Efficient Importance Sampling for Large Sums of IID Random Variables

Nadhir Ben Rached, Abdul-Lateef Haji-Ali, Gerardo Rubino, Raúl Tempone

To cite this version:

Nadhir Ben Rached, Abdul-Lateef Haji-Ali, Gerardo Rubino, Raúl Tempone. Efficient Importance Sampling for Large Sums of IID Random Variables. UQ 2022 - SIAM Conference on Uncertainty Quantification, Apr 2022, Atlanta, United States. pp.1-31. hal-03899073

HAL Id: hal-03899073
https://inria.hal.science/hal-03899073
Submitted on 15 Dec 2022
Efficient Importance Sampling for Large Sums of IID Random Variables

Nadhir Ben Rached

joint work with Abdul-Lateef Haji-Ali (Heriot-Watt University), Gerardo Rubino (INRIA, Rennes - Bretagne Atlantique), and Raúl Tempone (KAUST, RWTH Aachen)

April 12, 2022
Estimate using importance sampling (IS)

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i \leq \gamma\right),$$  \hspace{1cm} (1)

where $X_1, X_2, \cdots, X_N$ is a sequence of nonnegative independent and identically distributed (i.i.d) random variables (RVs).

Particularly interested in the rare event regime when $N$ is large and/or $\gamma$ is small.

Application: performance evaluation/prediction of wireless communication systems operating over fading channels (Simon and Alouini 2005).
Need the use of variance reduction techniques as the naive Monte Carlo (MC) estimator is computationally expensive in the regime of rare events.


- Logarithmically efficient estimator in the case of i.i.d. sums Log-normal RVs using exponential twisting (Asmussen et al. 2016).
- Under mild assumptions, logarithmically efficient in case of sums of correlated Log-normal RVs based on shifting the mean of the multivariate Gaussian distribution (Gulisashvili and Tankov 2016).
- Under an assumption, asymptotically vanishing relative error for sums of correlated Log-normal RVs by combining IS with the control variate techniques (Alouini et al. 2018).
- Two unified IS using hazard rate twisting (HRT) for sums of independent RVs. The first estimator is shown to be logarithmically efficient, whereas the second achieves bounded relative error under mild assumptions (Ben Rached et al. 2016).
Efficiency of previous estimators was studied when $N$ is fixed and $\gamma \to 0$.
Efficiency is considerably affected when $N$ increases.

**Aim:** efficiently estimate (1) in the rare event regime when $N$ is large and/or $\gamma$ is small.
Problem setting

Let \( \mathbf{x} = (x_1, \cdots, x_n)^t \) and \( h_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{N} f_X(x_i) \) be the joint PDF of the random vector \((X_1, \cdots, X_N)^t\). We consider the estimation of

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

where \( \mathbb{P}_{h_{\mathbf{x}}}(\cdot) \) is the probability under which the random vector \( \mathbf{X} = (X_1, \cdots, X_N)^t \) is distributed according to \( h_{\mathbf{x}}(\cdot) \).

**Motivating example:** write the quantity of interest as

\[
\mathbb{P}_{h_{\mathbf{x}}}
\left( \sum_{i=1}^{N} X_i \leq \gamma \right) = \mathbb{E}_{h_{\mathbf{w}}}
\left[ (F_X(\gamma))^N \mathbf{1}_{\left( \sum_{i=1}^{N} w_i \leq 1 \right)} \right]
\]

where \( w_i \) is equal in distribution to \( \frac{X_i}{\gamma} \) conditional on the event \( \{X_i \leq \gamma\} \), \( i = 1, 2, \cdots, N \), and \( h_{\mathbf{w}}(\mathbf{w}) = \prod_{i=1}^{N} f_w(w_i) \) with \( f_w(\cdot) \) is the PDF of \( w_i \).

Estimate the right-hand side term by the naive MC method.
For fixed $N$, it achieves the **bounded relative error** property with respect to the rarity parameter $\gamma$ for distributions that satisfy $f_w(x) \sim bx^p$ as $x$ approaches zero and for $p > -1$ and $b > 0$ (Ben Rached et al. 2018).

