Nested Sampling:

A multi-purpose numerical tool for science and machine learning

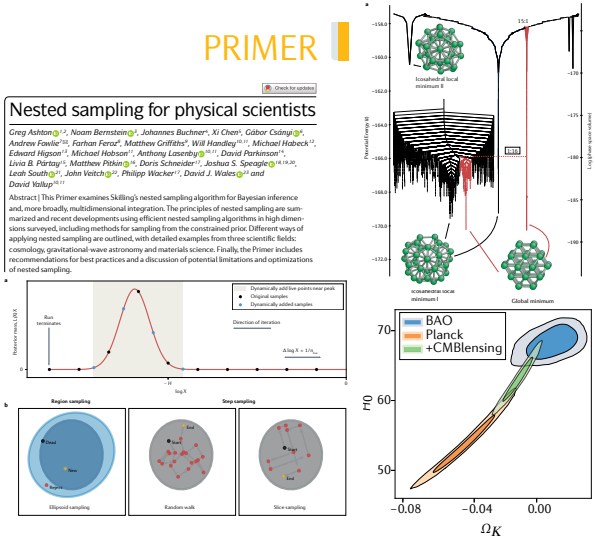
Will Handley
<wh260@cam.ac.uk>

Royal Society University Research Fellow & Turing Fellow
Astrophysics Group, Cavendish Laboratory, University of Cambridge
Kavli Institute for Cosmology, Cambridge
Gonville & Caius College
willhandley.co.uk/talks

20th March 2023

Highlight: state-of-the-art Nature review [NatRev]

- Invented by John Skilling in 2004.
- Recent Nature review primer on nested sampling led by Andrew Fowlie and assembled by the community.
- Showcases the current set of tools, and applications from chemistry to cosmology.
- In this talk
  - User guide to nested sampling
  - Particle physics applications
  - Cosmology applications

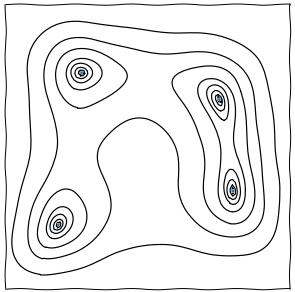


What is Nested Sampling?

- Nested sampling is a radical, multi-purpose numerical tool.
- Given a (scalar) function f with a vector of parameters θ, it can be used for:

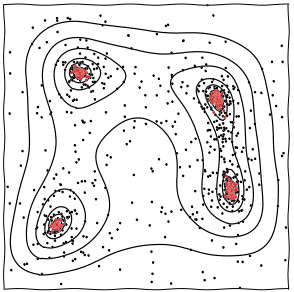
Optimisation

$$\theta_{\max} = \max_{\theta} f(\theta)$$



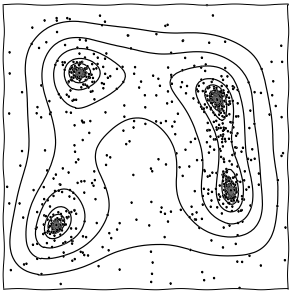
Exploration

draw/sample $\theta \sim f$



Integration

$$\int f(\theta) dV$$



The three pillars of Bayesian inference

Parameter estimation

What do the data tell us about the parameters of a model? *e.g. the size or age of a ΛCDM universe*

$$P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}$$

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Model comparison

How much does the data support a particular model? *e.g. ΛCDM vs a dynamic dark energy cosmology*

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

$$\frac{\mathcal{Z}_{\mathcal{M}}\Pi_{\mathcal{M}}}{\sum_m \mathcal{Z}_m \Pi_m}$$

$$\text{Posterior} = \frac{\text{Evidence} \times \text{Prior}}{\text{Normalisation}}$$

Tension quantification

Do different datasets make consistent predictions from the same model? *e.g. CMB vs Type IA supernovae data*

$$\mathcal{R} = \frac{\mathcal{Z}_{AB}}{\mathcal{Z}_A \mathcal{Z}_B}$$

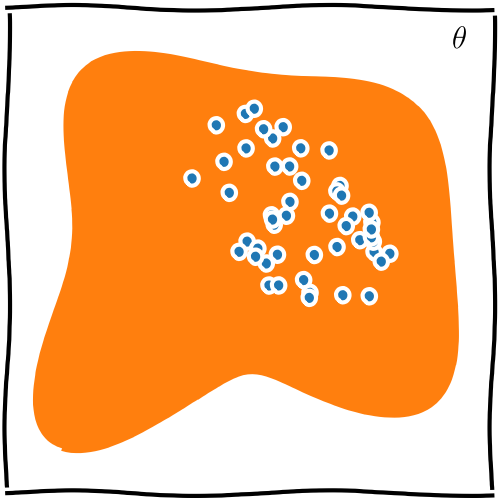
$$\begin{aligned}\log \mathcal{S} = &\langle \log \mathcal{L}_{AB} \rangle_{\mathcal{P}_{AB}} \\ &- \langle \log \mathcal{L}_A \rangle_{\mathcal{P}_A} \\ &- \langle \log \mathcal{L}_B \rangle_{\mathcal{P}_B}\end{aligned}$$

Why do sampling?

