A Note on Imprecise Monte Carlo over Credal Sets via Importance Sampling

Matthias C. M. Troffaes MATTHIAS.TROFFAES@DURHAM.AC.UK *Durham University, UK*

Abstract

This brief paper is an exploratory investigation of how we can apply sensitivity analysis over importance sampling weights in order to obtain sampling estimates of lower previsions described by a parametric family of distributions. We demonstrate our results on the imprecise Dirichlet model, where we can compare with the analytically exact solution. We discuss the computational limitations of the approach, and propose a simple iterative importance sampling method in order to overcome these limitations. We find that the proposed method works pretty well, at least in the example studied, and we discuss some further possible extensions.

Keywords: importance sampling; lower prevision; Monte Carlo; optimisation.

1. Introduction

Various sensible approaches to sampling for lower previsions can be found in the literature. Some of these are:

- two-level Monte Carlo sampling, where first one samples distributions over the (extreme points of the) credal set, and then samples from these distributions,
- sampling random sets, and then evaluating the resulting belief function (Moral and Wilson, 1996), and
- perform importance sampling from a reference distribution, and then solve an optimisation problem over the importance sampling weights (O'Neill, 2009; Fetz and Oberguggenberger, 2015; Zhang and Shields, 2016).

The first is inefficient, and only provides a non-conservative solution. The second is more efficient, but requires a large number of optimisation problems to be solved (one for each sample), and requires a suitable belief function approximation to be identified if one wants to apply this to arbitrary lower previsions. The third can be quite effective. For example, de Angelis et al. (2015) have successfully used sensitivity analysis over importance sampling weights with respect to the mean parameter of a normal distribution. Fetz and Oberguggenberger (2015) used importance sampling over both the mean and the variance parameters of a normal distribution using a 2-dimensional grid. A case study comparing a wide range of techniques, specifically aimed at reliability analysis, can be found in Oberguggenberger et al. (2009). Here, we are interested in seeing whether importance sampling can be performed over larger parameter spaces and distributions with non-trivial normalisation constants, using standard high-dimensional optimisation procedures.

Importance sampling in imprecise probability has been studied already in the '90s; see for example Moral and Wilson (1996); Cano et al. (1996); Hernández and Moral (1997) for some early works. In this paper, we follow O'Neill (2009), and look specifically at how we can use sensitivity analysis over the importance sampling weights directly in order to obtain sampling estimates, without needing to draw large numbers of samples, and without needing to solve large numbers of optimisation problems. Unlike O'Neill (2009), however, we do not just look at Bayesian sensitivity analysis, and admit arbitrary sets of distributions in our theoretical treatment. Also unlike for instance O'Neill (2009); de Angelis et al. (2015); Fetz and Oberguggenberger (2015); Zhang and Shields (2016), in this paper, we use self-normalised importance sampling instead of standard importance sampling, as we find that this drastically speeds up calculations.

The main contribution of this paper is a simple yet novel (as far as we know) iterative importance sampling method that requires far less computational power compared to standard importance sampling methods for sensitivity analysis, in the sense that far smaller samples can be used, and that far smaller optimisation problems need to be solved. The key novelty is the idea of iteratively changing the importance sampling distribution itself, in order to ensure that the final answer has an effective sample size that is as close as possible to the actual sample size.

No novel theory is proved in the paper, however we do demonstrate the method on a fully worked example. This leads us to conjecture that convergence of the technique can be established under certain circumstances.

Section 2 reviews the basic theory behind importance sampling. Section 3 looks at how sensitivity analysis can be applied on importance sampling. An example of this approach is discussed in section 4, and various issues are identified. Section 5 describes a simple way of addressing some of these issues. The example is revisited in section 6. Section 7 concludes the paper with a discussion and some further ideas for future research.

2. Importance Sampling

In this section, we review the basic ideas behind importance sampling. For the theory behind the results that are presented here, we refer to Owen (2013, Chapter 9).

Assume we have an i.i.d. sample $x_1, \dots, x_n$ drawn from a strictly positive probability density function q. Throughout the entire paper, we will consider many different probability density functions, but the sample $x_1, \dots, x_n$ will always be one drawn from q. Assume we have a real-valued function $f(x)$, and we would like to calculate the expectation of f with respect to some other probability density function p.

