Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 0946 – 8633

On non-asymptotic optimal stopping criteria in Monte Carlo simulations

Christian Bayer[1], Håkon Hoel[2], Erik von Schwerin[3], Raúl Tempone[2]

submitted: March 26, 2013

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: christian.bayer@wias-berlin.de

[2] Division of Mathematics
King Abdullah University of Science and Technology
Thuwal 23955-6900
Kingdom of Saudi Arabia
E-Mail: hakon.hoel@kaust.edu.sa
raul.tempone@kaust.edu.sa

[3] MATHICSE-CSQI
EPF de Lausanne
Switzerland
E-Mail: erik.vonschwerin@epfl.ch

No. 1774
Berlin 2013

2010 *Mathematics Subject Classification.* Primary 65C05; Secondary 62L12, 62L15.

Key words and phrases. Monte Carlo methods, optimal stopping, sequential stopping rules, non-asymptotic.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

We consider the setting of estimating the mean of a random variable by a sequential stopping rule Monte Carlo (MC) method. The performance of a typical second moment based sequential stopping rule MC method is shown to be unreliable in such settings both by numerical examples and through analysis. By analysis and approximations, we construct a higher moment based stopping rule which is shown in numerical examples to perform more reliably and only slightly less efficiently than the second moment based stopping rule.

1 Introduction

Given i.i.d. random variables $X_1, X_2, \ldots$ the typical way of approximating their expected value $\mu = \mathrm{E}[X]$ using M samples is the sample average

$$\overline{X}_M := \sum_{i=1}^{M} \frac{X_i}{M}.$$

We consider the objective of choosing M sufficiently large so that the error probability satisfies

$$\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right) \leq \delta, \tag{1}$$

for some fixed small constants $\mathrm{TOL} > 0$ and $\delta > 0$. Clearly, $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right)$ decreases as M increases, but at the same time the cost of computing $\overline{X}_M$ increases. From an application and cost point of view it is therefore of interest to derive theory or algorithms that will give upper bounds on M satisfying (1) that are not far too large. When a-priori information about the distribution of X is available, for example if X is a bounded r.v. with an explicitly given bound, it is possible to derive good theoretical upper bounds for M using Hoeffding type inequalities, cf. Hoeffding [7]. In the general case when no or little information of the distribution is given, little theory is however known, and the typical way of estimating $\mathrm{E}[X]$ using a sufficiently large number of samples M is through a sequential stopping rule. Below we give the general structure of the class of sequential stopping rules we have in mind.

(I) Generate a batch of M i.i.d. samples $X_1, X_2, \ldots, X_M$.

(II) Infer distributive properties of $\overline{X}_M$ from the generated batch of samples through higher order sample moments, e.g. the sample mean and the sample variance.

(III) Based on the sample moments, estimate the error probability. When, based on the estimated probability, (1) is violated, increase the number of samples M and return to step (I).
Else, break and accept M.

Certainly the most natural and important representative of this class of algorithms is given in Algorithm 1. The algorithm estimates the error probability by appealing to the Central Limit Theorem (CLT). Consequently, it only relies on the sample variance in addition to the sample mean. In particular, the algorithm only requires mild additional assumptions on X, namely square integrability.

In the literature, various second moment based sequential stopping rules have been introduced to estimate the steady-state mean of stochastic processes, see for example Law and Kelton [10, 11] for comparisons of the performance of different stopping rules and Bratley, Bennet, and Fox [2] for an overview. Second moment based sequential stopping rules generally tend to perform well in the asymptotic regime when $\mathrm{TOL} \to 0$. In fact, Chow and Robbins [3] proved that under very loose restrictions, second moment based sequential stopping rules such as Algorithm 1 are asymptotically consistent, meaning that for a fixed δ,

$$\lim_{\mathrm{TOL}\to 0} \mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right) = \delta,$$

Algorithm 1 Sample Variance Based Stopping Rule

Input: Number of samples M_0, accuracy TOL, confidence δ, the cumulative distribution function of the standard normal distributed r.v. $\Phi(x)$.
Output: $\overline{X}_M$.

Set $n = 0$, generate M_n samples $\{X_i\}_{i=1}^{M_n}$ and compute the sample variance

$$\overline{\sigma}^2_{M_n} := \frac{1}{M_n - 1} \sum_{i=1}^{M_n} (X_i - \overline{X}_{M_n})^2. \tag{2}$$

while $2\left(1 - \Phi(\sqrt{M_n}\,\mathrm{TOL}\,/\overline{\sigma}_{M_n})\right) > \delta$ **do**
  Set $n = n + 1$ and $M_n = 2M_{n-1}$.
  Generate a batch of M_n i.i.d. samples $\{X_i\}_{i=1}^{M_n}$.
  Compute the sample variance $\overline{\sigma}^2_{M_n}$ as given in (2).
end while
Set $M = M_n$, generate samples $\{X_i\}_{i=1}^{M}$ and compute the output sample mean $\overline{X}_M$. (See Section 2 for a motivation of the choice of the stopping criterion in the while loop above.)

and in Glynn and Whitt [5] the consistency property is proven to hold for such stopping rules applied to more general stochastic processes. In the non-asymptotic regime – when TOL and δ are finite – Hickernell et al. [6] recently developed a second moment based MC method which guarantees to meet condition (1) under the assumption that an upper bound is given for the kurtosis of the r.v. to be sampled prior to sampling it. When no moment bounds are given prior to sampling, however, Bahadur and Savage [1] proved that it is not possible to develop an algorithm guaranteed to meet condition (1). This is somewhat unsatisfactory as in applications the non-asymptotic regime with little prior information on the r.v. is a setting we believe is often encountered. While consistency is clearly a reassuring property in any case, in many situations one is in dire need of quantitative estimates of the error probability in the non-asymptotic regime, for instance when one tries to optimize the computational cost needed to meet a certain accuracy target using an adaptive algorithm. We could not find such a quantitative, non-asymptotic analysis of sequential stopping algorithms like Algorithm 1 in the literature.

In this work we demonstrate by numerical examples that second moment based stopping rules can fail convincingly in the non-asymptotic regime, especially when the underlying distribution X is heavy-tailed, see Section 2. We proceed by giving an error analysis of Algorithm 1 specifically in the non-asymptotic regime. We note a-priori that there are two obvious approximation errors in the underlying assumptions of Algorithm 1:

(I) The algorithm appeals to the CLT to approximate the tail probabilities for $\overline{X}_M$ even though M is finite.