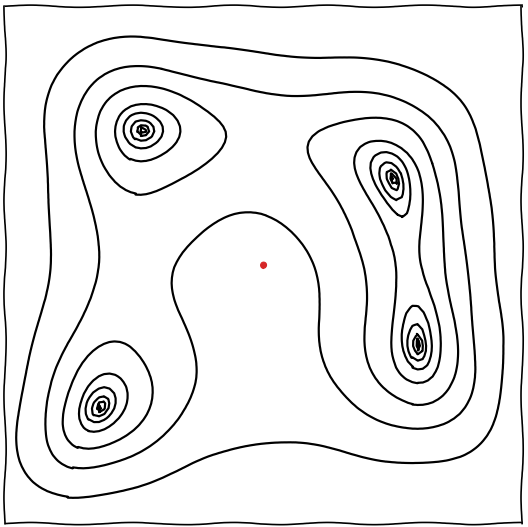
- The cornerstone of numerical Bayesian inference is working with **samples**.
- Generate a set of representative parameters drawn in proportion a distributionto the posterior $\theta \sim \mathcal{P}$.
- The magic of marginalisation $\Rightarrow$ perform usual analysis on each sample in turn.
- The golden rule is stay in samples until the last moment before computing summary statistics/triangle plots because

$$\boxed{f(\langle X \rangle) \neq \langle f(X) \rangle}$$

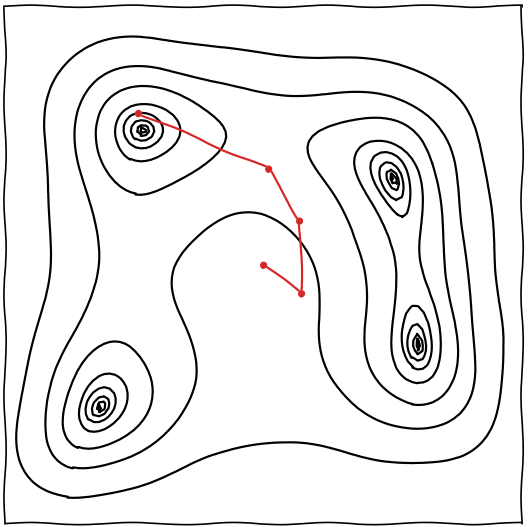
- Generally need $\sim \mathcal{O}(12)$ independent samples to compute a value and error bar.



