Nested Sampling tutorial

1 Motivation: exploring probability distributions

In astrophysics, we are often interested in exploring probability distributions $p(\theta)$. This can be accomplished in several different ways, such as:

- Metropolis-Hastings Monte Carlo,
- Hamiltonian Monte Carlo,
- Gibbs Sampling.

These methods are all *invariant* under scaling transformations of the distribution, $p(\theta) \to \alpha p(\theta)$ for some $\alpha \in \mathbb{R}^+$.

This may be useful, since the distributions we are often working with may have an *a-priori* unknown normalization.

When performing the analysis of data d according to a model M parameterized by some finite number of parameters θ, we employ Bayes' theorem, which is derived by expressing the joint probability $\text{prob}(d, \theta|M)$ in two different ways:

$$\text{prob}(d, \theta|M) = \mathcal{L}(d|\theta, M)\pi(\theta|M) = p(\theta|d, M)\mathcal{Z}(d|M)\,.$$

The entries in this equation are:

- the *likelihood* $\mathcal{L}$ of the data given the model and a specific choice of its parameters - it is a probability distribution over the data (i.e. $\int \mathcal{L}(d|\theta, M)\mathrm{d}d = 1$), but we interpret it as a function of the parameters θ while the data are fixed;
- the *prior* π - a representation of our beliefs about the model's parameters before seeing the data;
- the *posterior* p for the model's parameters - a representation of our updated beliefs about them after seeing the data;
- the *evidence* $\mathcal{Z}$ for the model - a normalization constant for the posterior, or the likelihood for the data marginalized over all the model's parameters.

Typically, it is cheap to evaluate the quantity $\mathcal{L}(d|\theta, M)\pi(\theta|M) \propto p(\theta|d, M)$ at any given point in parameter space.

All the aforementioned methods can give us samples from the unnormalized density $\mathcal{L}(d|\theta, M)\pi(\theta|M)$, but it is in trickier to estimate the **evidence** $Z(d|M)$.

This is what can be accomplished by *nested sampling*.

2 Nested sampling

Nested sampling, introduced by Skilling in the early 2000s (Skilling (2004), Skilling (2006)), is an algorithm designed to compute the *evidence* $\mathcal{Z}$ as well as the posterior distribution on the parameters.

The evidence is computed through the integral

$$\mathcal{Z}(d|M) = \int \mathcal{L}(d|\theta, M)\pi(\theta|M)\mathrm{d}^n\theta\,.$$

This integral is in n dimensions (where n is the dimensionality of the parameter space) and therefore difficult to work with. The idea in nested sampling is to rewrite it in terms of the auxiliary variable X, defined as the prior volume contained within a likelihood constraint $\mathcal{L} \geq L$:

$$X(L) = \int_{\mathcal{L}(d|\theta,M)\geq L} \pi(\theta|M)\mathrm{d}^n\theta\,.$$

Since the prior is a normalized probability distribution and the likelihood ranges from 0 to $L_{\max}$, this variable will range from $X(0) = 1$ to $X(L_{\max}) = 0$, and it will always be decreasing.[1] We can define its inverse, $L(X)$, and rewrite the likelihood integral through integration by parts: the evidence is the expectation value of the likelihood, therefore, starting from the definition of a Lebesgue integral:

$$\mathcal{Z}(d|M) = \int_0^{L_{\max}} X(L)\mathrm{d}L = XL|_{L=0}^{L=L_{\max}} - \int_{X(L=0)}^{X(L=L_{\max})} L(X)\mathrm{d}X = \int_0^1 L(X)\mathrm{d}X\,. \tag{1}$$

For a detailed discussion of the mathematical details, also see Ashton et al. (2022), Sivia and Skilling (2006).

[1] It will not necessarily be *strictly* decreasing, but that is not a conceptual issue. One can introduce jitter in the likelihood, varying it at each point by an inconsequential amount, in order to ensure that the decreasing condition is verified precisely.

2.1 Prior compression

Note that, by definition, X is directly related to probability mass: the prior probability of an interval $[X_0, X_1]$, i.e. the prior mass which has likelihood values $L(X_1) < \mathcal{L} < L(X_0)$, is

$$p([X_0, X_1]) = \int_{L(X_0)<\mathcal{L}(d|\theta,M)<L(X_1)} \pi(\theta|M)\mathrm{d}^n\theta = X_1 - X_0\,,$$

where we computed the integral on a likelihood "ring" by subtracting the inside volume from the outside one. This is useful to us, since it means that if we distribute points uniformly according to the prior their X values will also be uniformly distributed in $[0, 1]$. More formally, the probability we are assigning to X is the *push-forward* probability measure associated to the mapping from θ to X defined by $X(\theta) = X(L(\theta))$. For a discussion on this, see (Ashton et al. 2022, box 2).

