Iterative Importance Sampling for Estimating Expectation Bounds Under Partial Probability Specifications

Matthias C. M. Troffaes,[1*] Thomas Fetz,[2] and Michael Oberguggenberger[2]

[1] *Department of Mathematical Sciences, Durham University, UK*
[2] *Arbeitsbereich für Technische Mathematik, Universität Innsbruck, Austria*
[*] Corresponding author: matthias.troffaes@durham.ac.uk

Abstract

In this paper, we explore and enhance importance sampling techniques for calculating lower and upper expectations with respect to sets of probability distributions. We formalize an iterative algorithm that we proposed in earlier work, by formulating the algorithm as a procedure for identifying a fixed point. We show how the algorithm can break down under poor coverage of the sampling distribution, and explore simple methods to increase coverage and thereby improve the algorithm.

Keywords: importance sampling, reweighting, probability bounding, imprecise probability, lower prevision, Monte Carlo

1 Introduction

In many engineering problems, it can be hard to specify full probability densities for all parameters, due to lack of data and expert information. In such cases, we may prefer to work with partial probability specifications, or equivalently, sets of probability densities [1–3]. Typically, we may wish to estimate the lower expectation (lower prevision) $\theta_* \coloneqq \min_{t\in\mathcal{T}} \int h(x) f_t(x)\,\mathrm{d}x$ of some function h with respect to some parametrized family of probability density functions f_t, over all $t \in \mathcal{T}$. For example, in reliability engineering, h might be an indicator function of a failure region described by a limit state function g and then θ_* is the lower probability of failure. Upper probabilities of failure can be treated in the same way.

In earlier work [4–8] we studied how to estimate θ_* through importance sampling. In this paper, we formalize our work further by describing the sample as a parametrized function $x_t(V)$ of a fixed random variable V, so that $x_t(V)$ has the distribution f_t. This enables better control of the error, reduces the bias as shown in [8], and formalizes the technique of 'fixing the seed' across iterations of the optimisation steps [7]. Here, we study the convergence of the iterative importance sampling estimator developed in [6, 7] by formulating it as a fixed point of an operator. We contrast the iterative procedure with standard sampling, and we investigate how increased coverage of the sampling region can substantially improve the accuracy of the estimates. Examples demonstrate our approach.

2 Importance Sampling

Let f_t be a density parameterized by $t \in \mathcal{T}$. We are interested in estimating the lower and/or upper expectation of some function h with respect to f_t over all $t \in \mathcal{T}$:

$$\theta(t) \coloneqq \int h(x) f_t(x)\,\mathrm{d}x, \tag{1}$$

$$\theta_* \coloneqq \min_{t\in\mathcal{T}} \theta(t), \qquad \theta^* \coloneqq \max_{t\in\mathcal{T}} \theta(t). \tag{2}$$

We assume that samples from f_t can be generated as follows. We start from a random variable V (e.g. uniform in $[0,1]^k$), and a function x_t of V, such that

$$x_t(V) \sim f_t. \tag{3}$$

For example, if $t = (\mu, \sigma)$ and f_t is $N(\mu, \sigma^2)$, then $V = (U_1, U_2)$ could be a standard bivariate uniform on $[0,1]^2$, and

$$x_t(V) = \mu + \sigma\sqrt{-2\ln U_1}\cos(2\pi U_2) \tag{4}$$

will have the desired distribution [9]. Similar transformation methods are available for all standard distributions, and, in general, can be obtained for instance via inverse transform sampling. Specifically, if

F_t is the cumulative distribution function for the density f_t, and V is a standard uniform random variable on $[0,1]$, then

$$x_t(V) = F_t^{-1}(V) \sim f_t. \tag{5}$$

The reason for making the function x_t explicit is that we need to control the randomness throughout the algorithm that we will describe next. More specifically, we will need to describe the sample itself as a deterministic function of the parameter t.