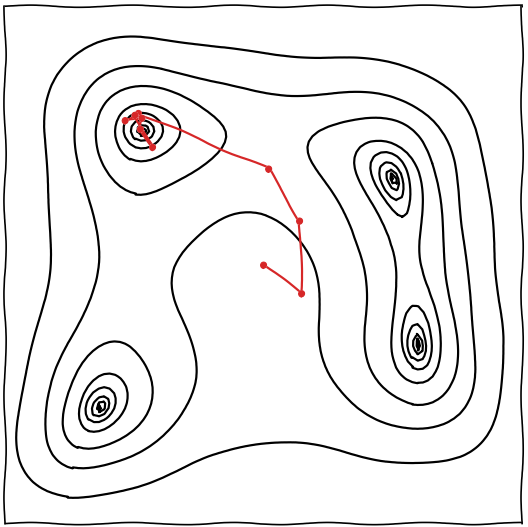
MCMC



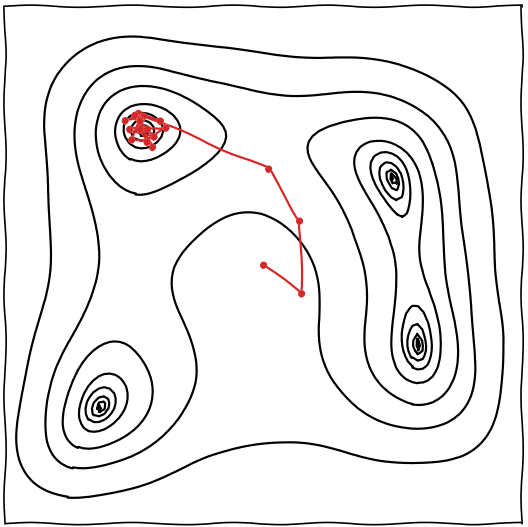
MCMC



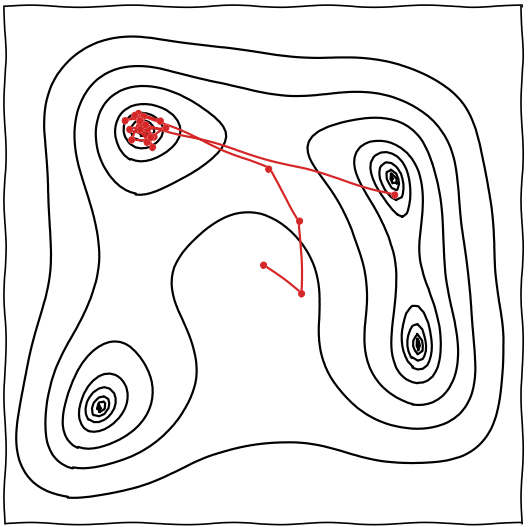
MCMC



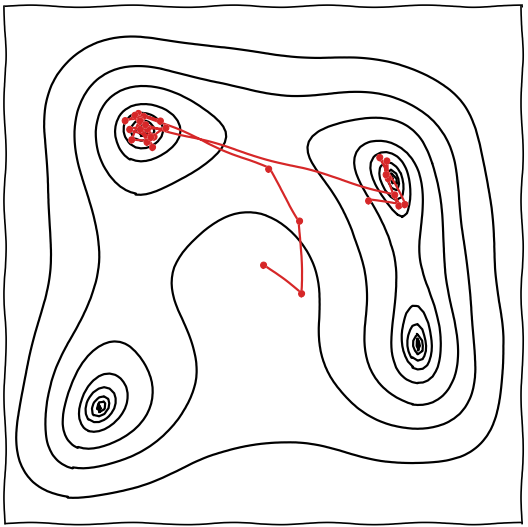
MCMC



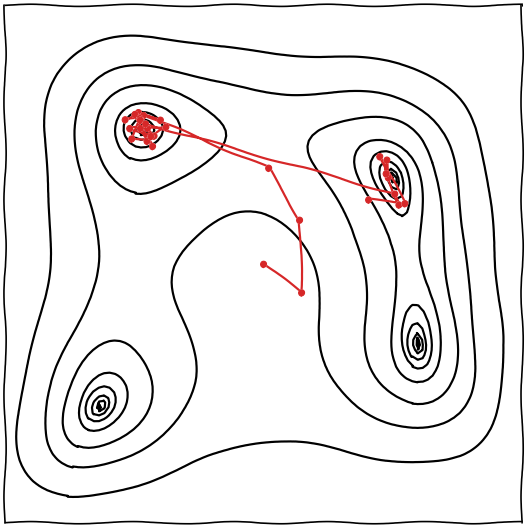
MCMC



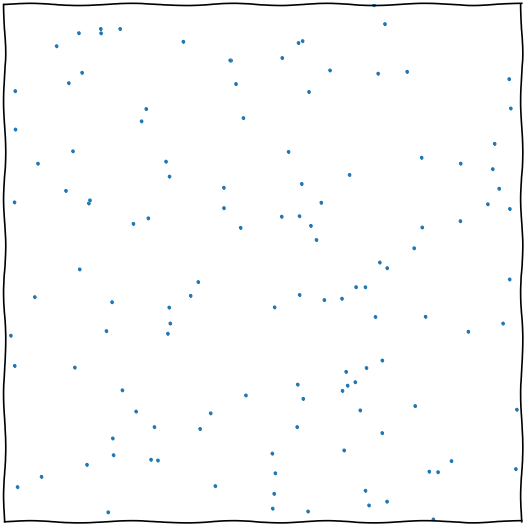
MCMC



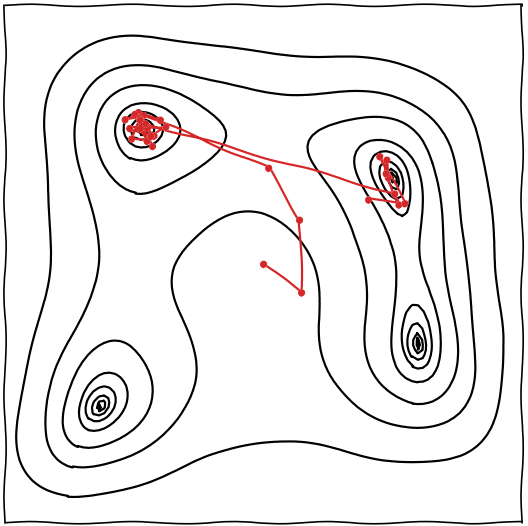
MCMC



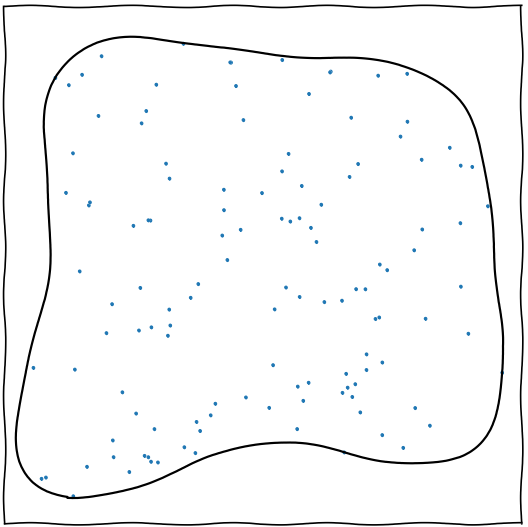