**But** using the Chernoff bound, we obtain for all $\eta > 0$

$$
\mathbb{P}_{h_w}(\sum_{i=1}^{N} w_i \leq 1) \leq \exp (\eta + N \log (\mathbb{E}_{f_w} [\exp(-\eta w)]))
$$

The squared coefficient of variation of $\hat{\alpha}(\gamma, N)$ is given by

$$
\frac{\text{var}_{h_w}[\hat{\alpha}(\gamma, N)]}{\alpha^2(\gamma, N)} = \frac{1 - \mathbb{P}_{h_w}(\sum_{i=1}^{N} w_i \leq 1)}{\mathbb{P}_{h_w}(\sum_{i=1}^{N} w_i \leq 1)}.
$$

When $\eta = 1$, the squared coefficient of variation is asymptotically equal to $1/\mathbb{P}_{h_w}(\sum_{i=1}^{N} w_i \leq 1)$ and is lower bounded by

$$
\exp (-1 - N \log (\mathbb{E}_{f_w} [\exp(-w)]))
$$

The efficiency of the estimator deteriorates when $N$ is large.
• When applicable, **exponential twisting** is expected to compare favorably, in most of the cases, to existing estimators.

• For light tails i.i.d. case and under regularity assumptions, **exponential twisting** is logarithmically efficient when the probability of interest is either $P(\sum_{i=1}^{N} X_i > \gamma)$ and $\gamma \to +\infty$ or $P(\sum_{i=1}^{N} X_i > \gamma N)$ and $N \to +\infty$ (Asmussen and Glynn 2007).

• In the left tail setting, **exponential twisting** achieves the logarithmic efficiency property in the case of i.i.d. Log-normal RVs when the probability of interest is $P(\sum_{i=1}^{N} X_i < N\gamma)$ and either $N \to +\infty$ or $\gamma \to 0$ (Asmussen et al. 2016).
Exponential twisting PDF minimizes the Kullback-Leibler (KL) divergence with respect to the underlying PDF under the constraint that the rare set \( \{ \mathbf{x} \in \mathbb{R}^N_+, \text{ such that } \sum_{i=1}^{N} x_i \leq \gamma \} \) is no longer rare (Ridder and Rubinstein 2007)

\[
\inf_{h^*_X \geq 0} \int h^*_X(\mathbf{x}) \log \left( \frac{h^*_X(\mathbf{x})}{h_X(\mathbf{x})} \right) d\mathbf{x} \\
\text{s.t} \quad \int h^*_X(\mathbf{x}) d\mathbf{x} = 1 \tag{3}
\]

\[
\mathbb{E}_{h^*_X} \left[ \sum_{i=1}^{N} X_i \right] = \gamma
\]

\[
h^*_X(\mathbf{x}) \geq 0, \quad x_i \geq 0 \text{ for all } i \in 1, 2, \cdots, N.
\]

The solution of this problem is given as

\[
h^*_X(\mathbf{x}) = \frac{h_X(\mathbf{x}) \exp \left( \theta^* \sum_{i=1}^{N} x_i \right)}{\mathbb{E}_{h_X} \left[ \exp \left( \theta^* \sum_{i=1}^{N} X_i \right) \right]}, \quad \mathbf{x} \in \mathbb{R}^N_+ \tag{4}
\]
The optimal twisting parameter $\theta^*$ solves

$$\frac{\mathbb{E}_{h_X} \left[ \sum_{i=1}^{N} X_i \exp \left( \theta^* \sum_{i=1}^{N} X_i \right) \right]}{\mathbb{E}_{h_X} \left[ \exp \left( \theta^* \sum_{i=1}^{N} X_i \right) \right]} = \gamma.$$ 

By writing $h_X^*(\mathbf{x}) = \prod_{i=1}^{N} f_X^*(x_i)$, the optimal density is given by exponentially twisting each univariate PDF $f_X(\cdot)$

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

with $M(\theta) = \mathbb{E}_{f_X}[\exp(\theta X)]$ and the optimal twisting parameter $\theta^*$ satisfies

$$\frac{M'(\theta^*)}{M(\theta^*)} = \frac{\gamma}{N}.$$

Since the left-tail is considered, $\theta^* \to -\infty$ as $\gamma \to 0$ and/or $N \to +\infty$ (Asmussen et al. 2016).
Observe the following drawbacks:

- Sampling according to $f^*_X(\cdot)$ is not straightforward → needs the use of an acceptance-rejection technique, the complexity of which can be dramatic when the probability of acceptance is relatively small.
- Computations are much simpler if the moment generating function $M(\theta)$ is known in closed-form. Such a requirement does not hold in general.
- $\theta$ does not have, in general, a closed-form expression, and hence, it should be approximated numerically.