Imagine that we can start from an i.i.d. sample $\Omega := (V_1, V_2, \ldots, V_n)$, to obtain an i.i.d. sample

$$x_s(V_1), \ldots, x_s(V_n) \tag{6}$$

from f_s, for some fixed $s \in \mathcal{T}$. Now, because

$$\int h(x) f_t(x)\, dx = \int \frac{f_t(x)}{f_s(x)} h(x) f_s(x)\, \mathrm{d}x, \tag{7}$$

we can use this sample from f_s to estimate the expectation of h with respect to f_t, for any $t \in \mathcal{T}$:

$$\hat{\theta}_{\Omega,s}(t) := \frac{1}{n} \sum_{i=1}^{n} w_{st}(x_s(V_i)) h(x_s(V_i)) \tag{8}$$

where

$$w_{st}(x) := \frac{f_t(x)}{f_s(x)}. \tag{9}$$

Note that $\hat{\theta}_{\Omega,s}(t)$ is an unbiased estimator for $\theta(t)$:

$$E(\hat{\theta}_{\Omega,s}(t)) = \theta(t). \tag{10}$$

If this normalisation constant of the probability density f_t is expensive to calculate, then there is substantial gain by only evaluating it once for each desired t (or s), and to reuse it across all terms in the sum. Alternatively, self-normalised importance sampling can be used:

$$\hat{\theta}'_{\Omega,s}(t) := \frac{\sum_{i=1}^{n} w'_{st}(x_s(V_i)) h(x_s(V_i))}{\sum_{i=1}^{n} w'_{st}(x_s(V_i))} \tag{11}$$

where $w'_{st}(x)$ are defined in the same way as the weights $w_{st}(x)$ but only up to a normalisation constant of the densities involved. This has the downside that the resulting estimate is only asymptotically unbiased [8].

A special case obtains when $s = t$. In that case, we have standard sampling:

$$\hat{\theta}_{\Omega}(t) := \hat{\theta}_{\Omega,t}(t) := \frac{1}{n} \sum_{i=1}^{n} h(x_t(V_i)) \tag{12}$$

because $w_{tt}(x) = 1$ for all x. This leads to the following estimators for θ_* and θ^* [8]:

$$\hat{\theta}_{*\Omega} := \hat{\theta}_{\Omega}(T_{*\Omega}) = \min_{t \in \mathcal{T}} \hat{\theta}_{\Omega}(t) \quad \text{and} \quad \hat{\theta}^*_{\Omega} := \hat{\theta}_{\Omega}(T^*_{\Omega}) = \max_{t \in \mathcal{T}} \hat{\theta}_{\Omega}(t), \tag{13}$$

where

$$T_{*\Omega} := \arg\min_{t \in \mathcal{T}} \hat{\theta}_{\Omega}(t) \quad \text{and} \quad T^*_{\Omega} := \arg\max_{t \in \mathcal{T}} \hat{\theta}_{\Omega}(t). \tag{14}$$

The estimates $\hat{\theta}_{\Omega}(t)$ will be highly correlated for different values of t, which helps reducing the bias, as shown in [8].

A difficulty with calculating $T_{*\Omega}$ or T^*_{Ω} is that we need to evaluate h at points $x_t(V_i)$, and these points will arbitrarily shift around as we optimize over t. With importance sampling, however, for fixed s, we only need to evaluate h for the points $x_s(V_i)$, independently of t. So if h is expensive to evaluate, then importance sampling is particularly useful, because we do not need to re-evaluate h for different t when optimizing over t. In addition, we retain the benefit that the estimates $\hat{\theta}_{\Omega,s}(t)$ will be highly correlated for different values of t, helping to reduce the bias [8].

With importance sampling, for each s, we have the following estimators for θ_* and θ^* [8]:

$$\hat{\theta}_{*\Omega}(s) := \hat{\theta}_{\Omega,s}(\tau_{*\Omega}(s)) = \min_{t \in \mathcal{T}} \hat{\theta}_{\Omega,s}(t) \quad \text{and} \quad \hat{\theta}^*_{\Omega}(s) := \hat{\theta}_{\Omega,s}(\tau^*_{\Omega}(s)) = \max_{t \in \mathcal{T}} \hat{\theta}_{\Omega,s}(t), \tag{15}$$

where

$$\tau_{*\Omega}(s) := \arg\min_{t \in \mathcal{T}} \hat{\theta}_{\Omega,s}(t) \quad \text{and} \quad \tau^*_{\Omega}(s) := \arg\max_{t \in \mathcal{T}} \hat{\theta}_{\Omega,s}(t). \tag{16}$$

The quality of the importance sampling estimates can be verified in the standard way via confidence intervals that are constructed through repeated sampling [8].

