{\mathbb{E}_{f_\omega^*}\left[\mathbb{1}_{\left(\sum_{i=1}^N \omega_i \le 1\right)} \prod_{i=1}^N \frac{(1-e^{-N})^2 f_\omega(\omega_i)^2}{N^2 e^{-2N\omega_i}}\right] \left(F_X(\gamma)\right)^{2N}}{\alpha^2(\gamma, N)} - 1$$

Approach Based on Maximum Entropy

The main limitation of the maximum entropy approach is that the underlying density $f_\omega(.)$ is not involved in the optimization problem.

In other words, applying the truncated exponential $f_\omega^*(.)$ as a biased density for all possible choices of $f_\omega(.)$ could lead to even higher variance than the crude Monte Carlo method.

As a result, another approach has to be developed.

Approach Based on Exponential twisting

- In this approach the biased densities are obtained by minimizing the Kullback-Leibler divergence of underlying and the biased densities.
- Let $h(\mathbf{x}) = \prod_{i=1}^{N} f_X(x_i)$ be the joint PDF of $(X_1, ..., X_N)^T$, where $\mathbf{x} = (x_1, ..., x_N)^T$. Then, the objective is to obtain the new joint density f^* by solving the following optimization problem,

$$\begin{aligned} &\inf_{f^* \geq 0} \int f^*(\mathbf{x}) \log\left(\frac{f^*(\mathbf{x})}{h(\mathbf{x})}\right) d\mathbf{x} \\ &s.t. \quad \int f^*(\mathbf{x}) d\mathbf{x} = 1 \\ &\mathbb{E}_{f^*}\left[\sum_{i=1}^{N} X_i\right] = \gamma \\ &f^*(x) \geq 0,\ x \geq 0 \end{aligned} \tag{3}$$

Approach Based on Exponential twisting

By applying the method of Lagrange multiplier, we obtain the following solution (Ridder and Rubinstein 2007),

$$f^*(\mathbf{x}) = \frac{h(\mathbf{x}) \exp(\theta \sum_{i=1}^{N} x_i)}{c(\theta)} \tag{4}$$

where $c(\theta) = \mathbb{E}_h[\exp(\theta \sum_{i=1}^{N} x_i)]$ is the normalizing factor and θ is the Lagrange multiplier that solves,

$$\frac{\mathbb{E}_h[\sum_{i=1}^{N} X_i \exp(\theta \sum_{i=1}^{N} x_i)]}{\mathbb{E}_h[\exp(\theta \sum_{i=1}^{N} x_i)]} = \gamma$$

Thus, the optimal density is given by exponentially twisting of each univariate density $f_X(.)$,

$$f_X^*(x) = \frac{f_X(x) e^{\theta x}}{M(\theta)}, \ x \geq 0$$

where $M(\theta) = \mathbb{E}_{f_X}[e^{\theta x}]$ is the moment generating function. Then, θ satisfies $\frac{M'(\theta)}{M(\theta)} = \frac{\gamma}{N}$.

Approach Based on Exponential twisting

While the change of measure based on exponential twisting is known to be optimal and expected to outperform the previous approaches, it has two main limitations:

1. $M(\theta)$ should be known in closed form expression. And this is not the case always. For example, the moment generating function of log-normal distribution is not known in general.
2. Sampling under the new measure f_X^* is not straight forward and might be expensive.

$\rightarrow$ Next, we are going to propose the alternative change of measure that lead to the same performance as the exponential twisting, but without its limitations.

This can be done by distinguishing 3 different cases.

Approach Based on Exponential twisting

(i) $f_X(x) = \mathcal{O}(1)$ as $x \to 0$:

Since the exponential twisting change of measure is given by,

$$f_X^*(x) \propto f_X(x).e^{\theta x}, \; x \geq 0$$

with $\theta \to -\infty$ as $\alpha(\gamma, N) \to 0$, we propose the following change of measure,

$$\tilde{f}_X(x) = \frac{e^{\theta x}}{\tilde{M}(\theta)}, \; x \geq 0$$

with $\tilde{M}(\theta) = -\frac{1}{\theta}$. Recall that θ should satisfies $\frac{\tilde{M}'(\theta)}{\tilde{M}(\theta)} = \frac{\gamma}{N}$. Thus, $\theta = -\frac{N}{\gamma}$.

Consequently, the proposed change of measure is,

$$\tilde{f}_X(x) = \frac{N}{\gamma} e^{-\frac{N}{\gamma}x}, \; x \geq 0$$

Approach Based on Exponential twisting

(ii) $f_X(x) = x^d g(x)$ with $g(x) = \mathcal{O}(1)$, as $x \to 0$ and $d > -1$:

The proposed change of measure for this case is,

$$\tilde{f}_X(x) = \frac{x^d e^{\theta x}}{\tilde{M}(\theta)}, \ x \geq 0$$

By comparing with the Gamma distribution,

$$Gamma(k, \beta) = \frac{1}{\beta^k \Gamma(k)} x^{k-1} e^{-\frac{x}{\beta}}$$

we obtain that the new measure corresponds to the Gamma distribution with shape parameter $d + 1$ and scale parameter $-\frac{1}{\theta}$. Also, it is clear that the normalizing factor is $\tilde{M}(\theta) = \frac{\Gamma(d+1)}{(-\theta)^{d+1}}$. Then, since $\frac{\tilde{M}'(\theta)}{\tilde{M}(\theta)} = \frac{\gamma}{N}$, we obtain,

$$\theta = -\frac{N}{\gamma}(d+1)$$

Approach Based on Exponential twisting

(iii) Distributions that does not approaches to 0 polynomially:

This case is the most difficult one and it is required to consider in a case by case basis.
$\rightarrow$ Take for example sum of i.i.d. standard log-normal random variables. (Asmussen et al. 2016) proposed the following procedure to apply the exponential twisting technique to this case,

1. They propose an unbiased estimator of the moment generating function $M(\theta)$
2. Then approximate the value of θ that satisfies $\frac{\tilde{M}'(\theta)}{\tilde{M}(\theta)} = \frac{\gamma}{N}$
3. Finally, they use acceptance-rejection to sample from the biased density (which might be expensive when N is large)

Approach Based on Exponential twisting

First, let $A = \bigcap_{i=1}^{N} \{X_i > \frac{\delta\gamma}{N}\}$ and rewrite the $\alpha(\gamma, N)$,

$$P\Big(\sum_{i=1}^{N} X_i \leq \gamma\Big) \approx \Big(1 - F_X(\frac{\delta\gamma}{N})\Big)^N P\Big(\sum_{i=1}^{N} X_i \leq \gamma | A\Big)$$

where $\delta \in (0, 1)$ is a fixed value that control the bias.

While there is a closed form expression for the first term in the RHS, it is needed to estimate the second term.

Recall that X_i are standard log-normal random variables. Thus, the density of $X_i | X_i > \frac{\delta\gamma}{N}$ is given by,

$$\bar{f}_X(x) = \frac{1}{x\sqrt{2\pi}} \frac{\exp(-\frac{(\log x)^2}{2})}{P(X_i > \frac{\delta\gamma}{N})}, \; x \geq \frac{\delta\gamma}{N}$$

where $P(X_i > \frac{\delta\gamma}{N})$ is the normalizing factor.

Approach Based on Exponential twisting

Therefore,

$$P\Big(\sum_{i=1}^{N} X_i \leq \gamma | A\Big) = P_{\bar{f}_X}\Big(\sum_{i=1}^{N} X_i \leq \gamma\Big)$$

Recall that the exponential twisting change of measure is,

$$\bar{f}_X^*(x) \propto \bar{f}_X(x) e^{\theta x}, \; x \geq \frac{\delta\gamma}{N}$$

Next, we apply the Taylor expansion to approximate $\bar{f}_X(.)$ in the interval $[\delta\gamma/N, \gamma]$,

$$\bar{f}_X(x) = \bar{f}_X\Big(\frac{\delta\gamma}{N}\Big) + \Big(x - \frac{\delta\gamma}{N}\Big)\bar{f}'_X\Big(\frac{\delta\gamma}{N}\Big) + o(x - \frac{\delta\gamma}{N})$$

Combining both equations lead to the following change of measure (with notation $\bar{f}_X = \bar{f}_X(\delta\gamma/N)$ and $\bar{f}'_X = \bar{f}'_X(\delta\gamma/N)$),

$$\tilde{f}_X(x) = \frac{\bar{f}_X e^{\theta x} + (x - \frac{\delta\gamma}{N})\bar{f}'_X e^{\theta x}}{\tilde{M}_X(\theta)}, \; x \geq \frac{\delta\gamma}{N}$$

Approach Based on Exponential twisting

Then,

$$\tilde{M}(\theta) = -\frac{\exp(\theta\delta\gamma/N)}{\theta}\bar{f}_X + \frac{\exp(\theta\delta\gamma/N)}{\theta^2}\bar{f}'_X$$

Now, it is time to go through the step 2 and estimate θ. Again, recall that θ satisfies $\frac{\tilde{M}'(\theta)}{\tilde{M}(\theta)} = \frac{\gamma}{N}$. Thus,

$$\theta = -\frac{\bar{f}_X - c\bar{f}'_X + \sqrt{(\bar{f}_X - c\bar{f}'_X)^2 + 8\bar{f}_X\bar{f}'_X c}}{2c\bar{f}_X}$$

where $c = \frac{\gamma}{N}(1-\delta)$

Finally, sampling from $\tilde{f}_X(.)$ can be done by writing,

$$\tilde{f}_X(x) = -\frac{\bar{f}_X\exp(\theta\delta\gamma/N)}{\tilde{M}_X(\theta)\theta}\tilde{f}_1(x) + \frac{\bar{f}'_X\exp(\theta\delta\gamma/N)}{\tilde{M}_X(\theta)\theta^2}\tilde{f}_2(x)$$

where $\tilde{f}_1(x) = -\frac{\theta\exp(\theta x)}{\exp(\theta\delta\gamma/N)}$ and $\tilde{f}_2(x) = -\frac{\theta^2(x-\delta\gamma/N)\exp(\theta x)}{\exp(\theta\delta\gamma/N)}$ are valid densities for $x > \delta\gamma/N$.

