Cosmological parameter estimation using Bayesian accelerated machine learning.

Aleksandr Petrosyan

December 5, 2019

Abstract

In the era of precision cosmology, physicists use astronomical data to extract precise measurements of our universe's global properties. In order to do this, we develop of advanced inference and machine learning algorithms. Cosmological parameter estimation is generally performed using Markov-Chain Monte-Carlo algorithms, such as Metropolis-Hastings, Hamiltonian Monte-Carlo or Nested sampling. This project will focus on extending recent work by the astrophysics group on prior re-partitioning to allow nested sampling to take the additional information usually provided to other algorithms into account, for accelerating inference. Such an extension for nested sampling would be highly desirable, and is the basis for the following research project.

1 Introduction

The standard cosmological model for the expansion and origin of the universe is the accepted ΛCDM[4] (Cold Dark Matter). It has six major parameters: *physical baryon density parameter; physical dark matter density parameter; the age of the universe; scalar spectral index; curvature fluctuation amplitude; and reionization optical depth*, which we will try to estimate. The data from which we estimate the parameters,[3] often contain a plethora of other less-important data e.g. the calibration of the equipment known, as nuisance parameters, in which case the parameter space can have more than 40 dimensions. The main issues with said nuisance parameters is that they cannot easily be decoupled from the cosmological parameters of ΛCDM.

Among the possible approaches to parameter estimation, of particular interest is Bayesian inference, which allows us not only to estimate the likelihood of the parameters having particular values, but also determine the validity of our physical model. As an added bonus, Bayesian inference automatically accounts for nuisance parameters in its formalism, including them in the analysis, but excluding them from the end results.Bayesian inference in such a high dimensional parameter space is carried out using a form of Markov-Chain Monte-Carlo integration, the most common example being the Metropolis-Hastings algorithm,[1] or more recently Nested Sampling.[7,12]

In the following sections we shall outline the basics of Bayesian parameter fitting (Section. 2), the basics of Nested Sampling (Section. 4) and outline the methods of Prior re-partitioning applied to Nested sampling (Section. 4.2).

2 Bayesian parameter estimation

Let's assume that a scientific theory has a model for a process m which has n parameters $\{\theta\}$ (we shall drop the braces from now on). The real world observations give us a some data D. To verify the theory we're interested in the posterior: $\mathcal{P} = P(\theta|D, M)$ and the evidence $Z = P(D|M)$. Usually, the predictions of the model can straightforwardly give us two things: the prior $\pi(\theta) = P(\theta|D)$ and the Likelihood $L(\theta) = P(D|\theta, M)$. These quantities are linked via Bayes' theorem:

$$Z\mathcal{P}(\theta) = L(\theta)\pi(\theta),$$

which is a straightforward result, with profound implications.[9]

It is important to note the significance of Z. Some[9] regard it as a simple normalisation factor, and deem it completely irrelevant to our analysis, indeed even the name "Evidence" is, while agreed-upon, not standardised like the terms "prior" and "posterior". Of course it's an important factor determining the fitness of the ΛCDM model, and we need it just as much if not more than the posterior distribution.

To evaluate the *evidence*, we shall make use of its definition as a normalisation factor:

$$Z = \int L(\theta)\pi(\theta)d\theta. \tag{1}$$

For a completely new theory, determining the probability of each parameter given the model is difficult, so the prior is usually uniform within some constraints.[12,7] The probability of data, conditional on the data and the model is usually a Gaussian distribution, due to the central limit theorem. Thus we can both estimate how well our model fits the data, i.e. how certain are we that the universe is indeed ΛCDM and determine the distribution of the parameters.

The above task is more computationally expensive thus the scientific community rarely if ever goes beyond simple approximate but computationally inexpensive parameter fitting algorithms. For example, most experimental results are quoted with a single symmetric uncertainty value (e.g. $a = 1.0 \pm 0.2$), and model suitability, if investigated at all, is done via the χ^2 test,[10] which makes further assumptions about the nature of experimental data distribution.

In Section 3, we shall explore various approaches to performing full Bayesian inference.

3 Evaluating the evidence

3.1 Full rasterisation

To find the evidence Z What we're doing is essentially evaluating a high-dimensional integral over the possible values of the parameters (Let's assume that all the physical quantities are normalised: $\{\theta\} \in [0, 1]$). Under such conditions, the integral can be approximated by the Riemann Sum, with higher accuracy of the integral as a result of a larger number of points.