3 Iterative Importance Sampling

An issue with the importance sampling estimates is that their quality can be very poor if $\tau_{*\Omega}(s)$ is far from s. A procedure for iteratively improving the choice of s was proposed in [6, 7]. The procedure essentially iteratively applies the operator $\tau_{*\Omega}$. Under the assumption that this iterative application

$$s^{(k+1)} = \tau_{*\Omega}(s^{(k)}), \quad k = 1, 2, \ldots \tag{17}$$

of the operator $\tau_{*\Omega}$ reaches a unique fixed point, say $S_{*\Omega}$, our improved lower estimator is:

$$\hat{\theta}^{\dagger}_{*\Omega} \coloneqq \hat{\theta}_{\Omega, S_{*\Omega}}(S_{*\Omega}) = \frac{1}{n}\sum_{i=1}^{n} h(x_{S_{*\Omega}}(V_i)). \tag{18}$$

In numerical examples discussed in [6–8], normally, a fixed point is indeed obtained after few steps. As we shall see, however, $\tau_{*\Omega}$ is not necessarily continuous, and therefore it is not guaranteed that a fixed point exists. Even if it is continuous, $\tau_{*\Omega}$ is not necessarily contracting, and therefore it is not guaranteed that a fixed point can be found by repeated application of $\tau_{*\Omega}$ on itself.

However, asymptotically, as the sample size n increases, we are tempted to conjecture that, under suitable conditions, $\tau_{*\Omega}$ should have a fixed point $S_{*\Omega}$, and both $T_{*\Omega}$ and $S_{*\Omega}$ should converge, in probability, to

$$t_* \coloneqq \arg\min_t \theta(t). \tag{19}$$

The intuition behind this conjecture is that $\hat{\theta}_{\Omega,s}$ converges in probability to θ as the sample size goes to infinity. In similar way we get the upper estimator $\hat{\theta}^{*\dagger}_{\Omega}$.

Example 1 *Here we consider the estimation of the upper probability of the event*

$$D = [-4, -3] \cup [-1, 1] \cup [3.1, 4.7], \tag{20}$$

with respect to the set of normal distributions with mean $t \in \mathcal{T} = [-7, 7]$ and variance $\sigma^2 = 2$. The probability of D, for each value of t, is depicted in fig. 1. The values of t where local maxima are achieved are indicated by vertical dotted lines. Note that the upper probability of D is simply the upper expectation of

$$h(x) \coloneqq \begin{cases} 1 & \text{if } x \in D, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

The probability $\theta(t) = \int_D f_t(x)\,\mathrm{d}x = \int_{\mathbb{R}} h(x) f(x)\,\mathrm{d}x$, is depicted in fig. 1 (left) for each value $t \in \mathcal{T}$. The exact maximum (upper probability) θ^* is equal to 0.5488524 and is achieved for $t = -0.0011136$.

Since the function h is not continuous, the approximation $\hat{\theta}_\Omega$, eq. (12), of θ is not continuous either. It is a step function with step sizes $1/n$. As we can see in fig. 1 (left) $\hat{\theta}_\Omega$ is a quite good approximation ($n = 1000$), but it is expensive to compute.

In fig. 2, we show the contour plots of $\hat{\theta}_{\Omega,s}(t)$ as a function of s and t, for three different sample sizes n. We also overlay the function $\tau^*_\Omega(s) = \arg\max_{t\in\mathcal{T}} \hat{\theta}_{\Omega,s}(t)$ as a function of s. Finally, we also show the result of iteratively applying τ^*_Ω, starting from $s^{(1)} = 6$. For smaller sample sizes, in this case, we have cycling. For larger sample sizes, we no longer have cycling. We can see in the intermediate case that we only pick up the local maximum through the iterative procedure, although it is the global maximum of $\hat{\theta}_{\Omega,S^*_\Omega}$ where $S^*_\Omega = 3.6494$. Table 1 provides the full numerical results of each iteration.

4 Increased Sampling Coverage

Unfortunately, in some already quite simple cases, the sample size required for $\hat{\theta}_{\Omega,s}$ to converge to θ can be excessively large. In particular, $\hat{\theta}_{\Omega,s}$ may not reflect at all the shape of θ especially when s is far from t. The cause of this behaviour is that sampling from f_s may not cover regions where f_t is located. We can address this by modifying the sampling distribution f_s in order to increase this coverage. There are various ways of doing this: we can use a convex mixture of the original distribution and an additional distribution with large variance, or if possible we may also simply inflate the variance of the distribution directly.