(II) In doing so, it uses the sample variance $\overline{\sigma}^2_M$ instead of the true variance σ^2.

To get a hold on the error probability (1) despite the fact that the distribution of the sample mean $\overline{X}_M$ is unknown, we again appeal to the central limit theorem, but we adjust the estimate by adding a Berry-Esseen type term, which extends the validity of the estimate to the non-asymptotic case, thereby dealing with the first approximation error. As the error probability (1) is a tail probability for the distribution of the sample mean and the Berry-Esseen theorem itself is rather aimed at being sharp at the center of the distribution, we appeal to non-uniform versions of the Berry-Esseen theorem, see Theorem 1.1 and Corollary 1.2 below. However, both intuition and numerical tests suggest that the approximation of the tail probabilities by the non-uniform Berry-Esseen theorem is far too pessimistic at least when the second approximation error is small, i.e., when the computed sample variance is actually close to the true variance. In this case, we adjust the normal distribution by a less stringent

extra term, which is obtained from an Edgeworth expansion of the distribution function of the sample mean $\overline{X}_M$, cf. Feller [4].[1]

Having identified possible origins of failure of Algorithm 1, we propose an improvement of Algorithm 1. However, this variant requires third and fourth sample moments, see Section 4. Finally, in Section 5, we test the new algorithm numerically. We find that the new stopping Algorithm 2 indeed satisfies the desired confidence level δ on the error probability (1) even when $\delta \ll \mathrm{TOL}$.

As already discussed above, we need to approximate the unknown distribution of a sample mean in a general, non-asymptotic regime. The uniform and non-uniform Berry-Esseen theorems provide quantitative bounds for the difference between the true distribution of the sample mean and its asymptotic limit, namely the normal distribution. The following Berry-Esseen bounds, which can be found in their classical formulations in Petrov [13], is here presented with the present optimal bound constants.

Theorem 1.1 (Uniform and Non-Uniform Berry-Esseen). *Suppose $X_1, X_2, \ldots$ are i.i.d. r.v. with $E[X] = 0$, $\sigma^2 = Var(X)$ and $\beta = \frac{E[|X|^3]}{\sigma^3} < \infty$. Then the following uniform bound*

$$\left| P\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) - \Phi(x) \right| \leq 0.3328 \cdot \frac{\beta + 0.429}{\sqrt{n}}$$

holds, cf. Shevtsova [14]. Furthermore, the following non-uniform bound holds

$$\left| P\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) - \Phi(x) \right| \leq 18.1139 \cdot \frac{\beta}{\sqrt{n}(1 + |x|^3)},$$

cf. Hickernell et al. [6], and the references therein.

For the purpose of this work, it will be useful to combine the uniform and non-uniform Berry-Esseen bound as follows.

Corollary 1.2 (Berry-Esseen). *Suppose $X_1, X_2, \ldots$ are i.i.d. r.v. with $E[X] = 0$. Then*

$$\left| P\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) - \Phi(x) \right| \leq \frac{C_{\mathrm{BE}}(x, \beta)}{\sqrt{n}}$$

where the bound function $C_{\mathrm{BE}} : \mathbb{R} \to [0, C_0]$ is defined by

$$C_{\mathrm{BE}}(x, \beta) := \min\left(0.3328 \cdot (\beta + 0.429), 18.1139 \cdot \frac{\beta}{(1 + |x|^3)} \right).$$

In the asymptotic regime, the distribution of $\mathrm{P}\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right)$ can be expressed by so called Edgeworth expansions. Here we present the one-term Edgeworth expansion.

Theorem 1.3 (Edgeworth expansion, cf. Feller [4] p. 541). *Suppose $X_1, X_2, \ldots$ are i.i.d. r.v. with a distribution which is not a lattice distribution and $E[X] = 0$, $\sigma^2 = Var(X)$ and $E\left[X^3\right] < \infty$. Then*

$$P\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) = \Phi(x) + \frac{(x^2 - 1)e^{-x^2/2}}{6\sqrt{2\pi n}} \frac{E\left[X^3\right]}{\sigma^3} + o\left(n^{-1/2}\right),$$

uniformly for $x \in \mathbb{R}$.

[1] Note that here we are introducing a gap in the analysis: the estimate based on the non-uniform Berry-Esseen theorem is reliable in the sense that it always leads to an upper bound of the error probability (1). For the Edgeworth expansion, however, there might be situations when the true error probability is underestimated, and, consequently, the accuracy target might still be missed. Numerical evidence, however, suggests that the estimate obtained from solely relying on the non-uniform Berry-Esseen theorem is usually by orders of magnitude too pessimistic. Apart from intrinsic reasons, one reason might be that the constants known in the non-uniform Berry-Esseen theorems might be far from being optimal. In the end, we think that the above compromise between Berry-Esseen type estimations and estimations based on the Edgeworth expansion might be a good compromise which retains the goal of reliably meeting the accuracy target – except maybe for very extreme situations – while keeping a certain level of efficiency. We note, however, that it is also possible to construct even more conservative stopping rules which are only based on the Berry-Esseen theorem.

2 Stopping rule failures

Suppose we seek to estimate $\mu = \mathrm{E}[X]$ using Monte Carlo simulation and we actually *do know* the variance $\sigma^2 = \mathrm{Var}(X)$. As before, our objective is to achieve $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right) \leq \delta$, for some fixed, small constants $\mathrm{TOL}, \delta > 0$. The CLT motivates the stopping rule

$$M = \frac{C_{\mathrm{CLT}}^2 \sigma^2}{\mathrm{TOL}^2}, \quad C_{\mathrm{CLT}} := \Phi^{-1}\left(\frac{2-\delta}{2}\right), \tag{3}$$

which would exactly fulfill our objective (1) in the asymptotic regime $M \to \infty$. Of course, this conflicts with our choice (3) for M, since we treat δ and TOL as finite constants. However, we can still estimate the probability in (1) using Corollary 1.2 and obtain

$$\begin{aligned}
\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right) &= \mathrm{P}\left(\sqrt{M}\frac{\left|\overline{X}_M - \mu\right|}{\sigma} > \frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right) \\
&\leq 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right)\right) + 2C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}, \beta\right)\frac{1}{\sqrt{M}} \\
&= 2(1 - \Phi(C_{\mathrm{CLT}})) + 2C_{\mathrm{BE}}(C_{\mathrm{CLT}}, \beta)\,\frac{1}{\sqrt{M}} \\
&= \delta + 2\frac{C_{\mathrm{BE}}(C_{\mathrm{CLT}}, \beta)}{\sigma C_{\mathrm{CLT}}}\,\mathrm{TOL}
\end{aligned} \tag{4}$$