If k points are uniformly distributed in an interval $[0, X^*]$, then the largest of them will have a X coordinate distributed as $X_{\max}/X^* \equiv t \sim \mathrm{beta}(k, 1) = kt^{k-1}$, where we defined the *compression ratio* t: when we discard the lowest-likelihood point with $X_{\max}$, we will have contracted the prior volume by a factor t.

> **Note 1: On the beta distribution**
>
> Here is a simple argument for why t, the largest out of k uniform random variates in $[0, 1]$, is distributed according to $t \sim kt^{k-1}$. The statement can equivalently be framed in terms of the cumulative distribution: $p(t \leq T) = T^k$.
> We can think of our k random variates as a point in the k-dimensional unit box $[0, 1]^k$, and since they are uniformly distributed there will be a one-to-one correspondence between volume within the box and probability mass. Then, the geometric meaning of the cumulative distribution is: what is the volume of the region corresponding to the event $t \leq T$, i.e. such that the maximum of the coordinates of our point is T? The answer is the box $[0, T]^k$, whose volume is T^k, thus proving our claim.
> In general, the n-th smallest out of k uniform random variates in $[0, 1]$ is distributed according toa beta distribution, specifically $\mathrm{beta}(k, n-k+1) \propto x^{n-1}(1-x)^{k-n}$; here we are looking at the k-th smallest (i.e. the largest).

In terms of $\log t$, the distribution reads

$$\frac{\mathrm{d}p}{\mathrm{d}\log t} = t\frac{\mathrm{d}p}{\mathrm{d}t} = kt^k = ke^{k\log t}$$

This distribution is shown in Figure 1.

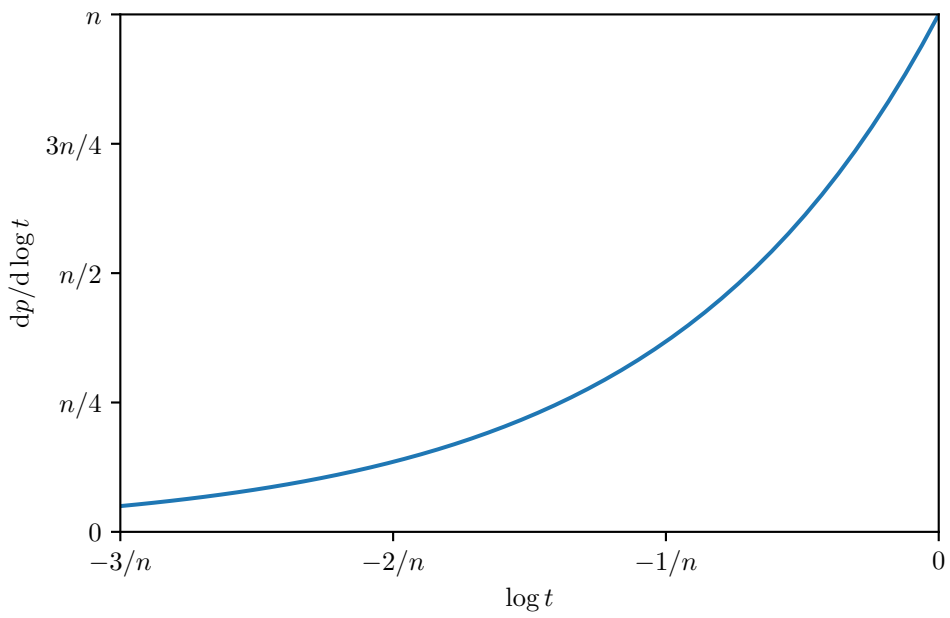

Figure 1: Volume compression

Its mean and standard deviation are:

$$\begin{aligned}\langle \log t \rangle &= \int_{-\infty}^{0} \log t \frac{\mathrm{d}p}{\mathrm{d}\log t} \mathrm{d}\log t \\ &= \int_{-\infty}^{0} \frac{x}{k} k e^x \frac{\mathrm{d}x}{k} \\ &= \frac{1}{k} \int_{-\infty}^{0} x e^x \mathrm{d}x \\ &= -\frac{1}{k}\end{aligned}$$

where $x = k \log t$, and

$$\begin{aligned}\langle (\log t)^2 \rangle - \frac{1}{k^2} &= \int_{-\infty}^{0} (\log t)^2 \frac{\mathrm{d}p}{\mathrm{d}\log t} \mathrm{d}\log t - \frac{1}{k^2} \\ &= \int_{-\infty}^{0} \frac{x^2}{k^2} k e^x \frac{\mathrm{d}x}{k} - \frac{1}{k^2} \\ &= \frac{1}{k^2} \int_{-\infty}^{0} x^2 e^x \mathrm{d}x - \frac{1}{k^2} \\ &= \frac{2-1}{k^2} = \frac{1}{k^2}\,.\end{aligned}$$

Compactly, we can then write $\log t = (-1 \pm 1)/k$.