For example, imagine that we wish to ensure that our sampling distribution covers the entire interval $[a, b]$ on the real line. For this purpose, we modify our sampling distribution to:

$$f^R_s(x) \coloneqq (1 - \alpha) f_s(x) + \frac{\alpha}{b - a} I_{x\in[a,b]} \tag{22}$$

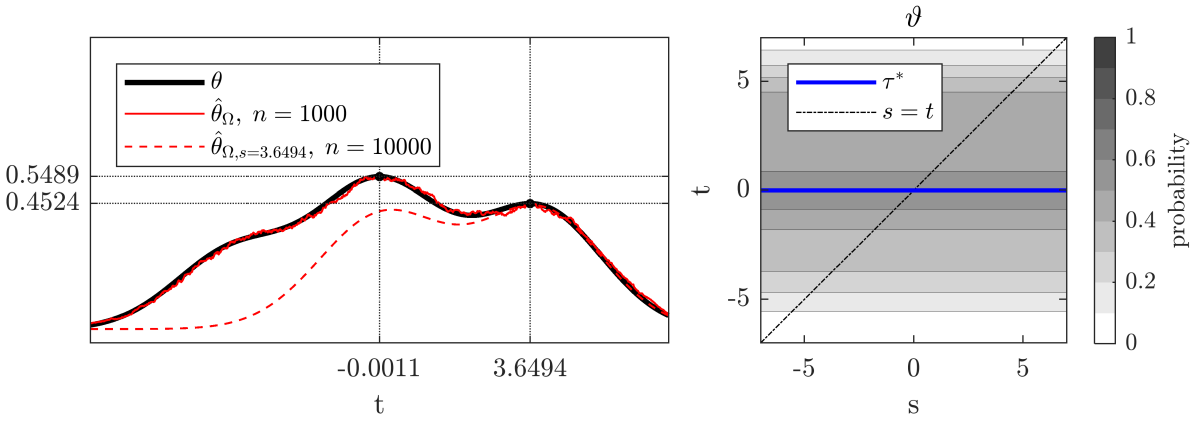

Figure 1: Exact probability θ, standard Monte Carlo estimation $\hat{\theta}_\Omega$, and estimation $\hat{\theta}_{\Omega,s}$ for a fixed value of $s = 3.6494$ where θ has a local maximum, as a function of $t \in \mathcal{T} = [-7, 7]$ (left). Contour plot of exact function $\vartheta(s,t) := \theta(t)$ and its maximum $\tau^*(s)$ (right), for comparison with the contour plots of the estimators in fig. 2.

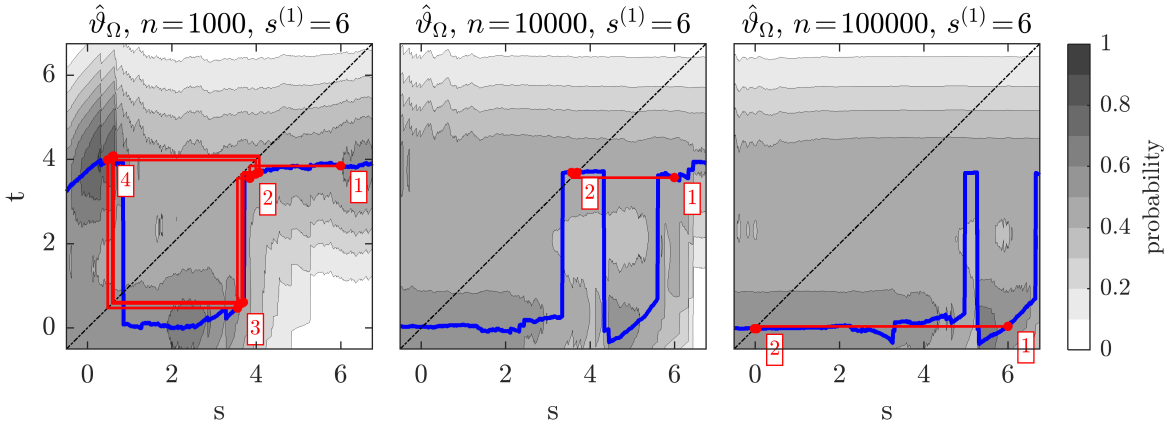

Figure 2: Contour plots of $\hat{\vartheta}_\Omega(s,t) := \hat{\theta}_{\Omega,s}(t)$ and depiction of $\tau^*_\Omega(s) = \arg\max_{t\in\mathcal{T}} \hat{\theta}_{\Omega,s}(t)$ (blue line) for three different sample sizes $n = 1000$, 10000, and 100000. The path of the iteration with starting value $s^{(1)} = 6$ is plotted as a red line. We have cycling (left), convergence to a local maximum (middle), and convergence to the global maximum (right).