In case $p = q$, by the central limit theorem, an approximate 95% confidence interval for the expectation of f with respect to q is then given by $\hat{\mu} \pm 1.96\hat{\sigma}/\sqrt{n}$ where

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} f(x_i) \qquad \hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(f(x_i) - \hat{\mu})^2 \tag{1}$$

Can we use the same sample $x_1, \dots, x_n$ drawn from q to get an estimate for the expectation of f with respect to $p \neq q$? The following equality gives a clue as to how we might do that:

$$\int f(x)p(x)dx = \int \frac{p(x)}{q(x)} f(x)q(x)dx = \int w_p(x)f(x)q(x)dx \tag{2}$$

where $w_p = p/q$. So, the expectation of f with respect to p is the same as the expectation of $w_p f$ with respect to q, and therefore an approximate 95% confidence interval for the expectation of f with respect to p is then given by $\hat{\mu}_p \pm 1.96\hat{\sigma}_p/\sqrt{n}$ where

$$\hat{\mu}_p := \frac{1}{n}\sum_{i=1}^{n} w_p(x_i)f(x_i) \qquad \hat{\sigma}_p^2 := \frac{1}{n-1}\sum_{i=1}^{n}(w_p(x_i)f(x_i) - \hat{\mu}_p)^2 \tag{3}$$

This estimate is called the *importance sampling estimate*.

Often, the normalisation constant of the densities is unknown, or is slow to compute, and we only know $w'_p = cp/q$ for some unknown value of c. In this case, we can use the *self-normalised importance sampling estimate*:

$$\hat{\mu}_p := \frac{\sum_{i=1}^n w'_p(x_i) f(x_i)}{\sum_{i=1}^n w'_p(x_i)} \qquad \hat{\sigma}_p^2 := \frac{1}{n-1} \frac{\frac{1}{n}\sum_{i=1}^n w'_p(x_i)^2 (f(x_i) - \hat{\mu}_p)^2}{\left(\frac{1}{n}\sum_{i=1}^n w'_p(x_i)\right)^2} \tag{4}$$

Although $\hat{\sigma}_p^2$ gives an indication of the quality of the estimate, one must be wary that $\hat{\sigma}_p^2$ is by itself only an approximation of the true error. An additional diagnostic to consider is the effective sample size, which can be calculated as follows:

$$n_p := \frac{\left(\sum_{i=1}^n w'_p(x_i)\right)^2}{\sum_{i=1}^n w'_p(x_i)^2} \tag{5}$$

Note that there are many different ways to define effective sample size and even more ways to define diagnostics for importance sampling. What matters for this paper is that a low n_p is bad, and that $n_p \simeq n$ is good. For an in-depth discussion about diagnostics for importance sampling, we refer to Owen (2013, Section 9.3).

3. Sensitivity Analysis

Importance sampling has many different uses, including variance reduction, numerical integration, and Bayesian inference. In this paper, we aim to study importance sampling in order to do inference over sets of distributions.

A key observation is that we can use importance sampling in order to estimate the lower prevision of a gamble f. O'Neill (2009) studied this technique already in a Bayesian setting. Here, we present the theory generally for an arbitrary set of probability density functions.

Say we have some set $\mathcal{M}$ of probability density functions. The *lower prevision* of f is then defined as

$$\underline{E}(f) := \min_{p \in \mathcal{M}} \int f(x) p(x) dx \tag{6}$$

where we assume that the minimum is achieved, for simplicity of presentation. But we know that $\hat{\mu}_p \pm 1.96\hat{\sigma}_p/\sqrt{n}$ provides a confidence interval for the integral on the right hand side. So, if

$$p^* := \arg\min_{p \in \mathcal{M}} \hat{\mu}_p \tag{7}$$

then $\hat{\mu}_{p^*} \pm 1.96\hat{\sigma}_{p^*}/\sqrt{n}$ provides a 95% confidence interval for $\underline{E}$ provided that p^* is equal to, or close enough to, the density that minimises the expectation in eq. (6). The key observation here is that we only need a single sample $x_1, \ldots, x_n$, and that the optimisation procedure operates on the weights only.

One issue with this method is that $\hat{\sigma}_{p^*}$ can be very large. So, the method will only work if $\hat{\sigma}_p$ remains reasonably bounded. From the literature on importance sampling for variance reduction, we know that good choices for q are those that are proportional to $|f|p$ (Owen, 2013, Chapter 9, p. 6). So, in case $\mathcal{M}$ covers a wide range of distributions p, it may be hard to identify a single sampling distribution q. Zhang and Shields (2016, Section 3) discuss ways of chosing optimal sampling distributions for credal sets.