This means that in the worst case, the actual error probability could be $\delta + O(\mathrm{TOL})$ instead of δ.[2] For instance, in situations where the statistical confidence in the result is more stringent than the accuracy so that $\delta \ll \mathrm{TOL}$, the asymptotically motivated choice of M in (3) could, granted the bound (4) is sharp, fail to deliver the expected level of confidence. For most r.v. however, the bound (4) is far too conservative, and one might ask whether it is reasonable to fear underestimating the error probability in the fashion we have described. The following numerical example shows the existence of r.v. for which the stopping rule (3) fails in the non-asymptotic regime

Example 2.1. The heavy-tailed Pareto-distribution has the probability distribution function

$$f(x) = \begin{cases} \alpha x_m^{\alpha} x^{-(\alpha+1)} & \text{if } x \geq x_m \\ 0 & \text{else,} \end{cases} \tag{5}$$

where $\alpha, x_m \in \mathbb{R}_+$ are respectively the shape and the scale parameter. The moments of $\mathrm{E}[X^n]$ for the Pareto r.v. only exists for $n < \alpha$ and, supposing $\alpha > 2$, its mean and variance are given by

$$\mu = \frac{\alpha x_m}{\alpha - 1} \text{ and } \sigma^2 = \frac{x_m^2 \alpha}{(\alpha-1)^2(\alpha-2)}.$$

It is further easy to derive that for a Pareto r.v. with $\alpha = 3 + \gamma$ and $0 < \gamma < 1$,

$$\beta = \frac{\mathrm{E}\left[|X-\mu|^3\right]}{\sigma^3} = O\left(\gamma^{-1}\right).$$

This implies that there exists r.v. for which the second summand of the bound (4) can become arbitrary large. So for such r.v. the stopping rule (3) might fail. Let us investigate by numerical approximations. Considering the distribution with $\alpha = 3.1$ (and $x_m = 1$), yields a heavy-tailed r.v. with known mean, variance and third moment. For a set of accuracies $\mathrm{TOL} \in [0.05, 0.2]$ and confidences $\delta = \mathrm{TOL}^{\ell}$, $\ell = 1$ and 2 we have numerically approximated $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right) \leq \delta$ using, in accordance with (3), the stopping rule

$$M = \left\lceil \frac{C_{\mathrm{CLT}}^2 \sigma^2}{\mathrm{TOL}^2} \right\rceil$$

The results, illustrated in Figure 1, show that when $\delta \ll \mathrm{TOL}$, the sought confidence is far from met.

[2] Note that C_{CLT} as defined in (3) grows only very slowly as δ decreases, since we have $C_{\mathrm{CLT}} < \sqrt{2\log(\delta^{-1})}$. So as a function of δ, the factor in front of TOL is approximately constant.

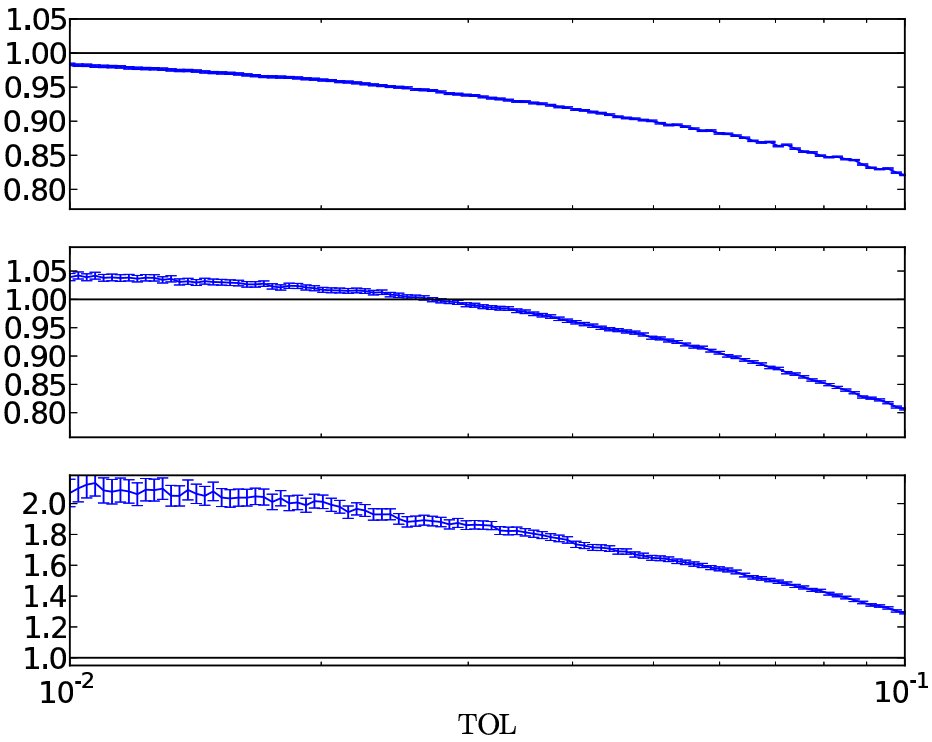

Figure 1: MC estimate using the stopping rule (3) for i.i.d. Pareto r.v. with parameters $\alpha = 3.1$ and $x_m = 1$. The quantity $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right)/\delta$ (blue line) is plotted for the settings $\delta(\mathrm{TOL}) = \mathrm{TOL}^\ell$ with $\ell = 0.5$ (top), $\ell = 1$ (middle), and $\ell = 2$ (bottom). The probability of failure is estimated by $\overline{p}_N(\mathrm{TOL}) = N^{-1}\sum_{i=1}^{N} \mathbf{1}_{\left|\overline{X}_M(\omega_i)-\mu\right|>\mathrm{TOL}}$ with $N = 10^7$, and the error bars for the estimate of $\overline{p}_N$ by $1.96\sqrt{\overline{p}_N(1-\overline{p}_N)/N}$. The stopping rule fails when $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right)/\delta > 1$. We observe that the smaller δ is relative to TOL, the larger is the probability of failure for the stopping rule.