Nested sampling



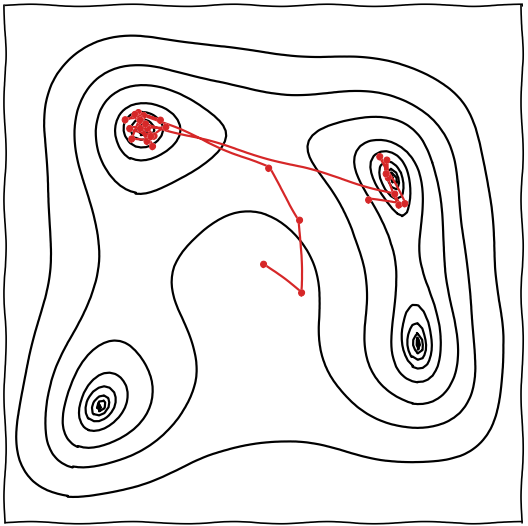
MCMC



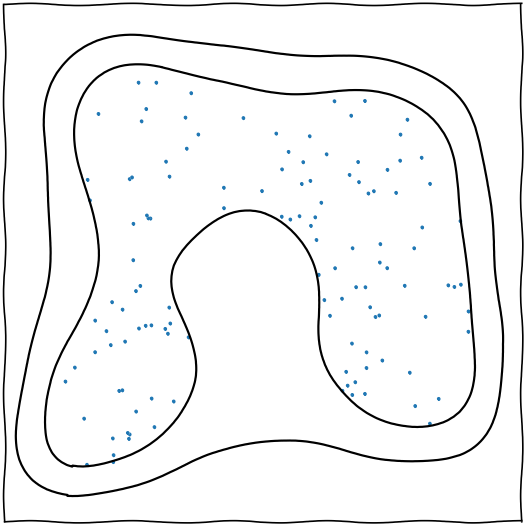
Nested sampling



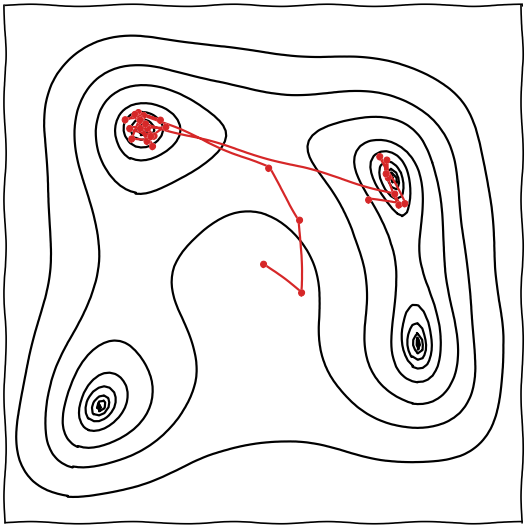
MCMC



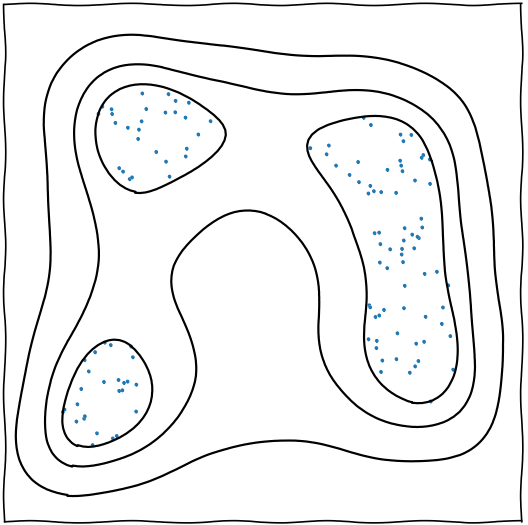
Nested sampling



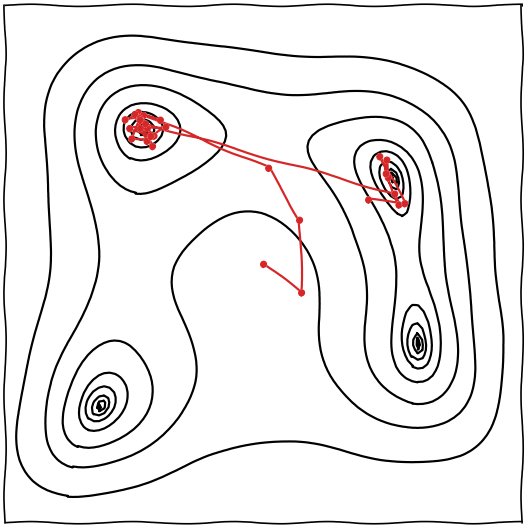
MCMC



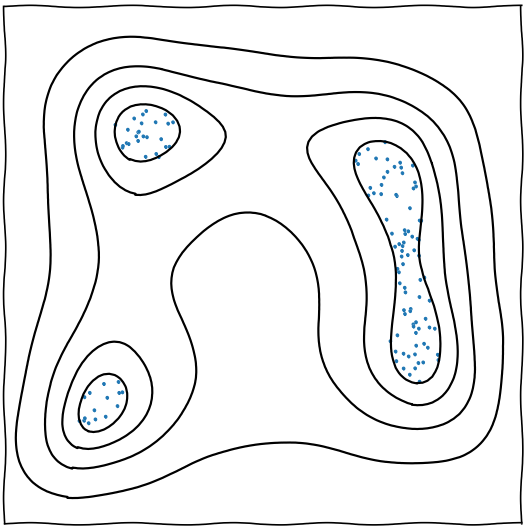
Nested sampling