Goal

Propose an alternative IS PDF that approximately yields, for certain classes of distributions and in the rare event regime corresponding to large $N$ and/or small $\gamma$, at least the same performance as the exponential twisting technique and at the same time does not introduce serious limitations.
$f_X(x) = x^p g(x)$ with $g(x) \sim b$ as $x$ goes to 0, $d > -1$, and $b > 0$

- Distributions in this class: Exponential, Gamma, Nakagami-m, Weibull, Generalized Gamma, Rice, Gamma-Gamma, $\kappa - \mu$.
- The change of measure that we consider is

$$\tilde{f}_X(x) = \frac{x^p \exp(\theta x)}{\tilde{M}(\theta)}, \ x \geq 0.$$  \hspace{1cm} (5)

- The IS PDF corresponds to the Gamma PDF with shape parameter $p + 1$ and scale parameter $-1/\theta$.
- The value of $\theta$ is chosen to be equal to $\tilde{\theta}$ such that

$$\tilde{\theta} = -\frac{N}{\gamma}(p + 1).$$  \hspace{1cm} (6)

- The proposed IS estimator of $\alpha(\gamma, N)$ using $M$ i.i.d. samples of $X$ from $\tilde{h}_X(x) = \prod_{i=1}^{N} \tilde{f}_X(x_i)$ is

$$\hat{\alpha}_{is,M}(\gamma, N) = \frac{1}{M} \sum_{k=1}^{M} \mathbf{1}_{(\sum_{i=1}^{N} X_i^{(k)} \leq \gamma)} (\tilde{M}(\tilde{\theta}))^N \prod_{i=1}^{N} \frac{f_X(X_i^{(k)}) \exp \left(-\tilde{\theta} X_i^{(k)}\right)}{(X_i^{(k)})^p}.$$
Main Result
For distributions satisfying $f_X(x) = x^p g(x)$ with $g(x) \sim b$ as $x$ goes to 0, $p > -1$, and $b > 0$ is a constant, the proposed approach achieves approximately, as $\gamma$ decreases to 0 and/or $N$ increases, the same performance as the one given by the exponential twisting without introducing serious limitations.
Distributions that do not approach 0 polynomially are much more
difficult to handle and need to be tackled on a case-by-case basis.

We consider the case of the sum of i.i.d. standard Log-normal RVs whose
PDF decreases to 0 at a faster rate than any polynomials.

The Gamma distribution with fixed shape parameter will not recover the
results given by the use of the exponential twisting technique.

In Asmussen et al. 2016, the exponential twisting was applied by:

1. Providing an unbiased estimator of the moment generating function.
2. Approximating the value of $\theta$.
3. Using acceptance-rejection to sample from the IS PDF.
We propose two estimators:

- The first is based on truncating the support $[0, +\infty]$ and only working on $[a, +\infty]$ with $a = \delta \gamma / N$
  → introduces a bias that needs to be controlled.

- In the regime of rare events, the biased estimator achieves approximately the same performances as the Gamma IS PDF with shape parameter equal to 2
  → propose a second estimator whose IS PDF is a Gamma PDF with optimized parameters.
Standard Log-normal case: biased estimator

- We rewrite the quantity of interest as

\[
\mathbb{P}_{h_X} \left( \sum_{i=1}^{N} X_i \leq \gamma \right) \approx \left( 1 - F_X \left( \frac{\delta \gamma}{N} \right) \right)^N \tag{Known in closed form}
\]

\[
\times \mathbb{P}_{h_X} \left( \sum_{i=1}^{N} X_i \leq \gamma \bigg| X_i > \frac{\delta \gamma}{N} \right), \tag{7}
\]

where \( \delta \in [0, 1) \).