The demonstrated stopping rule failure motivated us to study and develop ways of constructing more reliable stopping rules. In Section 3, we first analyze the stopping rule of Algorithm 1, and derive an approximate upper bound for the failure probability expressed in terms of M, TOL and δ. In Section 4, we develop a more reliable stopping rule algorithm, which in addition to sampled second moments of the r.v. in question also depends on sampled third and fourth moments. The paper is concluded with numerical examples comparing the reliability and computational cost of Algorithm 1 with the stopping rule developed in Section 4.

3 Error analysis for Algorithm 1

Example 2.1 illustrate that for some r.v. the stopping rule in Algorithm 1 does not meet the accuracy-confidence constraint (1). To construct a more reliable stopping rule, penalty terms have to be added to the stopping criterion in Algorithm 1. Some care should be taken to make the penalty terms of the right size: if too large penalties are added, the new stopping rule will be reliable but very inefficient, while if too small penalty terms are added, the algorithm will of course be efficient but unreliable.

In this section, we first derive an approximate upper bound for the failure probability

$$\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\,\middle|\, M\right) \tag{6}$$

corresponding to the stopping rule of Algorithm 1 conditional on the (random) final number of samples M. Clearly, the bound for (6) will also be a r.v. Using the bound for (6), we thereafter construct reasonable penalty terms to be added to the stopping criterion of our new stopping rule.

As a general idea, we are going to use a weighted average of one very reliable, but typically overly pessimistic error bound based on the Berry-Esseen theorem, and another error bound based on the Edgeworth expansion, which is typically too optimistic. The (critical) choice of the weights is based on the following consideration: If we have successfully estimated the variance of X to high accuracy, then we give ourselves some leeway for the bias and choose the more optimistic bound. If, on the other hand, we have already mis-estimated the variance σ^2, then we want to be highly conservative in our conditional bias estimate. Thus, our first step towards an upper bound for (6) is partitioning the probability (6) into two parts as follows

$$\begin{aligned}\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\,\middle|\, M\right) = {} & \mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\,\middle|\, M\right)\mathrm{P}\left(|\overline{\sigma}^2_M - \sigma^2| \geq \sigma^2/2\middle| M\right) \\ & + \mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\,\middle|\, M\right)\mathrm{P}\left(|\overline{\sigma}^2_M - \sigma^2| < \sigma^2/2\middle| M\right).\end{aligned} \tag{7}$$

In the next step, we will use two different (asymptotic) error bounds for $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\,\middle|\, M\right)$, a conservative one weighted by $\mathrm{P}\left(|\overline{\sigma}^2_M - \sigma^2| \geq \sigma^2/2\middle| M\right)$ and an efficient one weighted by $\mathrm{P}\left(|\overline{\sigma}^2_M - \sigma^2| < \sigma^2/2\middle| M\right)$. To derive the conservative error bound, we recall that in Algorithm 1 the samples used in the output estimate $\overline{X}_M$ and for $\overline{\sigma}_M$ are independent of the samples used to determine M. Keeping this in mind, we derive the following approximate upper bound

$$\begin{aligned}\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\,\middle|\, M\right) = {} & \mathrm{P}\left(\left|\overline{X}_n - \mu\right| > \mathrm{TOL}\right)\Big|_{n=M} \\ \lessapprox {} & 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right) + C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}, \beta\right)\frac{1}{\sqrt{M}}\right).\end{aligned} \tag{8}$$

Here the Berry-Esseen bound of Corollary 1.2 was used to derive the approximate bound of the last line.

For the weighting factor, we obtain the following equality

$$\mathrm{P}\left(|\overline{\sigma}^2_M - \sigma^2| \geq \sigma^2/2\middle| M\right) = \mathrm{P}\left(|\overline{\sigma}^2_n - \sigma^2| \geq \sigma^2/2\right)\Big|_{n=M}.$$

Furthermore, using Chebycheff's inequality for k-statistics to bound the variance of the sample variance, cf. Keeping [9], we derive that

$$\begin{aligned}\mathrm{P}\left(|\overline{\sigma}^2_M - \sigma^2| \geq \sigma^2/2\middle| M\right) = {} & \mathrm{P}\left(|\overline{\sigma}^2_n - \sigma^2| \geq \sigma^2/2\right)\Big|_{n=M} \\ \leq {} & 4\,\mathrm{E}\left[\frac{\left|\sigma^2 - \overline{\sigma}^2_n\right|^2}{\sigma^4}\right]\Bigg|_{n=M} \leq 4\frac{\sigma^4\left(\frac{2}{M-1} + \frac{\kappa}{M}\right)}{\sigma^4} = 4\left(\frac{2}{M-1} + \frac{\kappa}{M}\right).\end{aligned}$$

Here κ denotes the *kurtosis*, i.e.,

$$\kappa = \frac{\mathrm{E}\left[|X - \mu|^4\right]}{\sigma^4} - 3.$$

More generally, we might consider Markov's inequality for a different p than two – for instance, because the kurtosis fails to be finite. Then we can use

$$\mathrm{P}\left(|\overline{\sigma}^2_M - \sigma^2| \geq \sigma^2/2\middle| M\right) \leq 2^p \frac{\mathrm{E}\left[\left|\overline{\sigma}^2_M - \sigma^2\right|^p \,\middle|\, M\right]}{\sigma^{2p}}.$$

In particular, we are free to choose $p < 2$, if the fourth moment of the distribution might not exist. We conclude that

$$\mathsf{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2 \Big| M\Big) \leq \min\Big(1,\ 4\Big(\frac{2}{M-1} + \frac{\kappa}{M}\Big),$$
$$\frac{2^{p_1}}{\sigma^{2p_1}}\mathsf{E}\Big[\,|\overline{\sigma}_M^2 - \sigma^2|^{p_1} \,\Big|\, M\Big], \ldots, \frac{2^{p_n}}{\sigma^{2p_n}}\mathsf{E}\Big[\,|\overline{\sigma}_M^2 - \sigma^2|^{p_n} \,\Big|\, M\Big]\Big) =: C_P(2, p_1, \ldots, p_n; M), \quad (9)$$

for some finite sequence $p_1, \ldots, p_n > 0$, with the natural interpretation of $\kappa = \infty$, if fourth moments do not exist.