$N = 1000$, $s^{(1)} = 6$			
k	$s^{(k)}$	$\tau^*_\Omega(s^{(k)})$	$\hat{\theta}_{\Omega,s^{(k)}}(\tau^*_\Omega(s^{(k)}))$
1	6.0000	3.8500	0.4954
2	3.8500	3.5560	0.4432
3	3.5560	0.4760	0.5098
4	0.4760	3.9900	0.7443
5	3.9900	3.6540	0.4321
6	3.6540	0.5880	0.4958
7	0.5880	4.0740	0.7670
8	4.0740	3.6960	0.4388
9	3.6960	0.6160	0.4749
10	0.6160	4.0880	0.7775

$N = 10000$, $s^{(1)} = 6$			
k	$s^{(k)}$	$\tau^*_\Omega(s^{(k)})$	$\hat{\theta}_{\Omega,s^{(k)}}(\tau^*_\Omega(s^{(k)}))$
1	6.0000	3.5700	0.4581
2	3.5700	3.6960	0.4488
3	3.6960	3.6960	0.4467
4	3.6960	3.6960	0.4467

$N = 100000$, $s^{(1)} = 6$			
k	$s^{(k)}$	$\tau^*_\Omega(s^{(k)})$	$\hat{\theta}_{\Omega,s^{(k)}}(\tau^*_\Omega(s^{(k)}))$
1	6.0000	0.0420	0.5527
2	0.0420	-0.0140	0.5496
3	-0.0140	-0.0140	0.5497
4	-0.0140	-0.0140	0.5497

Table 1: Iteration steps.

where $I_{x\in[a,b]} = 1$ if $x \in [a,b]$ and 0 otherwise. To generate a sample from this distribution, we let

$$x_s^R(U,V) \coloneqq \begin{cases} a + (b-a)\frac{U}{\alpha} & \text{if } U < \alpha \\ x_s(V) & \text{if } U \geq \alpha \end{cases} \tag{23}$$

where U is a uniform $[0,1]$ variable, and V is as before.

Now, we start from an i.i.d. sample $\Omega \coloneqq (U_1, V_1, \ldots, U_n, V_n)$ to obtain an i.i.d. sample

$$x_s^R(U_1,V_1), \ldots, x_s^R(U_n,V_n) \tag{24}$$

from f_s^R, for some fixed $s \in \mathcal{T}$. As before, we can use this sample from f_s to estimate the expectation of h with respect to f_t, for any $t \in \mathcal{T}$:

$$\hat{\theta}_{\Omega,s}(t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} w_{st}(x_s^R(U_i,V_i))h(x_s^R(U_i,V_i)) \tag{25}$$

where

$$w_{st}(x) \coloneqq \frac{f_t(x)}{f_s^R(x)}. \tag{26}$$

Here too, $\hat{\theta}_{\Omega,s}(t)$ is an unbiased estimator for $\theta(t)$:

$$E(\hat{\theta}_{\Omega,s}(t)) = \theta(t). \tag{27}$$

Obviously, we should choose a and b in a way that we are sure to generate samples that cover the range of f_t across all $t \in \mathcal{T}$. For α, if we choose it too small, then we may not generate sufficient samples across the desired interval $[a,b]$. If we choose it too large, then f_s^R may be too far away from f_t regardless of our choice of s, thereby removing the opportunity to increase the accuracy of the estimate through the iterative procedure, as the estimator will no longer depend on s. A suggestion is to choose α in a way that for every $t \in \mathcal{T}$, there are at least a handful of samples that cover f_t.

