Opening up Nested Sampling

Andrew Fowlie

Xi'an Jiaotong Liverpool University

MaxEnt 2023 — 6 July 2023

A tribute from ChatGPT

In the realm of stats, there's a method we adore,
It's called nested sampling, let me tell you more.
It's a journey through likelihoods, a cosmic ride,
Exploring parameter space, with nothing to hide.

So let's raise a glass, to nested sampling's might,
A method that guides us through the statistical night.
With each iteration, we're closer to the truth,
Nested sampling, we salute your statistical sleuth!

Overview

1. Opening up the applications
2. Opening up the errors
3. Opening up the live points
4. Opening up the constrained sampling

A year of opening up after the pandemic

Opening up nested sampling

A year of opening up after the pandemic

I jinxed it because after submitting title I was no longer able to travel

Opening up the applications

Origins [1]

Bayesian Analysis (2006) **1**, Number 4, pp. 833–860

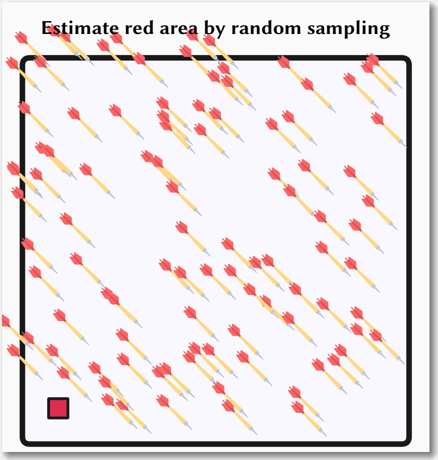
Nested Sampling for General Bayesian Computation

John Skilling*

Abstract. Nested sampling estimates directly how the likelihood function relates to prior mass. The evidence (alternatively the marginal likelihood, marginal density of the data, or the prior predictive) is immediately obtained by summation. It is the prime result of the computation, and is accompanied by an estimate of numerical uncertainty. Samples from the posterior distribution are an optional by-product, obtainable for any temperature. The method relies on sampling within a hard constraint on likelihood value, as opposed to the softened likelihood of annealing methods. Progress depends only on the shape of the "nested" contours of likelihood, and not on the likelihood values. This invariance (over monotonic re-labelling) allows the method to deal with a class of phase-change problems which effectively defeat thermal annealing.

Keywords: Bayesian computation, evidence, marginal likelihood, algorithm, nest, annealing, phase change, model selection

Volume and compression



We really need at least one dart to fall in red region

That won't work

- Estimate volume V by fraction of darts that fall in red region

$$\hat{V} = \frac{m}{n}$$

- Error of order Wald [2]

$$\frac{\Delta V}{V} = \sqrt{\frac{1/V}{n}}$$

- For fixed fractional uncertainty, number of samples scales as $1/V$
- Need $n \gtrsim 1/V$ at very least for reasonable estimate

NS for general compression computation

- NS rips an exponential factor out of the problem

$$\frac{\Delta V}{V} = \sqrt{\frac{\log 1/V}{n_{\text{live}}}}$$

- We expect though that number of samples

$$n = \frac{n_{\text{live}} \log 1/V}{\epsilon}$$

- Thus finally

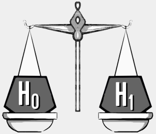
$$\frac{\Delta p}{p} = \sqrt{\frac{\log^2 1/V}{\epsilon n}}$$

NS for general frequentist computation

- The p-value is the probability of observing data as or more extreme than that observed, given the null hypothesis, H_0,

  $$p = P(\lambda \geq \lambda_{\text{Observed}} \,|\, H_0)$$

  where λ is a test-statistic that summarises the data and defines extremeness
- For a discovery in high-energy physics, we require $p < 10^{-7}$ as extraordinary claims require extraordinary evidence
- Compression from whole sampling space to region of size 10^{-7}
- Rejection sampling — throwing darts — requires too many samples, especially for high dimensional data



NS for approximate Bayesian computation

- What if our likelihood, $L(x) \equiv p(d_{\text{obs}} \,|\, x)$, is intractable?

$$\text{Cost}[\text{Evaluating } L(x)] \ggg \text{Cost}[\text{Sampling } d \sim p(d \,|\, x)]$$

- Approximate Bayesian computation
- For example, sample x, d from $p(x, d)$. Keep samples that lie within ϵ-ball of observed data
- Compression from whole joint space to ϵ-ball around observed data
- Rejection sampling requires too many samples, especially for high dimensional data