Next, we come to the optimistic bound of the error probability

$$\mathsf{P}\Big(|\overline{X}_M - \mu| > \mathrm{TOL} \,\Big|M\Big)$$

based on the Edgeworth expansion given in Theorem 1.3. The mild penalty obtained in this way is of the form

$$\begin{aligned}
&\mathsf{P}\Big(|\overline{X}_M - \mu| > \mathrm{TOL} \,\Big|M\Big) \\
&= \mathsf{P}\Big(\sqrt{n}\frac{|\overline{X}_n - \mu|}{\sigma} > \frac{\sqrt{n}\,\mathrm{TOL}}{\sigma}\Big)\Big|_{n=M} \\
&\lessapprox 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right) + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\sigma^2} - 1\right|\exp\left(-\frac{M\,\mathrm{TOL}^2}{\sigma^2}\right)}{6\sqrt{2\pi M}}\frac{\left|\mathsf{E}\big[(X-\mu)^3\big]\right|}{\sigma^3}\right).
\end{aligned} \quad (10)$$

Combining (8), (9) and (10), and noting that for all $x \in \mathbb{R}_+$ and $n \in \mathbb{N}$,

$$\frac{|x^2 - 1|\, e^{-x^2/2}}{6\sqrt{2\pi n}}\frac{\left|\mathsf{E}\big[(X-\mu)^3\big]\right|}{\sigma^3} \leq C_{\mathrm{BE}}(x,\beta)\,\frac{1}{\sqrt{n}},$$

we obtain the following approximate bound for failing to meet the accuracy of Algorithm 1 conditioned on the stopped number of samples M:

$$\begin{aligned}
\mathsf{P}\Big(|\overline{X}_M - \mu| > \mathrm{TOL} \,\Big|M\Big) \lessapprox\ & 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right)\right) + \\
& + 2C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}, \beta\right)\frac{1}{\sqrt{M}}C_P(2, p_1, \ldots, p_n; M) \\
& + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\sigma^2} - 1\right|\left|\mathsf{E}\big[(X-\mu)^3\big]\right|}{\exp\left(\frac{M\,\mathrm{TOL}^2}{2\sigma^2}\right) \times 3\sqrt{2\pi M}\sigma^3}\,(1 - C_P(2, p_1, \ldots, p_n; M)).
\end{aligned} \quad (11)$$

Remark 3.1. The bound from equation (11) is exact up to the higher order terms coming from the Edgeworth expansion. Note that $1 - C_P(2, p_1, \ldots, p_n; M)$ is not a lower bound for the complimentary error probability $\mathsf{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| < \sigma^2/2 \Big| M\Big)$, so we are even more conservative in our approach.

4 A higher moments based stopping rule

From the approximate stochastic error bound (11) we will in this section construct a new, more reliable stopping rule with a stopping criterion based on second, third, and fourth moments of the r.v. that is sampled. The sampled moments our new algorithm will depend on are (here represented in biased

form)

$$\overline{\sigma}_M := \sqrt{\sum_{i=1}^{M} \frac{(X_i - \overline{X}_M)^2}{M}}, \qquad \overline{\beta}_M := \sum_{i=1}^{M} \frac{|X_i - \overline{X}_M|^3}{M\overline{\sigma}_M^3}, \tag{12}$$

$$\widehat{\beta}_M := \sum_{i=1}^{M} \frac{(X_i - \overline{X}_M)^3}{M\overline{\sigma}_M^3}, \qquad \text{and} \quad \overline{\kappa}_M := \sum_{i=1}^{M} \frac{(X_i - \overline{X}_M)^4}{M\overline{\sigma}_M^4} - 3.$$

Replacing moments with sample moments in (11), we obtain a computable approximate stochastic error bound

$$\begin{aligned} &\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\middle| M\right) \\ &\lessapprox 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}\right)\right) + 2C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}, \overline{\beta}_M\right) \frac{1}{\sqrt{M}} \overline{C}_P(2, p_1, \ldots, p_n; M) \\ &\quad + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\overline{\sigma}_M^2} - 1\right| |\widehat{\beta}_M|}{\exp\left(\frac{M\,\mathrm{TOL}^2}{2\overline{\sigma}_M^2}\right) \times 3\sqrt{2\pi M}\overline{\sigma}_M^3} (1 - \overline{C}_P(2, p_1, \ldots, p_n; M)). \end{aligned} \tag{13}$$

The resulting approximate stochastic error bound will be implemented as the following stopping criterion in Algorithm 2:

$$\begin{aligned} &2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}\right)\right) + 2C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}, \overline{\beta}_M\right) \frac{1}{\sqrt{M}} \overline{C}_P(2, p_1, \ldots, p_n) \\ &\qquad + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\overline{\sigma}_M^2} - 1\right| |\widehat{\beta}_M|}{\exp\left(\frac{M\,\mathrm{TOL}^2}{2\overline{\sigma}_M^2}\right) \times 3\sqrt{2\pi M}\overline{\sigma}_M^3} (1 - \overline{C}_P(2, p_1, \ldots, p_n)) < \delta, \end{aligned} \tag{14}$$

where $\overline{C}_P(2, p_1, \ldots, p_n)$ is obtained from $C_P(2, p_1, \ldots, p_n)$ by replacing all moments of $\overline{\sigma}$ by their empirical counterparts.

We now present the new stopping rule algorithm.

Algorithm 2 Higher Moments Based Stopping Rule

Input: Accuracy TOL, confidence δ, and initial number of samples M_0.
Output: $\overline{X}_M$.

Set $n = 0$, generate i.i.d. samples $\{X_i\}_{i=1}^{M_n}$ and compute the sample moments $\overline{\sigma}_{M_n}$, $\overline{\beta}_{M_n}$, $\widehat{\beta}_{M_n}$ and $\overline{\kappa}_{M_n}$ and all (other) moments needed for $\overline{C}_P$ according to (12).

while Inequality (14) is not fulfilled. **do**

Set $n = n + 1$ and $M_n = 2M_{n-1}$.