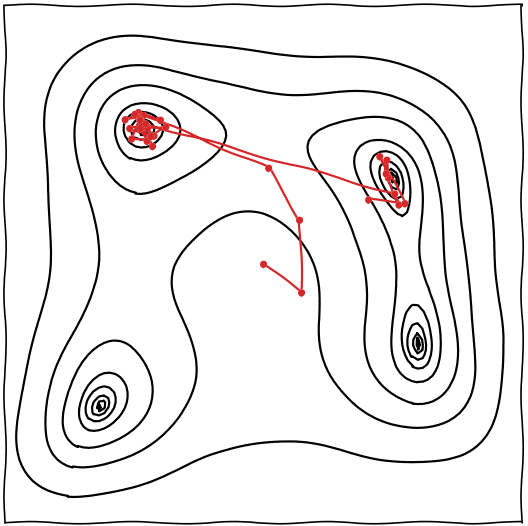
MCMC



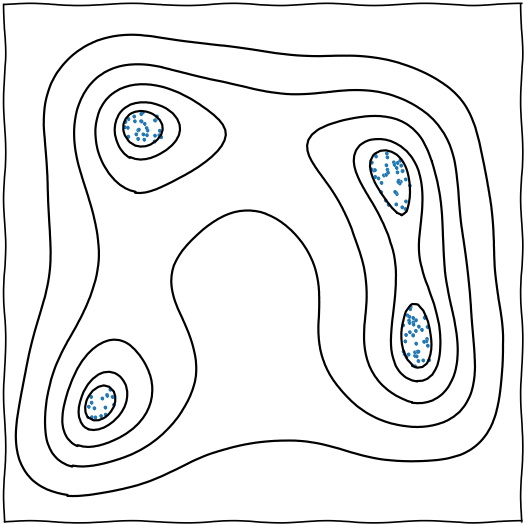
Nested sampling



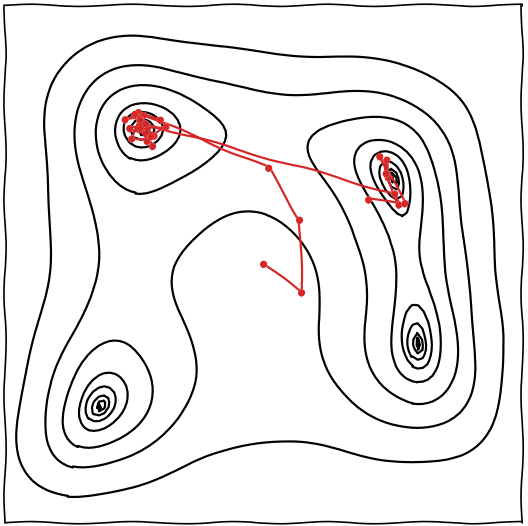
MCMC



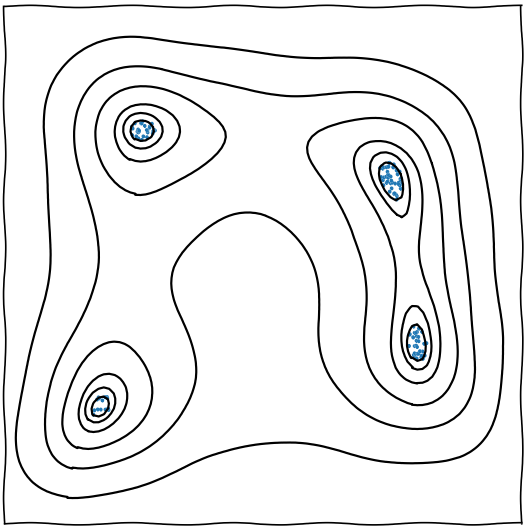
Nested sampling



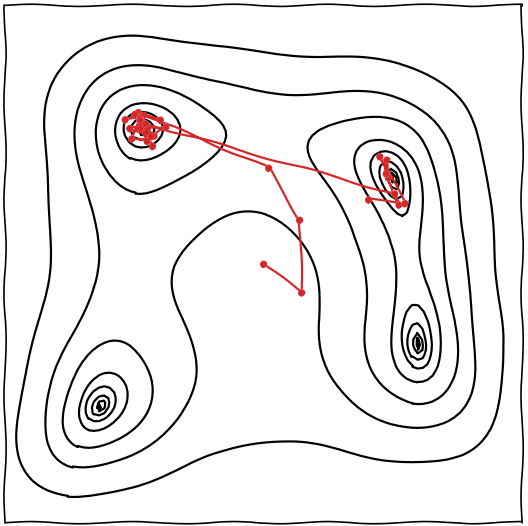
MCMC



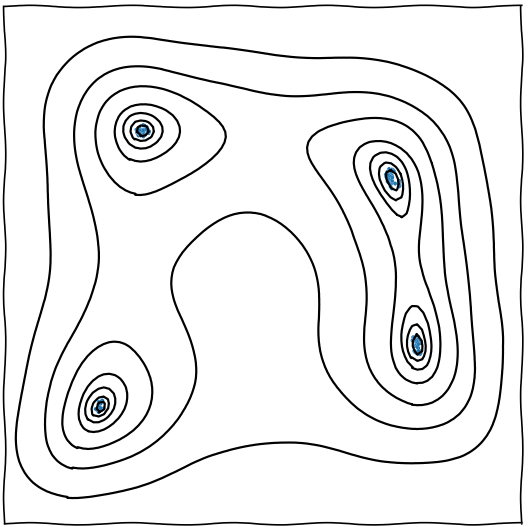
Nested sampling



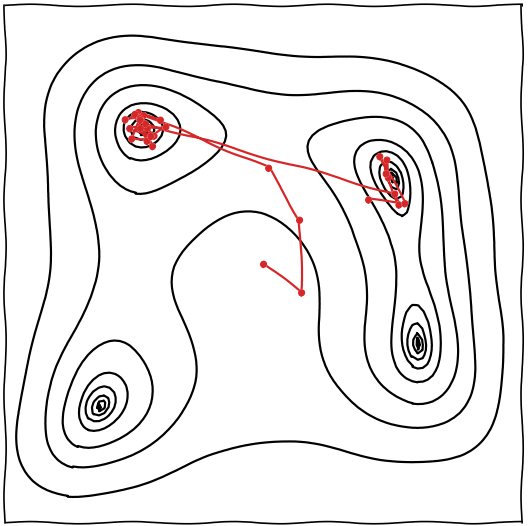
MCMC