NS for general compression computation

Consider joint distribution

$$p(x, d)$$

- For Bayesian computation compress on parameter space, (x, d). From whole prior, upwards in likelihood
- For frequentist computation compress on sampling space, (x, d). From whole sampling space, upwards in test-statistic [3]
- For ABC compress on joint space, (x, d). From whole joint space, towards ϵ-ball around observed data

NS tackles all these compression problems

Opening up the errors

NS errors

- Powerful heuristic argument that NS error related to information by

$$\Delta \log Z \simeq \sqrt{\frac{H}{n_{\mathrm{live}}}}$$

- Standard deviation of Z available analytically from moments of X
- We should be able to show that the analytic answer approximately equals the heuristic one

$$\frac{\mathrm{Stdev}[Z]}{Z} \simeq \Delta \log Z \simeq \sqrt{\frac{H}{n_{\mathrm{live}}}}$$

Approximate equivalence

- Indeed, carefully expanding the sums, we can show their approximate equivalence [4], provided that

$$\text{Mean}[\log X] \gg \text{Stdev}[\log X]$$

- If not, can drive differences between them
- Agrees with Skilling's original argument — that error estimate holds if $\log X$ contains single narrow peak
- The computations shed light on NS errors

Pathological cases

- NS error estimates can diverge as number of iterations increases
- Simple way of creating pathological cases

$$L(X) \propto \frac{f(\log X)}{X}$$

  where f is the density with infinite variance, e.g., a Cauchy distribution
- Finite evidence since

$$Z = \int L(X) dX = \int f(\log X) d \log X$$

- Though because $P(\log X) = f(\log X)$, this trick means that $\text{Stdev}[\log X]$ diverges

Opening up the live points

Sampling from the constrained prior

How can we find an independent sample from the constrained prior?

- Required for our estimates of the volume, X
- Region and step samplers are popular techniques
- How do we know if they really work?
- Failure to correctly sample from the constrained prior leads to faulty estimates

What if we knew the X of every sample?

- Suppose we knew the X of every sample, $X(L_i)$
- We could look at
$$f_i = \frac{X(L_i)}{X(L^\star)}$$
- This should be uniformly distributed from 0 to 1 as each new $X(L_i)$ should be uniformly distributed from 0 to $X(L^\star)$
- We could test whether the f indeed followed a uniform distribution
- But we don't know the X of any samples

What do we know?

- We do know the likelihood of every new sample, L_i, and that $X(L)$ is a monotonic function
- So we can rank the n_{live} points by $X(L_i)$ by ranking them by L_i
- The rank of every new sample, r, should be uniformly distributed from 1 to n_{live}
- It's just as likely to be the worst, second worst, ..., second best, best likelihood
- We can test whether the r indeed follow a discrete uniform distribution [5]
- The NS equivalent of $\hat{R}$ diagnostic for MCMC — important check
- Successfully diagnoses faulty runs and a feature in latest NS software

Opening up the constrained sampling

Machine Learning

Physicist's perspective – statistics is an interface between theory and experiment

$$\text{Theory} \rightleftharpoons \text{Statistics} \rightleftharpoons \text{Experiment}$$

Machine Learning

Many people are going to try to persuade you to replace statistics with machine learning

$$\text{Theory} \rightleftharpoons \text{Machine Learning} \rightleftharpoons \text{Experiment}$$

Should you listen?

They might want to replace traditional theory activities e.g., model-building at the same time,

$$\text{Machine Learning} \rightleftharpoons \text{Machine Learning} \rightleftharpoons \text{Experiment}$$

E.g., symbolic regression

Machine Learning

I haven't quite seen this yet

$$\text{Machine Learning} \rightleftharpoons \text{Machine Learning} \rightleftharpoons \text{Machine Learning}$$

Maybe experiment is safe for now

How to combine statistics and machine learning?

Should we change how we learn?

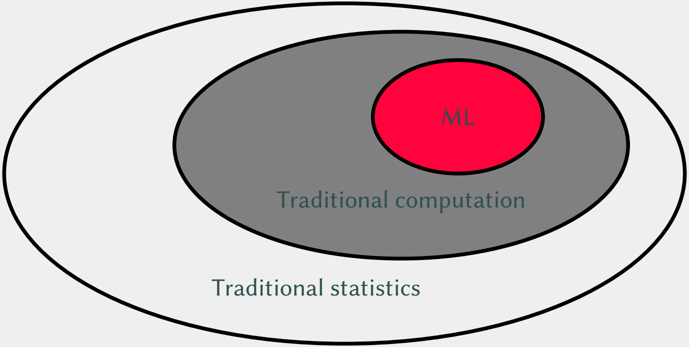
- No. Keep our statistical principles
- We should tell machine how to learn; not the other way around

Should we change how we compute?