For more general cases, a convex mixture of target distributions f_t could also be used. For example, if $\mathcal{T} = [0,1]$, then we could take

$$f_s^R(x) \coloneqq (1-\alpha)f_s(x) + \alpha\int_0^1 f_t(x)\,\mathrm{d}t \tag{28}$$

and

$$x_s^R(U,V) \coloneqq \begin{cases} x_{U/\alpha}(V) & \text{if } U < \alpha \\ x_s(V) & \text{if } U \geq \alpha \end{cases} \tag{29}$$

where U is a uniform $[0,1]$ variable, and V is as before. If need be, the integral can be approximated by a finite sum, and the sampling function can be adjusted accordingly. Theoretical arguments and further proposals for convex mixtures have been given in [10, §4.3].

The next example shows how we can improve coverage using variance inflation of the sampling distribution.

Example 2 *We consider the estimation of the lower probability of the event $D = (-\infty,-1]\cup[1,\infty)$ with respect to the set of normal distributions with mean $t \in [-7,7]$ and variance $\sigma^2 = 2$.*

The probability $\theta(t)$ is depicted in fig. 3 (left) for each value $t \in \mathcal{T}$. The minimum (lower probability) θ_* is equal to 0.4795 and is achieved for $t_* = 0$.

First, we consider $f_s^R = f_s$, that is, no increased coverage from the sampling distribution. The cheaper function $\hat{\theta}_{\Omega,s=0}$ using reweighting based on density $f_{s=0}^R$ provides a good approximation near the solution $t_* = 0$ but not far away from t_*. In this example, $\hat{\theta}_{\Omega,s=0}$ leads to a completely wrong global minimum 0.0282 achived at $t = -7$, as seen from the dashed curve in fig. 3 (left). The reason why this approximation is bad for t far away from $t_* = 0$ is that there are no or almost no sample points of the reweighting density $f_{s=0}^R$ in areas in D with high density of f_t for t going to ± 7.

Figures 5 and 6 show what happens if instead we take f_t^R to be normally distributed with mean t but with increased variance $\sigma^2 = 10$. We observe that $\hat{\theta}_{\Omega,s}$ provides a reasonably good approximation for θ across a much wider range of values for s and t. For the lowest sample size, 1000, we still observe cycling in the iterative method, as the variance is not sufficiently inflated to cover the more extreme ends of the range of the distributions that we are interested in. For larger sample sizes, the coverage becomes sufficient, and the iterative procedure produces a correct value.

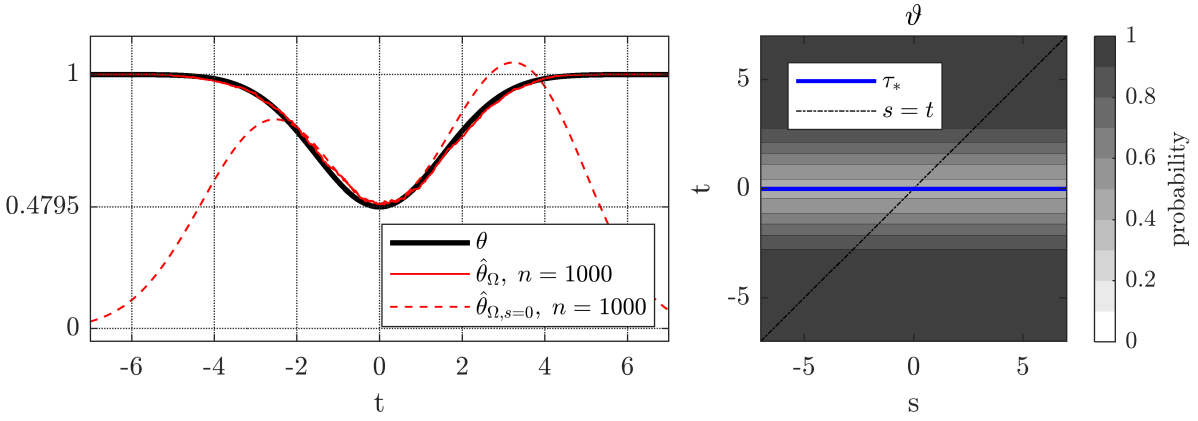

Figure 3: Exact probability θ, standard Monte Carlo estimation $\hat{\theta}_\Omega$ and estimation $\hat{\theta}_{\Omega,s}$ for a fixed value of s, as a function of $t \in \mathcal{T} = [-7, 7]$ (left). Contour plot of exact function $\vartheta(s,t) := \theta(t)$ and its minimum $\tau_*(s) := t^*$ (right), for comparison with the contour plots of the estimators in figs. 4 and 6.

