BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2010

Using TPA for Bayesian inference — Discussion

Iain Murray
University of Edinburgh, United Kingdom
`i.murray@ed.ac.uk`

[This discussion is intended to follow the Bayesian Statistics 9 paper by Huber and Schott (2010), possibly available at `http://www.uv.es/bernardo/Huber.pdf` *]*

TPA and Nested Sampling

In isolation, Algorithm 2.1 can be viewed as a special case of Nested Sampling. To recover TPA one could run Nested Sampling with the target distribution as its prior and with the likelihood to:

$$L(\theta) = \begin{cases} 1 & \theta \in B' \\ \epsilon/(1+e^{\beta(\theta)}) & \theta \notin B', \quad \text{where } \beta = \inf\{\beta' : \theta \in A(\beta')\}. \end{cases} \tag{1}$$

Skilling (2007) previously identified that the number of steps required to reach a given set is Poisson distributed. Huber and Schott suggest making this special case central, recasting all computations as finding the mass of a distribution on a set. Additional contributions are a theoretical analysis, two general ways of reducing problems to the required form and a link to annealing.

The resulting TPA methods *are* different from a straight application of Nested Sampling. For example, in both variants the initial sampling distribution is set to the posterior of an inference problem rather than the prior.

Theory vs. practice

This work is partly motivated by seeing the errors of Nested Sampling as hard to characterize. The difficulties I've had with Nested Sampling's error bars, which have otherwise been well calibrated, have been due to approximate sampling. I'm unconvinced that TPA offers an improvement.

The second point in Section 10 is incorrect. A Nested Sampling implementation might terminate near a maximum of the likelihood, however the answer is completely dominated by how long the algorithm took to reach the posterior region that contributes to the bulk of Z and the likelihoods there. Nested Sampling must be able to find typical samples from the posterior, but the presented TPA methods *start* by sampling from the posterior.

It is suggested that two forms of errors in Nested Sampling are removed. 1) Nested Sampling terminates when it doesn't appear as though further iterations are going to change the estimator significantly. If a hidden narrow spike actually is significant then wrong answers will result. However, a slice sampler for the posterior would also miss such a spike; TPA based on such samples would fare no better. 2) Although Nested Sampling contains classical numerical integration, upper and lower bounding rectangle rules can give limits on this error (ignoring issues with the final spike). It is easy to verify that errors from quadrature are irrelevant compared to Monte Carlo noise.

TPA's (ϵ, δ) procedure for choosing the number of runs is nice to have in theory. However in brief experiments I have not found it to be very practical. As is often the case with guarantees of this form, setting the mistake rate δ to reasonable values such as 0.05 leads, in practice, to errors much smaller than ϵ far more often than $1{-}\delta$. This means that (δ, ϵ) must be set very loose, or more computer time than really necessary will be used. Of course the (ϵ, δ) guarantee doesn't hold with approximate sampling used in real applications.

Parameter truncation

In the parameter truncation variant of TPA the posterior mass in a small region is estimated. The estimate is compared to the unnormalized probability to recover the implied normalizer Z. Estimating the mass of a special state is reminiscent of the family of methods introduced by Chib (1995), and I am concerned that it could suffer from the same problems (Neal, 1999).

The number of samples that Nested Sampling requires for a given accuracy scales with the square of the log-volume collapse (Murray, 2007). Parameter truncation compresses from the posterior to a small region, which generally has a different log-volume ratio than moving from the prior to the posterior.

Before thinking further about theoretical performance, I *just tried it.* I used a slice-sampling (Neal, 2003) based implementation on the (tiny) galaxy problem considered by Chib and Neal. After trying a few variations for picking the location and size of the final region, I could get answers in the right ball-park, but wasn't able to get reproducible enough answers to demonstrate whether the method was suffering the same problem as Chib's method. The approximate slice sampling that I could do in the time available caused the actual errors to vary by much more than theory would predict. In contrast I was able to get accurate answers with both Nested Sampling and the likelihood truncation version of TPA based on the same slice-sampling code.

Likelihood truncation

Likelihood truncation TPA explains the initially curious use of β throughout the paper. Traditionally β is used as an inverse temperature: to 'cool' or constrain a system one would increase β. However, TPA samples from successively more constrained subsets by *decreasing* its β. In the likelihood truncation variant the temperature analogy makes sense. While truncating the likelihood constrains the auxiliary space $\Omega \times [0, \infty)$, the marginal distribution on the parameter space Ω is usually more diffuse.

Another link to temperatures is given in the proposed method for constructing annealing schedules. Obtaining annealing schedules from the output of Nested Sampling is something I have attempted, in a more convoluted way (Murray, 2007). It will be interesting to see how the more straight-forward procedure presented here compares in practice.

Likelihood truncation TPA moves from sampling the posterior to the prior, whereas Nested Sampling starts by sampling the prior and terminates shortly after finding samples typical under the posterior. Having both methods could be useful: in the context of annealing methods looking for 'hysteresis', differences between cooling and heating curves, can be a useful diagnostic.

Likelihood truncation TPA and Nested Sampling aren't true reverses of each other. In particular, sampling from the likelihood truncated distributions with standard Markov chain methods will not work when there is a first-order phase transition, whereas Nested Sampling can work. As an aside: I have *once* seen a first-order phase transition in a real modeling application, although in that case the problem could be bypassed by re-representing the model.

Independent runs

Algorithm 2.1 specifies that independent runs are made and then combined. Performing runs in parallel is useful when sampling approximately to help set appropriate step-size parameters, which vary dramatically with β. Nested Sampling was explicitly presented with a multiple particle version. I have found that the multiple particle version is much less affected by errors due to using approximate sampling than the single particle version.

Summary

The core TPA algorithm is a simplified version of Nested Sampling with a single particle, for the purposes of theoretical analysis. In practice the presented theory doesn't apply because the required sampling operations are going to be performed approximately.

Huber and Schott have also presented novel methods that result from applying TPA to measuring different aspects of a target distribution. Some nice properties of Nested Sampling, robustness to first-order phase transitions and the multiple particle version, have been discarded. I have found it difficult to get reliable error bars from TPA, especially with the parameter truncation version, when using slice sampling. However, there are several ideas in this paper. My hope is that one or more of them inspire the development of useful tools, perhaps the method for constructing annealing schedules.

REFERENCES

Chib, S. (1995). Marginal likelihood from the Gibbs output *J. Amer. Statist. Assoc.* **90**(432), 1313–1321.

Murray, I. (2007). Advances in Markov chain Monte Carlo methods. PhD Thesis. University College London.

Neal, R. M. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". Available from `http://www.cs.toronto.edu/~radford/chib-letter.html`

Neal, R. M. (2003). Slice sampling. *Ann. Statist.* **31**(3), 705–767 (with discussion).

Skilling, J (2007). Nested Sampling for Bayesian computations. *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press.