Generate M_n i.i.d. samples $\{X_i\}_{i=1}^{M_n}$ and compute the sample moments $\overline{\sigma}_{M_n}$, $\overline{\beta}_{M_n}$, and $\overline{\kappa}_{M_n}$ and all (other) moments needed for $\overline{C}_P$.

end while

Set $M = M_n$, generate i.i.d. samples $\{X_i\}_{i=1}^{M}$ and return the sample mean $\overline{X}_M$.

Remark 4.1. The reasoning in Section 3 could also be used to construct alternative stopping rule algorithms. For instance, instead of using the respective probabilities for misestimation of σ, $\mathrm{P}\left(|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2 \middle| M\right)$ and $\mathrm{P}\left(|\overline{\sigma}_M^2 - \sigma^2| < \sigma^2/2 \middle| M\right)$ as respective weights for the conservative and the optimistic bound for the misestimation error probability for the mean μ itself, we could also take the argument more literally and

use the pessimistic error bound if we have misestimated the variance and the optimistic one otherwise, i.e., using random indicator functions instead of their expectations as weights. We leave this as a remark.

5 Numerical experiments

In the numerical experiments we will estimate the mean of four differently distributed r.v. by using three different stopping rules: the sample variance based stopping rule in Algorithm 1, the new higher moments based stopping rule in Algorithm 2 and the stopping rule recently introduced by Hickernell et al. [6] which they prove performs reliably, with guaranteed conservative fixed width confidence intervals, provided an upper bound for the kurtosis of the sampled r.v. is known prior to sampling.

5.1 Experimental setup and implementational details

The distributions of the r.v. we will consider will be the light-tailed Uniform, the Exponential, the heavy-tailed Normal-inverse Gaussian and the heavy-tailed Pareto distribution. The initial number of samples used in Algorithm 1 and 2 is set to $M_0 = 32$ for all experiments. The version of Algorithm 2 we will implement for these experiments uses the only sample moments defined in (12), i.e., $C_P = C_P(2; M)$. For the implementation of Hickernell et al.'s algorithm an upper bound on the kurtosis is required, so we will assume the exact kurtosis of the r.v. is known prior to sampling and set the "variance inflation factor" $\mathfrak{C} = 1.1$, cf. [6]. Let us stress that the assumption of knowing the exact value of kurtosis prior to sampling is only made in the runs for Hickernell et al.'s algorithm, and due to this, its performance is presented in a slightly more favorable light than Algorithm 1 and 2's.

For each of the three algorithms we will numerically estimate the probability of failure $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right)$ by the MC outer loop

$$\overline{p}_N(\mathrm{TOL}, \delta) := N^{-1} \sum_{i=1}^{N} \mathbf{1}_{\left|\overline{X}_M(\omega_i) - \mu\right| > \mathrm{TOL}} \tag{15}$$

on a 100×100 grid of $(\mathrm{TOL}_i, \delta_j) \in [10^{-2}, 10^{-1}] \times [10^{-3}, 10^{-1}]$ values where $\mathrm{TOL}_i = 10^{-(1+i/99)}$ and $\delta_i = 10^{-(1+2i/99)}$ for $i = 0, 1, \ldots, 99$. The number of outer loop samples N we are able to use in an experiment will depend on both the cost of sampling the given r.v. and the cost of the used stopping algorithm. The error in estimating the probability of failure by $\overline{p}_N(\mathrm{TOL}, \delta)$ is estimated by the standard deviation approximation of a sampled Bernoulli r.v.

$$\frac{\sqrt{\overline{p}_N(\mathrm{TOL}, \delta)(1 - \overline{p}_N(\mathrm{TOL}, \delta))}}{\sqrt{N}}.$$

At $(\mathrm{TOL}_i, \delta_j)$ points where for a given algorithm $\overline{p}_N(\mathrm{TOL}_i, \delta_j)/\delta_j > 1$, that algorithm is considered unreliable. So to visualize the domain of reliability for an algorithm we plot the function $\overline{p}_N/\delta$. Furthermore, Table 1 provides the following values

$$\max_{i,j=0,1,\ldots,99} \frac{\overline{p}_N(\mathrm{TOL}_i, \delta_j)}{\delta_j}, \quad \text{and} \quad \max_{i,j=0,1,\ldots,99} \frac{\sqrt{\overline{p}_N(\mathrm{TOL}_i, \delta_j)(1 - \overline{p}_N(\mathrm{TOL}_i, \delta_j))}}{\sqrt{N}\delta_j} \tag{16}$$

as estimates of an algorithm's maximum unreliability and the uncertainty in the unreliability estimate, respectively.

The computer code for the algorithms is written in the Java programming language and uses the "Stochastic Simulation in Java" library to sample the r.v. in parallel with the LFSR258 pseudo random number generator, cf. [12]. The experiments were run in parallel using 8 threads on an Intel Xeon(R) CPU X5550, 2.67GHz, and plots are made in Python using the open source plotting library Matplotlib, cf. [8].

5.2 Results

Figures 2, 3, 4, and 5 visualize the results of the numerical experiments and Table 1 contains numerical and parameter values used in the experiments.

For the Pareto distributed r.v. considered in Figure 2, we see that while Algorithm 2 is reliable everywhere, Algorithm 1 is not reliable in the region where TOL is large and $\delta \ll \mathrm{TOL}$. Furthermore, the complexity of Algorithm 2 is only slightly higher than Algorithm 1's. (Hikcernell et al.'s algorithm requires an upper bound on the kurtosis prior to sampling to be applicable, so that algorithm is not applicable in this experiment.)

For the Normal-inverse Gaussian distributed r.v. studied in Figure 3, we see that Algorithm 1 is unreliable in the region where TOL is large. Algorithm 2 and Hickernell et al.'s algorithm are on the other hand both very reliable. Hickernell et al.'s algorithm does however seem to be more reliable than is required, and due to this overkill the complexity of Hickernell et al.'s algorithm is much higher than Algorithm 2's. The Normal-inverse Gaussian distributed r.v. is quite expensive to generate computationally, therefore we have had to reduce the number of outer samples N substantially in this experiment.