Needless to say that if we choose to rasterise with n points, the number of samples for D dimensional parameter space is n^D. Thus these techniques are inefficient compared to Nested Sampling.[7]

3.2 Metropolis-Hastings and other forms of Markov-chain Monte-Carlo

These approaches are a form of rejection sampling, thus only a subset of points is used in the evaluation of the integral. Markov Chain Monte-Carlo algorithms also suffer from arbitrary halting criteria.[1,6] For example, when simulating an Ising Spin array, the time taken to fully equilibrate the system near critical temperature is not easily predictable, and determining whether a system has equilibrated at later stages is simply not practical.[11]

Another drawback to MC-MC methods is that they're not easily made concurrent.[6] In my previous work, I have shown that the multi-process scaling of such parallel implementations is sub-optimal, and a more significant speedup can be obtained simply by running several simulations as full POSIX processes in parallel.[11]

3.3 The need for nested sampling

As we can see, all naive approaches brute force the multidimensional integral. In other words, we may see a speedup by sampling from representative points, and avoiding other non-representative samples as much as possible.

This was the basis for the paper due to John Skilling,[12] on a new machine learning technique, that allowed to minimise the number of so-called live points.

4 Nested sampling

There were multiple improvements[7,5,8] on the original algorithm, so we shall only present the ideas and avoid the unnecessary and obsolete details as much as possible.

We want to accumulate a quantity called *prior mass* (for reasons that will become clear soon), defined as:

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta) d\theta, \tag{2}$$

which is the cumulant prior mass covering all likelihood values greater than λ. Note that increasing λ decreases the value for $X(\lambda)$ from 1 to 0.

By exploiting the existence of the inverse function $L(X(\lambda)) \equiv \lambda$, we can simplify the evidence to

$$Z = \int_0^1 L(X) dX. \tag{3}$$

Thus by sampling randomly, given N points, we expect that the approximate value of the integral (2) is given by a Riemann sum:

$$Z = \sum_i^{N^2} \frac{L_i}{N}. \tag{4}$$

[See 12] for a more detailed explanation and example.

Next we might be interested in the order of convergence of such a method, so we are interested in the order of the error terms.

By using a more clever approximation for the integral[12] (e.g. using a trapezoidal rule) we can show that the error falls as $O(N^2)$.

4.1 Basic algorithm

We start by taking m random samples from the distribution (i.e. evaluating the likelihood for a single set of random physical parameter values). We shall refer to them as *live points*.

On each iteration, we take the point with the lowest likelihood and record it. Then we pick new values for each of the parameters according to some criteria, that have a higher likelihood than the old point. The old point now becomes a *dead point* while the new value is added to the *live points*.

As a result, a human observing the set of live points would notice that the latter preferentially occupy areas of high likelihood. Moreover there will be a contour of constant likelihood passing through each of the points (dead and live) and the live points will necessarily occupy the contours of highest likelihood.

Moreover, since we've picked points at random inside a D dimensional hyper-cube, the volumes delineated by the contours will correspond roughly to $\frac{1}{m}$-th of the total volume of the hyper-cube. This allows us to estimate the probabilities delineated by the contours and effectively evaluate the likelihood integral.

Moreover this being an approximate method, we can also estimate the error within each probability and incorporate that uncertainty into the analysis.

So bringing it all together we have the algorithm 1.

Algorithm 1: Nested Sampling. Credit[12]

Start with $\{\theta_1, \ldots, \theta_N\} \in$ prior;
initialise $Z = 0, X_0 = 1$;
for $i = 1, 2, \ldots,$ **to** j **do**
  set $L_i =$ lowest of current likelihood values,
  set $X_i = \exp(-i/N)$ (*crude*) `/* or sample it to get uncertainty,` `*/`
  set $w_i = X_{i-1} - X_i$ (*simple*) `/* or` $(X_{i-1} - X_{i+1})/2$ `(trapezoidal),` `*/`
  increment Z **by** $L_i w_i$
  replace point of lowest likelihood **with** new one drawn from within $L(\theta) > L_i$, `/* in proportion to the prior` $\pi(\theta)$`.` `*/`
end
increment Z **by** $N^{-1}\ (L(\theta_1) + ... + L(\theta_N))X_j$.

4.2 Improvements: Posterior re-partitioning

Looking at the algorithm[2] leads us to a few important conclusions. The shape of the prior matters, and as such if the prior and the posterior are the same, the algorithm should converge more rapidly, than with a flat prior.Moreover the prior is a function of what our parameters' definition. In other words a co-ordinate transformation can change the shape of the prior. Thus a clever choice of such a transformation of physical parameters θ can result in a significant speedup.

As to why it's called re-partitioning, because the true physical quantities of interest are really only dependent on the product $L(\theta)\pi(\theta)$ and which is which (i.e. where do we partition the likelihood from the prior) is only a matter of choice, thus we can easily come up with a more tractable form by moving the boundary — $\tilde{\pi}(\theta)$ as long as the product $L(\theta)\pi(\theta) = \tilde{L}(\theta)\tilde{\pi}(\theta)$ is the same.