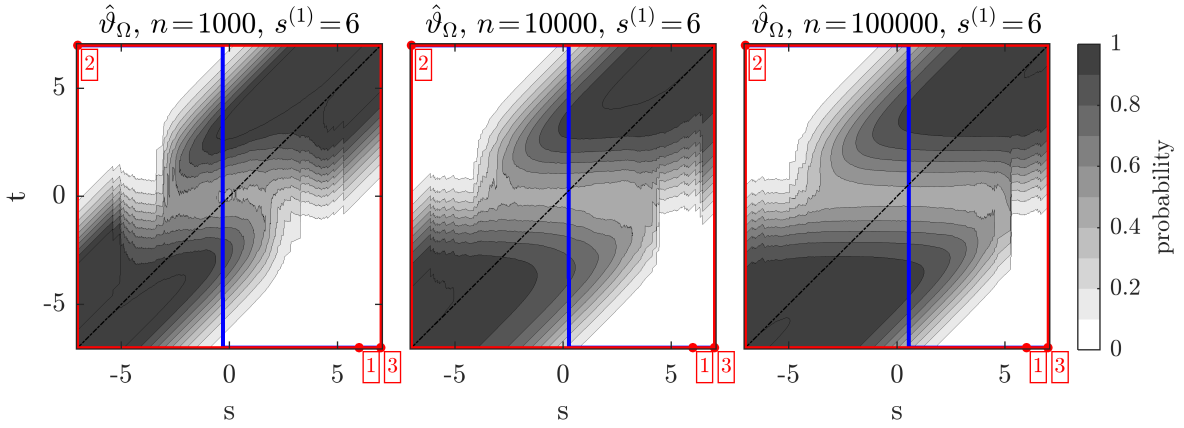

Figure 4: Contour plots of $\hat{\vartheta}_\Omega(s,t) := \hat{\theta}_{\Omega,s}(t)$ and depiction of $\tau_{*\Omega}(s) = \arg\min_{t\in\mathcal{T}} \hat{\theta}_{\Omega,s}(t)$ (blue line) for three different sample sizes $n = 1000$, 10000, and 100000. The path of the iteration with starting value $s^{(1)} = 6$ is plotted as a red line. We have cycling in all cases, due to lack of coverage of the sampling distribution in regions where t is far from s.

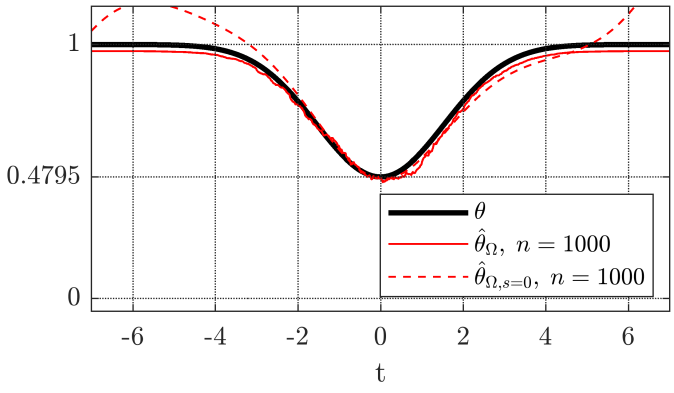

Figure 5: Exact probability θ, standard Monte Carlo estimation $\hat{\theta}_\Omega$ and estimation $\hat{\theta}_{\Omega,s}$ for a fixed value of $s = 0$, as a function of $t \in \mathcal{T} = [-7, 7]$.