A second problem is that, in general, there is no single sampling distribtution q that can guarantee a good effective sample size for all p in $\mathcal{M}$. Consequently, with this approach, even if we try to chose q optimally, the effective sample size at p^* can still become extremely low.

A third problem is that p^* as determined by eq. (7) may not be close at all to the density that minimises the expectation in eq. (6), especially when the effective sample size is low. In that case, $\hat{\mu}_{p^*} \pm 1.96\hat{\sigma}_{p^*}/\sqrt{n}$ may not provide a very accurate confidence interval on $\underline{E}$. O'Neill (2009, Section 7) derived some explicit statistical bounds on the absolute and relative errors, but these bounds only cover standard (not self-normalising) importance sampling.

4. Example

As a first example, we demonstrate the use of importance sampling for sensitivity analysis on the imprecise Dirichlet model, similar to the one studied in O'Neill (2009).

Denote the k-dimensional unit simplex by Δ. Consider an unknown parameter $x \in \Delta$, say, modelling the probabilities of some multinomial process. Consider the following class of probability density functions on x:

$$p(x \mid t) = \frac{\Gamma(s)}{\prod_{j=1}^{k}\Gamma(st_j)}\prod_{j=1}^{k} x_j^{st_j-1} \tag{8}$$

with hyperparameters $s > 0$ and $t \in \Delta$—these are Dirichlet distributions. We are interested in finding the lower expectation of some function $f(x)$, over all $t \in \mathcal{T} \subseteq \Delta$ and with $s = 2$ fixed.

Note that in our notation, we will parameterise everything in terms of t rather than in terms of p. So $w_t := w_{p(\cdot|t)}$, $\hat{\mu}_t := \hat{\mu}_{p(\cdot|t)}$, $n_t := n_{p(\cdot|t)}$, and so on.

For $q(x)$, we take the Dirchlet distribution with uniform $\tilde{t}_j = 1/k$ and with the same value for $\tilde{s} = 2$. An alternative option is to take $\tilde{s} = \alpha k$ with $0 < \alpha < 1$, say $\alpha = 1/2$. This will incur a bias for sampling towards the extremes, i.e. make the tails heavier. Experimentally, we observed that increasing the variance of the reference distribution can increase the effective sample size.

In order to apply importance sampling, we need to calculate the weight function. The unnormalised weights are:

$$w_t(x) = p(x \mid t)/q(x) \propto \prod_{j=1}^{k} x_j^{st_j-\tilde{s}\tilde{t}_j} = w_t'(x) \tag{9}$$

In this case, we have a very simple closed analytical expression for $w_t'(x)$. Note that we could also use $w_t(x)$ directly, however evaluating the normalisation constants requires several evaluations of the Gamma function, and slows down the optimisation procedure considerably. The optimisation problem for the lower expectation can be written as

$$t^* = \arg\min_{t\in\mathcal{T}} \frac{\sum_{i=1}^{n} w_t'(x_i)f(x_i)}{\sum_{i=1}^{n} w_t'(x_i)} \tag{10}$$

As a numerical example, we take $k = 5$, $\mathcal{T} = \{t \in \Delta \colon t_j \geq 0.1\}$, and $f(x) = x_1 + 2x_2 + 5x_3 + 4x_4 - 3x_5$. In this case, we know that the exact expectation of f, for fixed t, is given by

$$E(f) = t_1 + 2t_2 + 5t_3 + 4t_4 - 3t_5. \tag{11}$$

So, the lower prevision of f over all $t \in \mathcal{T}$ is clearly achieved for $t^* = (0.1, 0.1, 0.1, 0.1, 0.6)$, and is given by

$$\underline{E}(f) = 0.1 + 2 \times 0.1 + 5 \times 0.1 + 4 \times 0.1 - 3 \times 0.6 = -0.6 \tag{12}$$

The next table summarizes our simulation results for $\tilde{s} = k/2 = 2.5$:

n	5	50	500	5000
$\hat{\mu}_{t^*}$	1.50	0.13	-0.85	-0.29
$\hat{\sigma}_{t^*}$	0.11	3.18	10.83	10.74
$\hat{\sigma}_{t^*}/\sqrt{n}$	0.048	0.45	0.48	0.15
n_{t^*}	1.104	15.016	6.061	141.67
t_1^*	0.1	0.1	0.17	0.1
t_2^*	0.57	0.1	0.1	0.1
t_3^*	0.1	0.1	0.1	0.1
t_4^*	0.1	0.1	0.1	0.1
t_5^*	0.13	0.6	0.53	0.6

The code was implemented in R. The `constrOptim` function was used to do the actual optimisation, through the downhill simplex method. The cases $n = 5$ and $n = 50$ give a result instantly, for $n = 500$, the simulation took about 10 seconds, and for $n = 5000$, the simulation took about 200 seconds. The bottleneck is clearly the optimisation procedure. We emphasize that we have not tried to write the fastest possible code, and there might still be good opportunities for optimisation.