- Let \( \bar{f}_X(\cdot) \) be the PDF of \( X_i|\{X_i > \frac{\delta \gamma}{N}\} \), \( i = 1, 2, \ldots, N \)

\[
\mathbb{P}_{h_X} \left( \sum_{i=1}^{N} X_i \leq \gamma \bigg| X_i > \frac{\delta \gamma}{N} \right) = \mathbb{P}_{\bar{h}_X} \left( \sum_{i=1}^{N} X_i \leq \gamma \right),
\]

with \( \bar{h}_X(x) = \prod_{i=1}^{N} \bar{f}_X(x_i) \). The exponential twisting change of measure is then given by

\[
\bar{f}_X^*(x) \propto \bar{f}_X(x) \exp(\theta x), \quad x \geq \frac{\delta \gamma}{N}.
\]
Standard Log-normal case: biased estimator

- The approximate exponential twisting change of measure is given by
  \[
  \tilde{f}_X(x) = \frac{\bar{f}_X \exp(\theta x) + (x - \frac{\delta \gamma}{N}) \bar{f}'_X \exp(\theta x)}{\tilde{M}(\theta)}, \quad x \geq \frac{\delta \gamma}{N},
  \]
  (8)
  with \( \bar{f}_X = \bar{f}_X(\frac{\delta \gamma}{N}) \) and \( \bar{f}'_X = \bar{f}'_X(\frac{\delta \gamma}{N}) \). We assume that \( \frac{\delta \gamma}{N} < \exp(-1) \) to ensure that \( \bar{f}'_X > 0 \).

- Through a simple computation, we get
  \[
  \tilde{M}(\theta) = -\frac{\exp(\theta \delta \gamma/N)}{\theta} \bar{f}_X + \frac{\exp(\theta \delta \gamma/N)}{\theta^2} \bar{f}'_X.
  \]

- The value of \( \theta \) that solves \( \frac{\tilde{M}'(\theta)}{\tilde{M}(\theta)} = \frac{\gamma}{N} \) is given by
  \[
  \theta = -\frac{\bar{f}_X - c \bar{f}'_X + \sqrt{(\bar{f}_X - c \bar{f}'_X)^2 + 8 \bar{f}_X \bar{f}'_X c}}{2c \bar{f}_X},
  \]
  with \( c = \frac{\gamma}{N} (1 - \delta) \).
Standard Log-normal case: biased estimator

- The question that remains is related to controlling the bias through a proper choice of the parameter $\delta$.
- Let $\alpha_1(\gamma, N) = \left(1 - F_X(\frac{\delta \gamma}{N})\right)^N \mathbb{P}_{h_X}\left(\sum_{i=1}^N X_i \leq \gamma \mid X_i > \frac{\delta \gamma}{N}, \forall i\right)$

$$\left|\frac{\alpha(\gamma, N) - \hat{\alpha}_{1,\text{is},M}}{\alpha(\gamma, N)}\right| \leq \frac{\alpha(\gamma, N) - \alpha_1(\gamma, N)}{\alpha(\gamma, N)} + \left|\frac{\alpha_1(\gamma, N) - \hat{\alpha}_{1,\text{is},M}}{\alpha_1(\gamma, N)}\right|,$$ (9)

where $\hat{\alpha}_{1,\text{is},M}$ is the IS estimator of $\alpha_1(\gamma, N)$ based on $M$ i.i.d. realizations sampled according to $\tilde{h}_X(x) = \prod_{i=1}^N \tilde{f}_X(x_i)$.