For the uniformly and exponentially distributed r.v. studied in Figure 4 and 5, respectively, all three algorithms are reliable and Algorithm 1 and 2 perform very similarly in terms of reliability. We also see that for both of these experiments the complexity measured in terms of the average number of r.v. used to generate $\overline{X}_M$ is very similar for Algorithm 1 and 2. This observation is at odds with the corresponding measured average computer run times given in Table 1 which show Algorithm 2's average run time is considerably longer than Algorithm 1's. The reason for the disagreement between the complexity measurements and the run time measurements is that for uniformly distributed r.v. (and to some weaker degree also for exponentially distributed r.v.) the generation of r.v. is not so costly for the computer program, and therefore do also other arithmetical operations contribute considerably to the run time of the computer program in these experiments.

We further note that the complexity contours of Algorithm 1 and 2 in the figures seem proportional to

$$\widetilde{M}(\mathrm{TOL}, \delta) = \left(\frac{\Phi^{-1}(1 - \delta/2)}{\mathrm{TOL}} \right)^2,$$

up to rounding. This property, which by the CLT is expected in the asymptotic regime, can be verified by comparing the contour of the function $\widetilde{M}(\mathrm{TOL}, \delta)$ to the complexity contour of Algorithm 1 and 2, respectively. The sharp changes in complexity that we observe in the contour plots for Algorithm 1 and 2, especially visible in Figure 4 and 5, are due to the doubling procedure ($M_n = 2M_{n-1}$) which is implemented in these two algorithms.

Considered r.v.	Algorithm	Algorithm performance			
		N	$\max \overline{p}_N/\delta$	$\max(\overline{p}_N(1-\overline{p}_N)/N)^{1/2}/\delta$	runtime$/N$
Pareto, cf. Figure 2 ($\sigma = 1$ and $\kappa = \infty$)	Alg 1	10^6	3.47	0.058804	0.134328s
	Alg 2	10^6	0.688217	0.021043	0.189814s
Normal-inv. Gaussian cf. Figure 3 ($\sigma = 1, \kappa = 123$)	Alg 1	4×10^5	11.897500	0.171435	2.772212s
	Alg 2	4×10^5	0.757546	0.031652	2.768470s
	Hick. et al.'s	2×10^4	0.000602	0.000602	71.960966s
Uniform $U(-\sqrt{3}, \sqrt{3})$, cf. Figure 4 ($\sigma = 1$ and $\kappa = -6/5$)	Alg 1	4×10^6	0.971062	0.015194	0.015740s
	Alg 2	4×10^6	0.971051	0.015192	0.045130s
	Hick. et al.'s	4×10^5	0.274425	0.003862	0.088797s
Exponential $\lambda = 1$, cf. Figure 5 ($\sigma = 1$ and $\kappa = 6$)	Alg 1	2×10^6	0.921406	0.019866	0.041282s
	Alg 2	2×10^6	0.913977	0.018935	0.071302s
	Hick. et al.'s	2×10^5	0.205100	0.003412s	0.212759s

Table 1: We compare the reliability and complexity of three MC algorithms approximating the mean of given r.v. The function $\overline{p}_N(\mathrm{TOL}, \delta)$ is an esitmate of the probability of failure and N is the number samples samples used to compute $\overline{p}_N$, cf. (15). The values $\max \overline{p}_N/\delta$ and $\max(\overline{p}_N(1-\overline{p}_N)/N)^{1/2}/\delta$ estimates of the maximum unreliability its uncertainty, cf. (16). The expression runtime$/N$ is an estimate of the average time it takes for one thread of the implemented computer program of to compute a single output $\overline{X}_M$ at *all* points $(\mathrm{TOL}_i, \delta_j)$, $i, j = 0, 1, \ldots, 99$.