Unsurprisingly, the $n = 5$ case is quite bad: t^* is completely off, and the stimate is completely off the chart. Also the error is underestimated substantially, due to the very small effective sample size. The $n = 50$ case fares better. Interestingly, t^* is fully correctly identified. However, the effective sample size is not too high, and the actual estimate is still quite far off, due to the variance once more being underestimated.

Intriguingly, the $n = 500$ case has a lower effective sample size than the $n = 50$ case, and a worse t^*. Nevertheless, the estimate is reasonably correct, and at least the actual value lies inside the 95% confidence interval in this case. The $n = 5000$ case gives the correct estimate for t^*, and again the actual value lies just at the edge of the 95% confidence interval.

5. Iterated Importance Sampling

We have seen that a single importance sampling distribution q may not provide a good effective sample size across the entire set of distributions $\mathcal{M}$, even if n is quite large. For instance, in the numerical example, with $n = 500$ we still only had $n_{t^*} \simeq 6$, and with $n = 5000$ we had only $n_{t^*} \simeq 141$.

What we conclude from this is that plain sensitivity analysis over our importance sampling does not work very well, even in simple cases. Next we discuss some extensions of the proposed procedure in order to make it work.

Even though the estimates are quite bad, our numerical experimentation shows that the correct t^*, or nearly correct t^*, can be identified already with lower n. So, rather than increasing n in order to guarantee a high n_{t^*}, one idea is to iterate the procedure so that $q(x)$ eventually converges to $p(x|t^*)$ where t^* is the actual optimal choice. If $q(x)$ is equal to $p(x|t^*)$, then all weights are identical, and $n = n_{t^*}$. Also, in this case, it turns out that the optimisation in eq. (10) runs very quickly, because we are already near the optimal solution.

Here is how we might implement this in practice:

1. Set t to some reasonable initial value.
2. Generate sample from $q(x) \coloneqq p(x \mid t)$.
3. Find optimal t_* through eq. (10).
4. Check if n_{t^*} is close to n. If yes, stop.
5. Set $t = t_*$, and return to item 2.

One suggestion is to take the same value for n through each step, however a case could be made for chosing a lower value for n, and then simply to repeat the final step of the procedure for a large value of n in order to obtain a final accurate estimate. Another option might be to increase the value of n as the algorithm converges closer to the correct t^*.

6. Example Revisited

Let us apply the proposed iterative procedure on our Dirichlet example. For simplicity, we chose a fixed value of $n = 141$; this corresponds roughly to our earlier n_{t^*} when $n = 5000$, so provides a good basis for comparison of computational efficiency. The next table summarises the results:

iteration	1	2	3
$\hat{\mu}_{t^*}$	0.062	-0.39	-0.63
$\hat{\sigma}_{t^*}$	4.28	2.00	1.76
$\hat{\sigma}_{t^*}/\sqrt{n}$	0.36	0.17	0.15
n_{t^*}	21.60	105.93	141.00
t_1^*	0.16	0.1	0.1
t_2^*	0.1	0.1	0.1
t_3^*	0.1	0.1	0.1
t_4^*	0.1	0.1	0.1
t_5^*	0.54	0.6	0.6

The entire simulation took only 6 seconds, compared to 200 seconds from before for the same effective sample size.

We see that the simulation converges in just 3 steps. In the first step, we get fairly close to the correct t^*, even though the effective sample size $n_{t^*} \simeq 22$ is pretty low. The second step uses this t^* to draw samples, and as this distribution is much closer to the actual optimal distribution, the effective sample size increases substantially. In this step, we also identify the correct value for t^*. The last step uses the correct distribution for sampling, and gets a full effective sample size.