Lemma: On the choice of $\delta$

The following expression of $\delta(\epsilon, N, \gamma)$

$$\delta(\epsilon, N, \gamma) = \frac{N}{\gamma} \exp\left(\Phi^{-1}\left(\frac{\epsilon}{2N} \left(\frac{\Phi(\log(\gamma/N))}{\Phi(\log(\gamma))} \right)^N\right)\right),$$ (10)

where $\Phi(\cdot)$ is the CDF of the standard Normal distribution, ensures that the relative bias is less than $\epsilon/2$. 
Standard Log-normal case: The Gamma family as an IS PDF

- For a sufficiently small value of $\delta$, the biased estimator achieves approximately the same performance as the Gamma IS PDF with shape parameter equal to 2 $\rightarrow$ suggests using the Gamma family as IS PDF.
- Note that the advantage of using the Gamma family as IS PDFs is that the estimator is unbiased.
- Recall that the Gamma PDF is given by

$$
\tilde{f}_X(x) = \frac{x^{k-1} \exp(-x/\theta)}{\Gamma(k)\theta^k}, \quad x > 0,
$$

where $\theta > 0$ and $k > 0$ are the scale and shape parameters.
- The value of $\theta$ is chosen to be equal to $\theta = \frac{\gamma}{Nk}$ to ensure that the expected value of each of the $X_i$’s, $i = 1, 2, \cdots, N$, under the PDF $\tilde{f}_X(\cdot)$ is equal to $\frac{\gamma}{N}$.
- The likelihood ratio is then given by

$$
\mathcal{L}(x_1, x_2, \cdots, x_N) = \frac{(\Gamma(k)\theta^k)^N \exp\left(\frac{\sum_{i=1}^{N} x_i}{\theta} - \frac{1}{2} \sum_{i=1}^{N} (\log(x_i))^2\right)}{\prod_{i=1}^{N} x_i^k (\sqrt{2\pi})^N}.
$$
Standard Log-normal case: The Gamma family as an IS PDF

- The second moment of the IS estimator is bounded by
  \[
  \mathbb{E}_{\tilde{h}_X} \left[ \mathcal{L}^2(X_1, X_2, \cdots, X_N) \mathbf{1}_{(\sum_{i=1}^N X_i \leq \gamma)} \right] 
  \leq \left( \frac{\Gamma(k)(\frac{\gamma}{Nk})^k}{\sqrt{2\pi}} \right)^{2N} \exp(2kN + k^2N).
  \]

- Using Stirling’s formula, we get for some constant \( C \)
  \[
  \mathbb{E}_{\tilde{h}_X} \left[ \mathcal{L}^2(X_1, X_2, \cdots, X_N) \mathbf{1}_{(\sum_{i=1}^N X_i \leq \gamma)} \right] 
  \lesssim C \exp(N(k^2 - 2k \log(N/\gamma) - \log(k))).
  \]

- \( k \) is chosen such that it minimizes the above right-hand side term
  \[
  k^* = \frac{1}{2} \left( \log\left(\frac{N}{\gamma}\right) + \sqrt{(\log\left(\frac{N}{\gamma}\right))^2 + 2} \right).
  \] (12)

- When \( N \) is large and/or \( \gamma \) is small, we obtain \( k^* \sim \log\left(\frac{N}{\gamma}\right) \).
We define the squared coefficient of variation of an unbiased estimator \( \hat{\alpha}(\gamma, N) \) of \( \alpha(\gamma, N) \) as

\[
SCV(\hat{\alpha}(\gamma, N)) = \frac{\text{var} [\hat{\alpha}(\gamma, N)]}{\alpha^2(\gamma, N)}.
\] (13)

From the CLT, the number of required samples to meet \( \epsilon \) relative statistical error with 95% confidence is equal to \((1.96)^2 SCV(\hat{\alpha}(\gamma, N))/\epsilon^2\).

**Weibull Case**

- \( X_i, i = 1, 2, \cdots, N, \) are distributed according to the Weibull distribution whose PDF is

\[
f_X(x) = \frac{k}{\lambda} \left( \frac{x}{\lambda} \right)^{k-1} \exp \left( -\left( \frac{x}{\lambda} \right)^k \right), \ x > 0.
\]

- The comparison is made with respect to the hazard rate twisting (HRT) approach of Ben Rached et al. 2016.
Weibull Case

When $k = 1.5$, $\lambda = 1$, $\gamma = 0.5$, and $N = 12$, the proposed approach is approximately $270$ times more efficient than the one based on HRT.