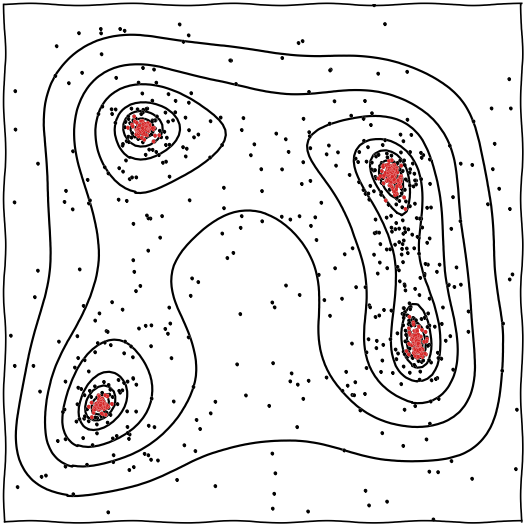
Nested sampling



MCMC

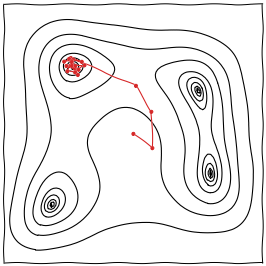


Nested sampling



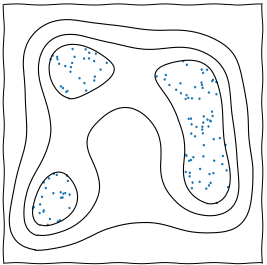
MCMC

- Single "walker"
- Explores posterior
- Fast, if proposal matrix is tuned
- Parameter estimation, suspiciousness calculation
- Channel capacity optimised for generating posterior samples