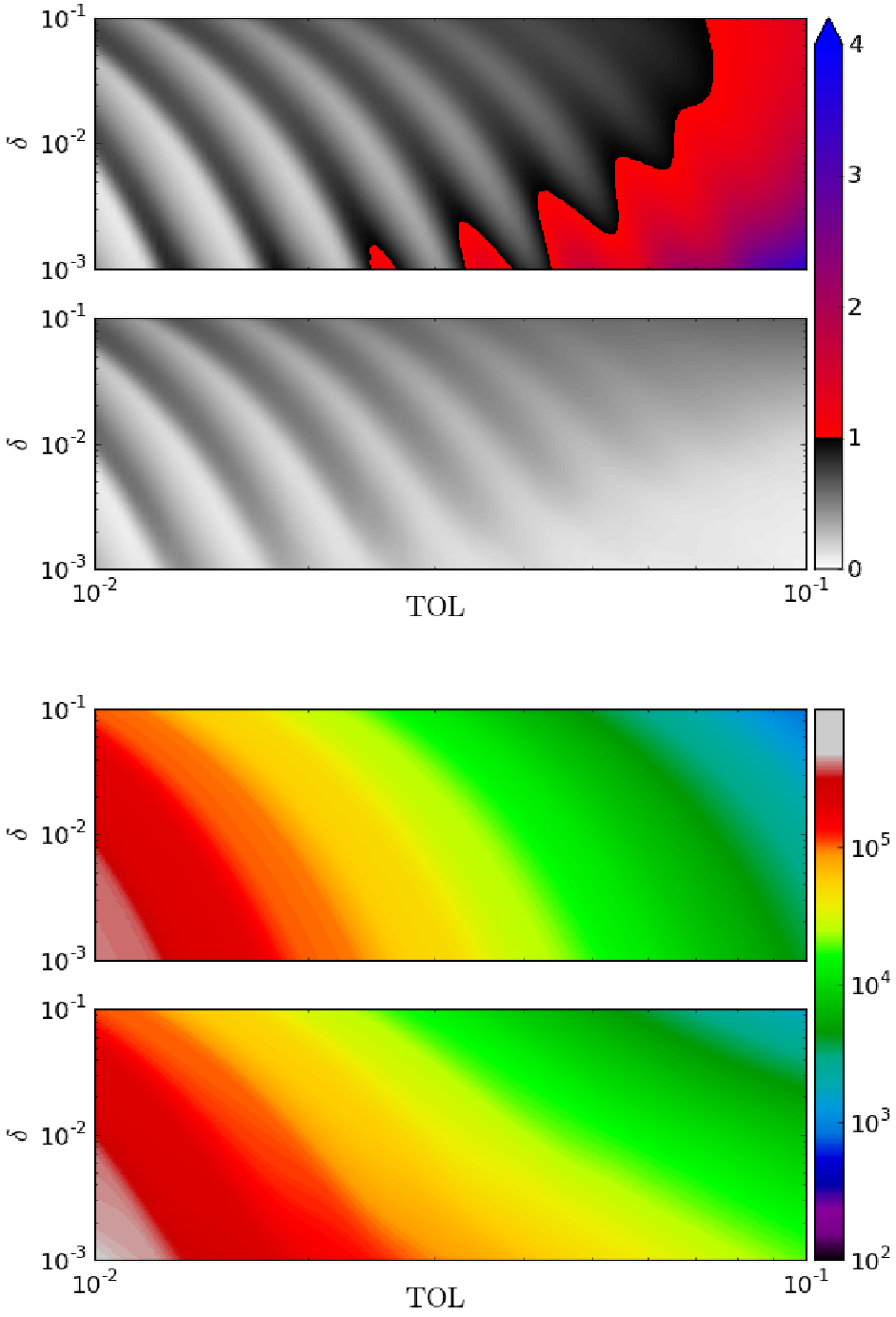

Figure 2: (**Pareto Distribution**) We sample a Pareto distributed r.v. with parameters $\alpha = 3.1$ and $x_m = (\alpha - 1)\sqrt{(\alpha - 2)/\alpha}$. Top two plots: Plots of $\overline{p}_N(\mathrm{TOL}, \delta)/\delta$, cf. (15), where where $\overline{X}_M$ respectively is computed by Algorithm 1 in the top plot, Algorithm 2 in the bottom plot. We consider the algorithm unreliable at (TOL, δ) points where $\overline{p}_N/\delta > 1$. The bottom two plots are corresponding plots of the respective algorithms' complexity in terms of the average number of r.v. samples required to generate $\overline{X}_M$ at the given (TOL, δ).

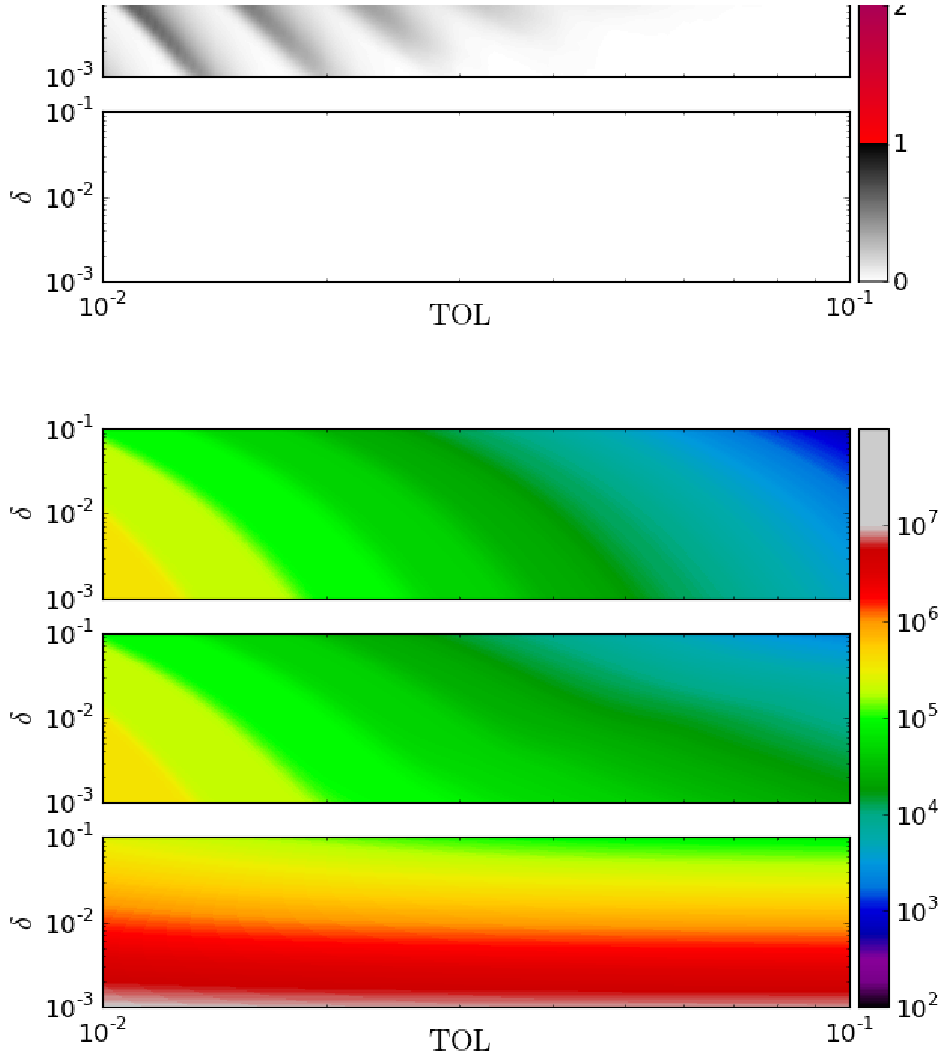

Figure 3: (**Normal-inverse Gaussian Distribution**) We sample a Normal-inverse Gaussian distributed r.v. with parameters $\alpha = 3, \beta = \sqrt{\alpha^2 - 1}$, $\gamma = 1$, $\delta = \alpha^{-2}$, and $mu = -\beta/\gamma$. This yields a r.v. with standard deviation $\sigma = 1$ and $\kappa = 223$. Top three plots: Plots of $\overline{p}_N(\mathrm{TOL}, \delta)/\delta$, cf. (15), where samples of $\overline{X}_M$ respectively are computed by Algorithm 1 in the top plot, Algorithm 2 in the middle plot, and Hickernell et al.'s algorithm in the bottom plot. We consider the algorithm unreliable at (TOL, δ) points where $\overline{p}_N/\delta > 1$. The bottom three plots are corresponding plots of the respective algorithms' complexity in terms of the average number of r.v. samples required to generate $\overline{X}_M$ at the given (TOL, δ) point.