We also ran the simulations using standard (not self-normalised) importance sampling. In that case, the entire simulation took 86 seconds, which is almost a factor 15 slower than the self-normalised version. Undoubtedly this is due to the computational expense of calculating the normalisation constant during the optimisation. Unless the normalisation constant is trivial, self-normalised importance sampling will outperform standard importance sampling for sensitivity analysis over the weights. In addition, the self-normalised estimator has also better consistency properties, even though it has a higher theoretical variance (Owen, 2013, Section 9.2).

7. Discussion and Conclusion

We have described how sensitivity analysis over importance sampling can be used to estimate lower previsions. The key observation that makes this possible is that importance sampling allows us to estimate means not just from the distribution that we are sampling from, but from an entire neighbourhood of distributions around the sampling distribution. Through straightforward optimisation over the importance sampling weights, we can therefore estimate lower previsions without having to, say, draw samples from all extreme points of the credal set. The technique is simple, seems largely unknown in the community, and is readily applicable for medium sized problems.

We saw that a naive application of sensitivity analysis around the weights may not work very well, due to poor effective sample sizes especially when the optimal distribution is far away from the sampling distribution. We suggested simple yet novel solution for this problem: an iterative procedure which naturally moves the sampling distribution towards the optimal distribution. We demonstrated how this led to a much quicker estimate with far less computational power required.

Whilst the procedure that we have described will work well for medium sized problems, we foresee that for really large scale problems, the effective sample size may still be too limited to ensure that the optimal distribution can be identified at all. In such cases, perhaps the credal set could scale throughout the algorithm, in order to ensure a reasonable effective sample size, and therefore to help convergence of the algorithm.

Another idea is to use importance sampling to explore only a very small region of $\mathcal{M}$, but then to use the resulting derivative information to move q in the right direction. A problem with this however is that the derivatives obtained are quite noisy, and in practice we have not found a good way of using these noisy derivatives to ensure convergence.

Obviously, this note only gave an initial exploration of what is possible with sensitivity analysis over the importance sampling weights. It would be interesting to try out these methods on large scale problems. Moreover, it would be great to develop theoretical guarantees and diagnostics for convergence. Finally, it would be interesting to see if the importance sampling as described could be integrated into Markov chain Monte Carlo methods for full robust Bayesian inference over large sets of priors.

Acknowledgements

The author would like to thank Jonathan Rougier, Louis Aslett, Ullrika Sahlin, and Rasmus Bååth for stimulating discussions on the topic of importance sampling for imprecise probability. The author is also grateful to the reviewers for their extremely useful and constructive comments, and in particular for the suggestion of various references relevant to the topic.

References

J. E. Cano, L. D. Hernández, and S. Moral. Importance sampling algorithms for the propagation of probabilities in belief networks. *International Journal of Approximate Reasoning*, 15(1):77–92, 1996.

M. de Angelis, E. Patelli, and M. Beer. Advanced line sampling for efficient robust reliability analysis. *Structural Safety*, 52, Part B:170–182, 2015. ISSN 0167-4730. doi:10.1016/j.strusafe.2014.10.002.

T. Fetz and M. Oberguggenberger. Imprecise random variables, random sets, and Monte Carlo simulation. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *ISIPTA '15: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applica-*

tions, pages 137–146, 2015. URL `http://www.sipta.org/isipta15/data/paper/12.pdf`.

L. D. Hernández and S. Moral. Mixing exact and importance sampling propagation algorithms in dependence graphs. *International Journal of Intelligent Systems*, 12(8):553–576, Aug. 1997.

S. Moral and N. Wilson. Importance sampling algorithms for the calculation of Dempster-Shafer belief. In *Proceedings of IPMU-96 Conference*, volume 3, pages 1337–1344, 1996.

M. Oberguggenberger, J. King, and B. Schmelzer. Classical and imprecise probability methods for sensitivity analysis in engineering: A case study. *International Journal of Approximate Reasoning*, 50(4):680–693, 2009. ISSN 0888-613X. doi:http://dx.doi.org/10.1016/j.ijar.2008.09.004.

B. O'Neill. Importance sampling for Bayesian sensitivity analysis. *International Journal of Approximate Reasoning*, 50(2):270–278, 2009. ISSN 0888-613X. doi:http://dx.doi.org/10.1016/j.ijar.2008.03.015.

A. B. Owen. *Monte Carlo theory, methods and examples*. 2013. URL `http://statweb.stanford.edu/~owen/mc/`.

J. Zhang and M. D. Shields. Efficient propagation of imprecise probabilities. In *7th International Workshop on Reliable Engineering Computing*, pages 197–209, 2016. URL `http://rec2016.rub.de/papers.html`.