Nested sampling

- Ensemble of "live points"
- Scans from prior to peak of likelihood
- Slower, no tuning required
- Parameter estimation, model comparison, tension quantification
- Channel capacity optimised for computing partition function



Nested sampling

- Sequentially update a set S of n samples:

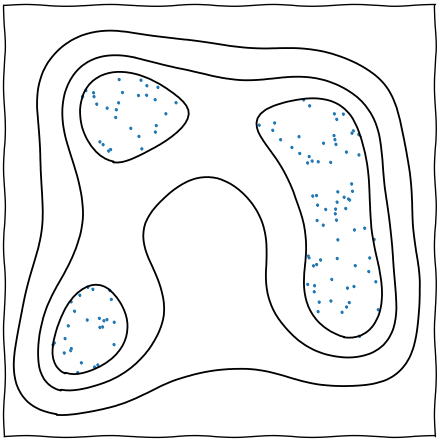
  S_0: Generate n samples uniformly over the space (from the prior π).

  S_{i+1}: Delete the lowest likelihood sample in S_i, and replace it with a new uniform sample with higher likelihood.

- Requires one to be able to sample uniformly within a region, subject to a *hard likelihood constraint*:

$$\{\theta \sim \pi : \mathcal{L}(\theta) > \mathcal{L}_*.\}$$

- This procedure optimises (multimodally), and can calculate the evidence & posterior weights.
- The evolving ensemble of live points allows algorithms to perform self-tuning and mode clustering.



Integration in Physics

- Integration is a fundamental concept in physics, statistics and data science:

Partition functions

$$Z(\beta) = \int e^{-\beta H(q,p)} dq dp$$

Path integrals

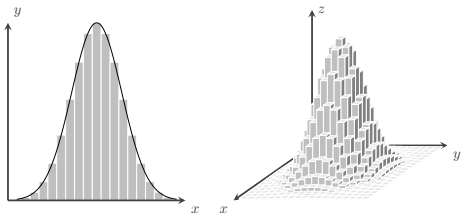
$$\Psi = \int e^{iS} \mathcal{D}x$$

Bayesian marginals

$$\mathcal{Z}(D) = \int \mathcal{L}(D|\theta)\pi(\theta) d\theta$$

- Need numerical tools if analytic solution unavailable.
- High-dimensional numerical integration is hard.
- Riemannian strategy estimates volumes geometrically:

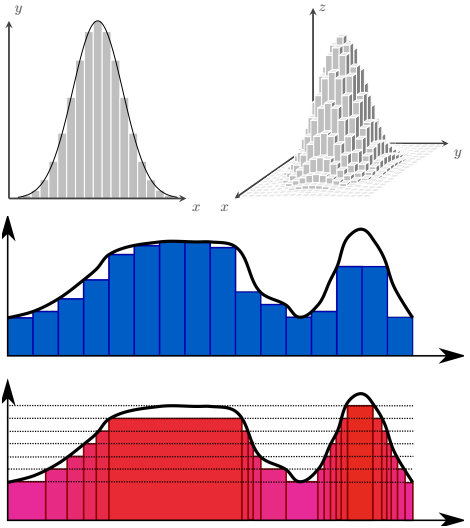
$$\int f(x) d^n x \approx \sum_i f(x_i) \Delta V_i \sim \mathcal{O}(e^n)$$



- Curse of dimensionality $\Rightarrow$ exponential scaling.
- Nested sampling integrates **probabilistically**.

Integration in high dimensions

- Nested sampling can compute the Bayesian evidence $\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta$
- Numerical integration $\int f(x)dV$ in high dimensions is hard.
- `scipy.integrate(...)` is unusable in more than four dimensions.
- This is due to the curse of dimensionality: need to sum $\sim N^d$ units to compute $\approx \sum_i f(x_i)\Delta V_i$.
- Additionally, estimating volumes with geometry becomes exponentially hard as d increases.
- *Aside:* **Riemannian integration** *(blue) is taught as standard. An orthogonal approach (red) [usually theoretical] is* **Lesbesgue integration**.