Table of Contents

Numerical Results

Consider the problem

$$\alpha(\gamma, N) = P\Big(\sum_{i=1}^{N} X_i \leq \gamma\Big)$$

where X are i.i.d. and drawn from the Weibull distribution.
In order to apply the optimal exponential twisting, one requires to compute θ from $\frac{M'(\theta)}{M(\theta)} = \frac{\gamma}{N}$, where $M(\theta) = \sum_{n=0}^{\infty} \frac{\theta^n \lambda^n}{n!} \Gamma(1 + \frac{n}{k})$ which makes the task of computing the θ extremely difficult (if not impossible).
Since the density of Weibull distribution is given by $f_X(x) = \frac{k}{\beta}(\frac{x}{\beta})^{k-1} \exp(-(\frac{x}{\beta})^k)$ i.e. $f_X(x) = x^{k-1} g(x)$, we use the change of measure,

$$\tilde{f}_X(x) = \frac{x^{k-1}\exp(\theta x)}{\tilde{M}(\theta)}, \ x \geq 0$$

with $\tilde{M}(\theta) = \frac{\Gamma(k)}{(-\theta)^k}$ and $\theta = -\frac{N}{\gamma} k$.

Numerical Results

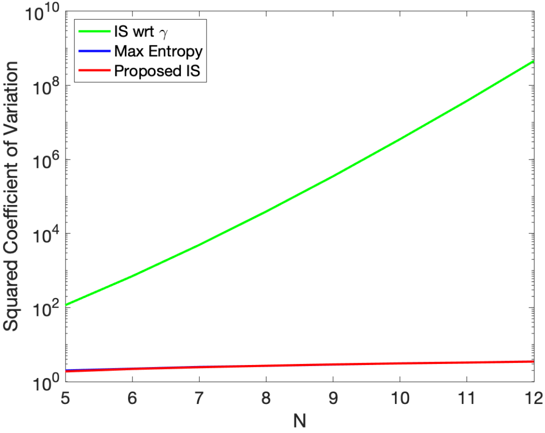

Figure: SCV as a function of N where X_i are i.i.d. Exponential RVs with rate $\lambda = 1$ and $\gamma = 0.01$

Numerical Results

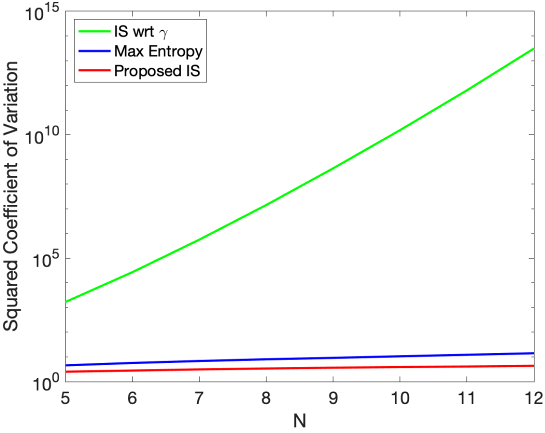

Figure: SCV as a function of N where X_i are i.i.d. Weibull RVs with rate $\lambda = 1, k = 1.5$ and $\gamma = 0.5$

Numerical Results

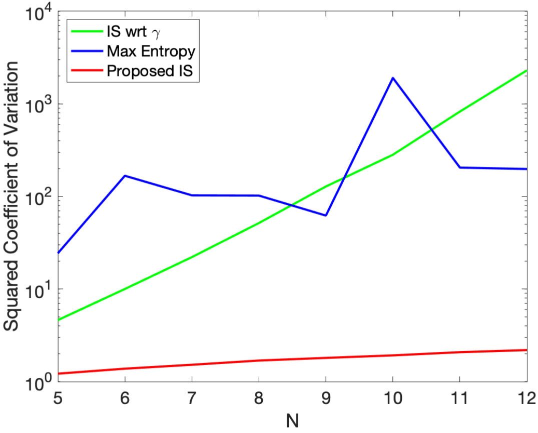

Figure: SCV as a function of N where X_i are i.i.d. Weibull RVs with rate $\lambda = 1, k = 0.5$ and $\gamma = 0.01$

Table of Contents

Conclusion

- We discuss 3 different approaches based on importance sampling to estimate rare-event probabilities.
- The approach based on minimizing the Kullback-Leibler divergence outperform the others.
- We develop the alternative change of measure that yields the same performance as the optimal change of measure, but without its computational limitations.

Future directions:

- How to choose δ?
- How to control the bias?
- What is the effect of δ on variance?

References

1. Ridder, Ad and Rubinstein, Reuven (2007)
Minimum cross-entropy methods for rare-event simulation

2. Asmussen, Søren and Jensen, Jens Ledet and Rojas-Nandayapa, Leonardo (2016)
Exponential family techniques for the lognormal left tail