2.2 Computing evidence and posterior

If we sample k points from the prior (with total volume X_0) and remove the one with the worst likelihood, the remaining prior volume will be $X_1 = t_1 X_0$, where $\log t_1$ has an expectation value of $-1/k$. This will be true at every iteration, so (Skilling 2006, sec. 5)

$$\log X_i = \sum_{j \leq i} \log t_j \approx -\frac{i}{k} \pm \frac{\sqrt{i}}{k}$$

The prior is compressed exponentially, which is desirable since the typical set for the posterior is often several orders of magnitude smaller than the prior.

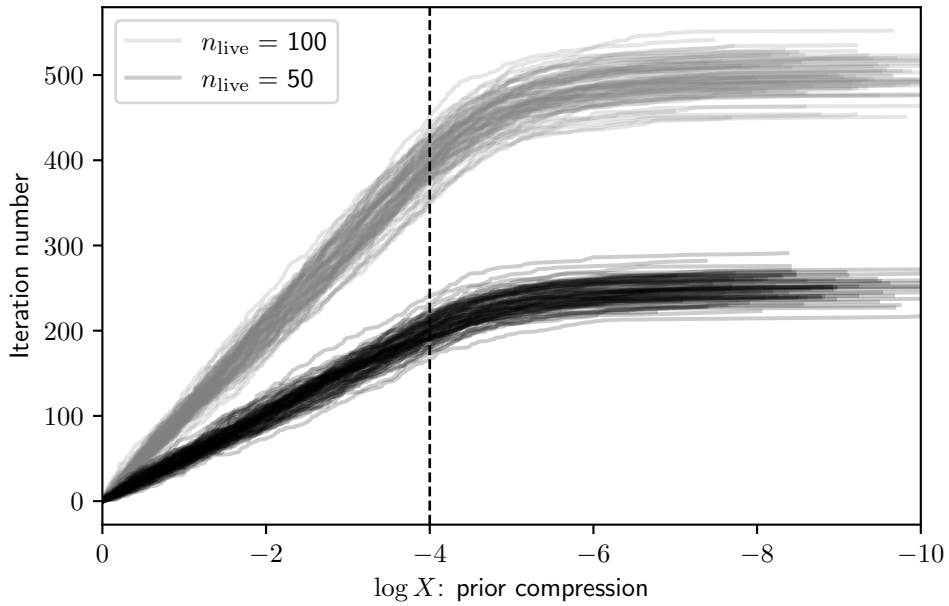

Figure 2: Volume compression by iteration. The number of live points is denoted here as $n_{\text{live}} = k$. Note how the horizontal slices are thinner when we have more live points: increasing them will decrease the uncertainty on the prior compression obtained at a fixed iteration number. After reaching a logarithmic volume compression of $-\log X = 4$ the runs are truncated, and the remaining points are shown: these are distributed uniformly and not geometrically in X, therefore they taper off in this logarithmic plot.

This allows us to construct a procedure to evaluate the integral in Equation 1, as long as we are able to do the following:

1. sample uniformly k points from the prior
2. find the lowest-likelihood point θ_1 and discard it: now the prior volume is approximately $X_1 = e^{-1/k}$; let us denote its likelihood as L_1
3. sample a new point θ_2 uniformly from the prior, constrained to $L_2 = L(\theta_2) > L_1$
4. repeat from point 2.

We are constructing a sequence of points θ_i with corresponding likelihoods L_i and approximate associated interior prior volumes $X_i \approx e^{-i/k}$. At this stage, we still need to discuss how step 3 is performed, as well as when the iteration should stop - for now, suppose that happens at some stage $n_{\rm iter}$.

For the evidence, we can simply use a trapezoidal integration rule:

$$\mathcal{Z}(d|M) \approx \sum_{i=1}^{n_{\rm iter}} w_i L_i = \sum_{i=1}^{n_{\rm iter}} \frac{X_{i+1} - X_{i-1}}{2} L_i$$

It can be shown that this sum converges to the correct value in the $n_{\rm iter} \to \infty$ and $n_{\rm live} \to \infty$ limit.

These weights can also be used to estimate the posterior as follows: take the points θ_i and assign to each of them a weight

$$p_i = \frac{w_i L_i}{Z} \,.$$

It is readily seen that these weights add to 1, and heuristically they are computed as one might expect, by multiplying the prior mass and likelihood value for each of the regions we have subdivided the prior volume into.

Often one is be interested in a set of equally-weighted samples: these can be obtained, for example, through density estimation methods.

3 Exercise

Suppose we have a set of data points (x_i, y_i) with uncertainties σ_y on the y values.