- Not sure, lots of hype, though personally sceptical
- Computation should be principled; how else could it be trustworthy?
- We require uncertainty quantification and methods that generalise

How to combine statistics and machine learning?

Incorporate ML in well-defined mini-problems in our computation



Don’t sacrifice uncertainty quantification or reliability

Sampling from constrained prior

- NS perfectly open to this synergy between principled computation and ML
- Mini-problem in NS: sampling from the constrained prior
- You can do it more or less however you like
- See for example, ref. [6, 7] for normalising flow based sampling from constrained prior
- Check results using insertion index test

Conclusions

Nested sampling, we salute your statistical sleuth!

- Compression is a common problem; NS applicable to any compression problem
- Heuristic NS errors analysis is justified, though there are pathological cases
- Hidden information in live points allows cross check of NS runs — use it
- ML versus NS is false dichotomy; NS + ML possible by using ML for sampling from constrained prior
- Please see our review paper [8]

Nested sampling song I

(Verse 1)
In the realm of stats, there's a method we adore,
It's called nested sampling, let me tell you more.
It's a journey through likelihoods, a cosmic ride,
Exploring parameter space, with nothing to hide.
(Chorus)
Nested sampling, oh what a thrill,
Searching for the best fit, we can't sit still.
From the prior to the posterior, we navigate,
Finding the peak, it's a statistical fate.
(Verse 2)
Like a detective, we start with a prior guess,
Sampling points, we're on a quest, no time to rest.
Step by step, we dig deeper into the core,
Uncovering regions of likelihood we adore.
(Chorus)

Nested sampling song II

Nested sampling, oh what a thrill,
Searching for the best fit, we can't sit still.
From the prior to the posterior, we navigate,
Finding the peak, it's a statistical fate.
(Bridge)
As the iterations go by, the prior starts to shrink,
Leaving behind the most probable, it makes us think.
Evaluating evidence, quantifying the strength,
Nested sampling, it goes to any length.
(Chorus)
Nested sampling, oh what a thrill,
Searching for the best fit, we can't sit still.
From the prior to the posterior, we navigate,
Finding the peak, it's a statistical fate.
(Verse 3)
From astronomy to machine learning, it's a game-changer,

Nested sampling song III

Revolutionizing science, like a cosmic rearranger.
With nested sampling, our minds are blown away,
Unlocking insights, with every passing day.
(Chorus)
Nested sampling, oh what a thrill,
Searching for the best fit, we can't sit still.
From the prior to the posterior, we navigate,
Finding the peak, it's a statistical fate.
(Outro)
So let's raise a glass, to nested sampling's might,
A method that guides us through the statistical night.
With each iteration, we're closer to the truth,
Nested sampling, we salute your statistical sleuth!

[1] J. Skilling, *Nested sampling for general bayesian computation*, *Bayesian Analysis* **1** (2006) .

[2] L.D. Brown, T.T. Cai and A. DasGupta, *Interval estimation for a binomial proportion*, *Statistical Science* **16** (2001) .

[3] A. Fowlie, S. Hoof and W. Handley, *Nested Sampling for Frequentist Computation: Fast Estimation of Small p-Values*, *Phys. Rev. Lett.* **128** (2022) 021801 [2105.13923].

[4] A. Fowlie, Q. Li, H. Lv, Y. Sun, J. Zhang and L. Zheng, *Nested sampling statistical errors*, *Mon. Not. Roy. Astron. Soc.* **521** (2023) 4100 [2211.03258].

[5] A. Fowlie, W. Handley and L. Su, *Nested sampling cross-checks using order statistics*, *Mon. Not. Roy. Astron. Soc.* **497** (2020) 5256 [2006.03371].

[6] A. Moss, *Accelerated Bayesian inference using deep learning*, *Mon. Not. Roy. Astron. Soc.* **496** (2020) 328 [1903.10860].

References II

[7] M.J. Williams, J. Veitch and C. Messenger, *Nested sampling with normalizing flows for gravitational-wave inference*, *Phys. Rev. D* **103** (2021) 103006 [2102.11056].

[8] G. Ashton et al., *Nested sampling for physical scientists*, *Nature* **2** (2022) [2205.15570].