For example,[2] consider the following re-partitioning:

$$\tilde{\pi}(\theta) = \frac{\pi(\theta)^\beta}{Z_\pi(\beta)}, \tag{5}$$

$$\tilde{L}(\theta) = L(\theta)\pi(\theta)^{(1-\beta)} Z_\pi(\beta), \tag{6}$$

where $Z_\pi(\beta) = \int \pi(\theta)^\beta d\theta$.

In this particular case we expect a speedup due to the re-scaled prior that is allowed to change according to the distribution, but with a trade-off of adding another parameter to the fit. Of course, the larger the number of already considered parameters, the more favourable the trade-off.

Additionally this particular re-partitioning scheme doesn't produce the expected result in the case where the prior has hard cutoffs.

5 Goal of the project

Thus the project shall need to be preoccupied with the following three main avenues. First we shall aim to test Nested Sampling in application to parameters which were generated using a toy distribution and evaluating the effects of different choices of priors and parameter values.

We shall then focus on attempting to improve the performance of said nested sampling, by considering other methods of re-partitioning: i.e. attempting a similar trick on the posterior half of the Bayes' theorem.

Accordingly we shall also investigate other re-partitioning schemes. In particular we shall consider cases where there are more parameters, and hard cutoffs, to cope with the limitations of the partitioning presented in.[2]

More concretely I shall:

- Run Polychord on data generated with an N-dimensional Correlated Gaussian and uniform prior.
- Implement the same with a multivariate Gaussian prior, and check that it is indeed faster than with a uniform prior.
- Investigate *mixture-model* prior re-partitioning, i.e. a re-partitioning scheme where $\tilde{\pi} = \beta\pi + (1-\beta)\pi_{\text{Gaussian}}$
- Attempt to apply to real data from Planck.[3]

Achieving all of the above will make the project at least a partial success. However, given enough time, and sufficient resources, there could be extra work that would be useful for this project to be a complete success.

- Investigate applicability of other mixture models,[8]
- Investigate applicability of re-partitioning schemes with a higher number of tuning parameters.
- Investigate the potential improvements arising from performing said operations algorithmic at compile-time.

References

[1] Gerhard Arminger and Bengt O. Muthén. "A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the metropolis-hastings algorithm". In: *Psychometrika* 63.3 (Sept. 1998), pp. 271–300. ISSN: 1860-0980. DOI: 10.1007/BF02294856.

[2] Xi Chen, Farhan Feroz, and Michael Hobson. *Bayesian automated posterior repartitioning for nested sampling*. 2019. eprint: arXiv:1908.04655.

[3] Planck Collaboration et al. *Planck 2018 results. VI. Cosmological parameters*. 2018. eprint: arXiv:1807.06209.

[4] J. J. Condon and A. M. Matthews. "*Lambda*CDM Cosmology for Astronomers". In: *Publications of the Astronomical Society of the Pacific* 130.989 (June 2018), p. 073001. DOI: 10.1088/1538-3873/aac1b2. URL: https://doi.org/10.1088/1538-3873/aac1b2.

[5] Farhan Feroz, Michael P. Hobson, and Michael Bridges. "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics". In: 2009.

[6] John Geweke and Hisashi Tanizaki. "Bayesian estimation of state-space models using the Metropolis–Hastings algorithm within Gibbs sampling". In: *Computational Statistics & Data Analysis* 37.2 (Aug. 2001), pp. 151–170. DOI: 10.1016/s0167-9473(01)00009-3.

[7] W. J. Handley, M. P. Hobson, and A. N. Lasenby. "polychord: next-generation nested sampling". In: *Monthly Notices of the Royal Astronomical Society* 453.4 (Sept. 2015), pp. 4384–4398. ISSN: 0035-8711. DOI: 10.1093/mnras/stv1911. eprint: http://oup.prod.sis.lan/mnras/article-pdf/453/4/4384/8034904/stv1911.pdf.

[8] Edward Higson et al. "Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation". In: *Statistics and Computing* (2018), pp. 1–23.

[9] Harold Jeffreys. *Scientific inference*. Cambridge: Cambridge University Press, 2010. ISBN: 978-0-521-18078-8.

[10] Karl Pearson. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (July 1900), pp. 157–175. DOI: 10.1080/14786440009463897. URL: https://doi.org/10.1080/14786440009463897.

[11] Aleksandr Petrosyan. "Simulation of Ferromagnetic behaviour using the Metropolis-Hastings algorithm within the Ising model." In: *NSTP Part III Furter Work* (2018).

[12] John Skilling. "Nested sampling for general Bayesian computation". In: *Bayesian Anal.* 1.4 (Dec. 2006), pp. 833–859. DOI: 10.1214/06-BA127. URL: https://doi.org/10.1214/06-BA127.