The value of $\alpha(\gamma, N)$ ranges approximately from $10^{-20}$ to $10^{-6}$.

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

**Figure 2:** Squared coefficient of variation as a function of $N$ where $X_i$ are i.i.d. Weibull RVs with rate $\lambda = 1$, $k = 0.5$, and $\gamma = 0.01$.

The value of $\alpha(\gamma, N)$ ranges approximately from $10^{-16}$ to $10^{-6}$.
Weibull Case

The value of $\alpha(\gamma, N)$ ranges approximately from $10^{-16}$ to $10^{-6}$ for $N = 8$ and from $10^{-22}$ to $10^{-8}$ for $N = 10$.

**Figure 3:** Squared coefficient of variation as a function of $\gamma$ where $X_i$ are i.i.d. Weibull RVs with rate $\lambda = 1$, $k = 1.5$.

Approximately 18 (respectively 64) times more efficient than the HRT one when $N = 8$ (respectively $N = 10$) and $\gamma = 0.2$. 
Log-normal Case

- The standard Log-normal PDF is given by
  \[ f_X(x) = \frac{1}{x \sqrt{2\pi}} \exp\left(-\frac{(\log(x))^2}{2}\right), \quad x > 0. \]

- Comparison with exponential twisting proposed in Asmussen et al. 2016.

**Figure 4**: Squared coefficient of variation as a function of \( N \) where \( X_i \) are i.i.d. standard Log-normal RVs with \( \gamma = 0.5 \), and \( \epsilon = 0.05 \). When \( N = 9 \), approximately 30 times more efficient than the one based on exponential twisting.

The value of \( \alpha(\gamma, N) \) ranges approximately from \( 10^{-20} \) to \( 10^{-2} \).
Log-normal Case

The quantity of interest $\alpha(\gamma, N)$ ranges approximately from $10^{-15}$ to $10^{-6}$ for $N = 8$ and from $10^{-21}$ to $10^{-9}$ for $N = 10$.

Figure 5: Squared coefficient of variation as a function of $\gamma$ where $X_i$ are i.i.d. standard Log-normal random variables with $\epsilon = 0.05$. Approximately 15 times (respectively 35) more efficient that the exponential twisting one when $N = 8$ (respectively $N = 10$) and $\gamma = 0.6$. 
Log-normal Case

- In order to include the computing time in our comparison, we define the Work Normalized Relative Variance (WNRV) metric of an unbiased estimator $\hat{\alpha}(\gamma, N)$ of $\alpha(\gamma, N)$ (Ben Rached et al. 2018)

$$\text{WNRV}(\hat{\alpha}(\gamma, N)) = \frac{SCV(\hat{\alpha}(\gamma, N))}{M} \times \text{computing time in seconds.}$$

- The computing time is the time in seconds needed to get an estimator of $\alpha(\gamma, N)$ using $M$ i.i.d. samples of $\hat{\alpha}(\gamma, N)$.

- When comparing two estimators, the one that exhibits less WNRV is more efficient than the other estimator: it achieves less relative error for a given computational budget, or equivalently it needs less computing time to achieve a fixed relative error.
Log-normal Case

Figure 6: WNRV as a function of $\gamma$ where $X_i$ are i.i.d. standard Log-normal RVs with $\epsilon = 0.05$.

For $N = 10$ and $\gamma = 0.6$, the approach based on using the Gamma IS PDF is approximately 35 times (respectively 340 times) more efficient than the one based on exponential twisting when using the squared coefficient of variation metric (respectively the WNRV metric).
Developed efficient importance sampling estimators to estimate rare event probabilities corresponding to the left-tail of the CDF of large sums of nonnegative i.i.d. RVs.

Achieved at least the same performance as the exponential twisting technique.

The Gamma PDF with suitably chosen parameters achieves for most of the well-practical distributions substantial amount of variance reduction, and at the same time avoids the restrictive limitations when using the exponential twisting technique.

Validated numerically the efficiency in the rare event regime corresponding to large $N$ and/or small $\gamma$.

**A possible future direction:** study the asymptotic properties of the proposed estimators as $N \to \infty$ and/or $\gamma \to 0$.

References


References


References


Thank you for your attention