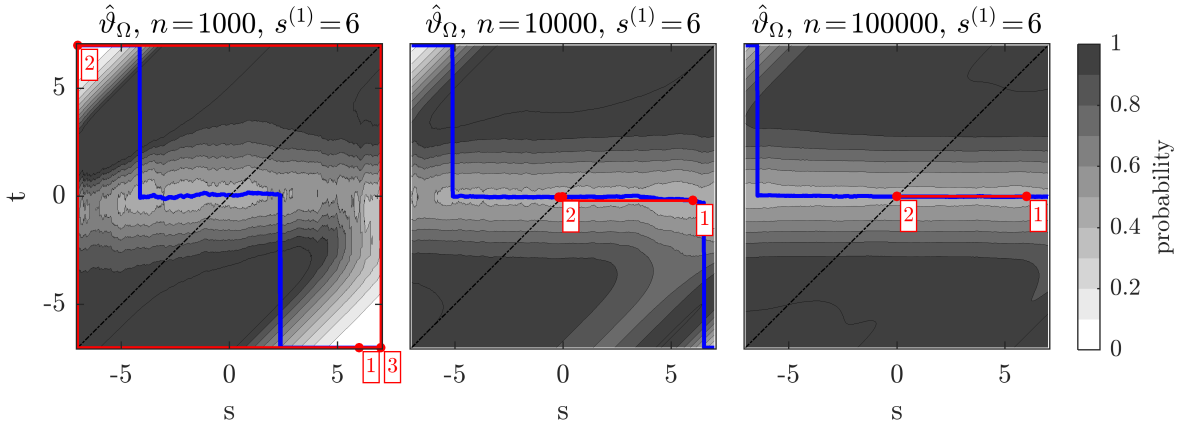

Figure 6: Contour plots of $\hat{\vartheta}(s,t) := \hat{\theta}_{\Omega,s}(t)$ and depiction of $\tau_{*\Omega}(s) = \arg\min_{t\in\mathcal{T}} \hat{\theta}_{\Omega,s}(t)$ (blue line) for three different sample sizes $n = 1000$, 10000, and 100000. The path of the iteration with starting value $s^{(1)} = 6$ is plotted as a red line. We still have cycling for sample size 1000, although $\hat{\theta}_{\Omega,s}(t)$ is clearly already a lot closer to $\theta(t)$ for a much wider range of values for s and t. For the larger sample sizes, the iterative procedure converges quickly.

5 Conclusion

We set out to explore importance sampling techniques for calculating lower and upper expectations with respect to sets of probability distributions. We revisited the iterative algorithm proposed in [6, 7], and put it on a better mathematical foundation, by formulating it as a fixed point of an operator. We provided some intuition under which this operator has a fixed point, and thereby provides a good estimator.

We explored some numerical examples, and found that the procedure breaks down when the sampling distribution provides insufficient coverage. We proposed three simple methods to increase coverage. Nevertheless, quantifying the conditions under which importance sampling can provide a sufficiently accurate estimate under a wide range of importance sampling distributions remains an important open question.

Acknowledgments

This work is partially supported by the H2020 Marie Curie ITN, UTOPIAE, Grant Agreement No. 722734.

References

[1] J. O. Berger, in *Robustness of Bayesian Analyses*, edited by J. B. Kadane (Elsevier Science, Amsterdam, 1984), 63–144.

[2] P. Walley, *Statistical Reasoning with Imprecise Probabilities* (Chapman and Hall, London, 1991).

[3] M. C. M. Troffaes and G. de Cooman, *Lower Previsions*, Wiley Series in Probability and Statistics (Wiley, 2014), ISBN 978-0-470-72377-7, http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470723777.html.

[4] T. Fetz and M. Oberguggenberger, in *ISIPTA '15: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, edited by T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur (2015), 137–146, http://www.sipta.org/isipta15/data/paper/12.pdf.

[5] T. Fetz and M. Oberguggenberger, Imprecise random variables, random sets, and monte carlo simulation., *Int. J. of Approximate Reasoning Reasoning* **78**, 252 (2016).

[6] T. Fetz, in *12th International Conference on Structural Safety & Reliability*, edited by C. Bucher, B. R. Ellingwood, and D. M. Frangopol (2017), 493–502.

[7] M. C. M. Troffaes, in *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, edited by A. Antonucci, G. Corani, I. Couso, and S. Destercke (PMLR, 2017), vol. 62 of *Proceedings of Machine Learning Research*, 325–332, http://proceedings.mlr.press/v62/troffaes17a.html.

[8] M. C. M. Troffaes, Imprecise Monte Carlo simulation and iterative importance sampling for the estimation of lower previsions (2018), submitted.

[9] G. E. P. Box and M. E. Muller, A note on the generation of random normal deviates, *The Annals of Mathematical Statistics* **29**, 610 (1958).

[10] J. Zhang and M. D. Shields, On the quantification and efficient propagation of imprecise probabilities resulting from small datasets, *Mechanical Systems and Signal Processing* **98**, 465 (2018), ISSN 0888-3270.