We will compare two simple *nested* models, namely a parabolic one with $y = f(x; a, b, c) = ax^2 + bx + c$ and a linear one with $y = f(x; b, c) = bx + c$. Model comparison can also be performed for non-nested models, but the nested case has some nice properties we will showcase later.

Let us generate the data and plot it: it will come from the parabolic model.

We can then analyze this data, using a Gaussian likelihood and setting (somewhat arbitrarily) uniform priors for the parameters a, b and c in the range $[-10, 10]$:

- $\pi(a, b, c) = 1/20^3$ in that range and 0 otherwise for the parabola;

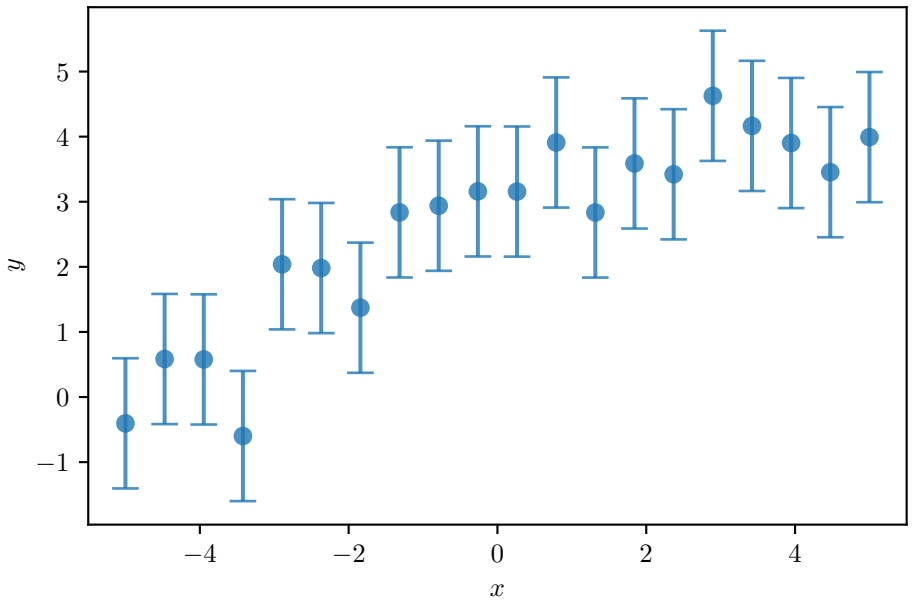

Figure 3: Toy model, with data coming from a parabolic function and Gaussian errors.

- $\pi(b, c) = 1/20^2$ in that range and 0 otherwise for the parabola;

The log-likelihood will be

$$\log \mathcal{L}(\text{data}|\theta, \text{model}) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - f(x_i;\theta))^2}{\sigma_y^2} - \frac{n}{2}\log(2\pi\sigma_y^2).$$

Here is a reference python implementation of the likelihoods and prior transform:

3.1 Inference

Perform Bayesian inference for the two models with your favourite variant of Nested Sampling, which will yield estimates for the posterior distributions of the parameters as well as the evidences. Options include:

1. dynesty
2. ultranest
3. PyMultiNest
4. bilby
5. nessai
6. ...

3.2 Bayes Factor

Compute the Bayes Factor: which model is preferred over the other? It is defined as follows:

$$\log \mathrm{BF}^{\mathrm{parabola}}_{\mathrm{line}} = \log Z_{\mathrm{parabola}} - \log Z_{\mathrm{line}} \,.$$

3.3 Savage-Dickey Density Ratio

The models we are considering are nested, so we have a comparison point for our Bayes Factor: the Savage-Dickey density ratio. Show that this estimate is consistent with the Bayes Factor computed with Nested Sampling. It is computed as follows:

$$\mathrm{BF}^{\mathrm{parabola}}_{\mathrm{line}} = \frac{p(a=0|d, M_{\mathrm{parabola}})}{\pi(a=0|M_{\mathrm{parabola}})} \,.$$

References

Ashton, Greg, Noam Bernstein, Johannes Buchner, Xi Chen, Gábor Csányi, Andrew Fowlie, Farhan Feroz, et al. 2022. "Nested Sampling for Physical Scientists." *Nature Reviews Methods Primers* 2 (1): 1–22. https://doi.org/10.1038/s43586-022-00121-x.

Sivia, Devinderjit, and John Skilling. 2006. *Data Analysis: A Bayesian Tutorial*. Oxford University Press. https://books.google.com?id=lYMSDAAAQBAJ.

Skilling, John. 2004. "Nested Sampling." In *AIP Conference Proceedings*, 735:395–405. Garching (Germany): AIP. https://doi.org/10.1063/1.1835238.

———. 2006. "Nested Sampling for General Bayesian Computation." *Bayesian Analysis* 1 (4): 833–59. https://doi.org/10.1214/06-BA127.