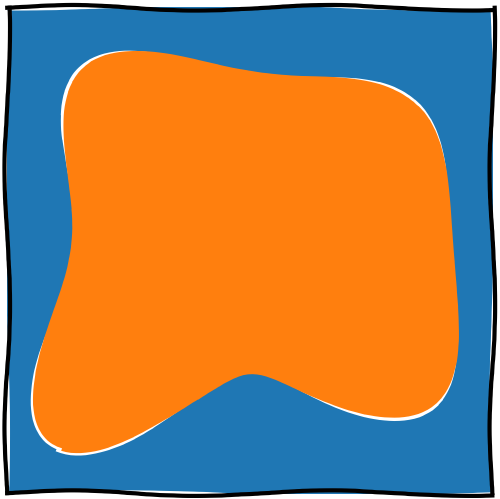


Probabalistic volume estimation

- Key idea in NS: estimating volumes probabilistically

$$\frac{V_{\text{after}}}{V_{\text{before}}} \approx \frac{n_{\text{in}}}{n_{\text{out}} + n_{\text{in}}}$$

- This is the **only** way to calculate volume in high dimensions $d > 3$.
  - Geometry is exponentially inefficient.
- This really is the unique selling point of nested sampling.

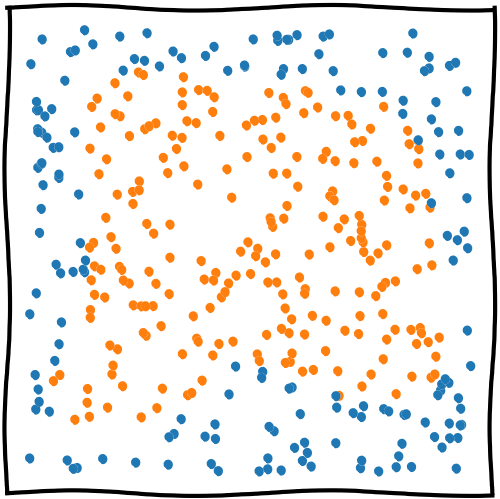


Probabalistic volume estimation

- Key idea in NS: estimating volumes probabilistically

$$\frac{V_{\text{after}}}{V_{\text{before}}} \approx \frac{n_{\text{in}}}{n_{\text{out}} + n_{\text{in}}}$$

- This is the **only** way to calculate volume in high dimensions $d > 3$.
  - Geometry is exponentially inefficient.
- This really is the unique selling point of nested sampling.



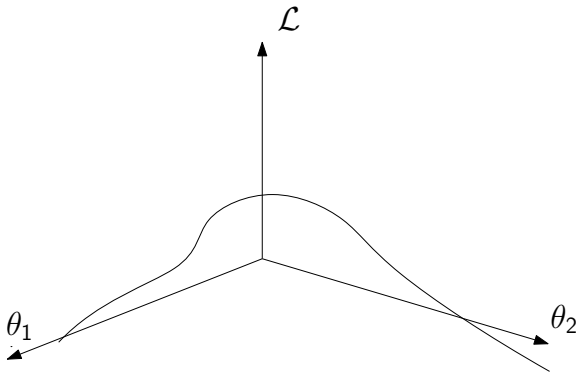
(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$



- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$

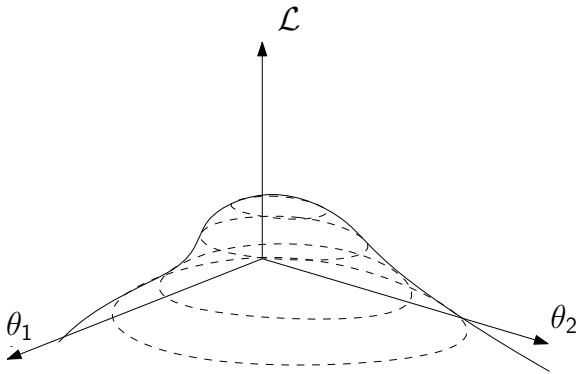
(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$



- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$

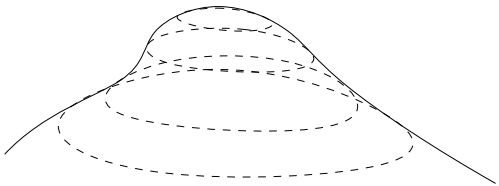
(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$



- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$

(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

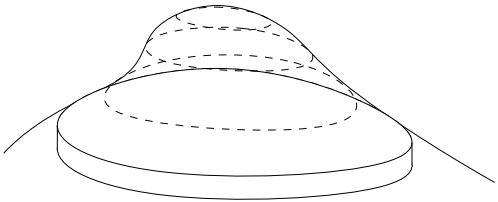
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

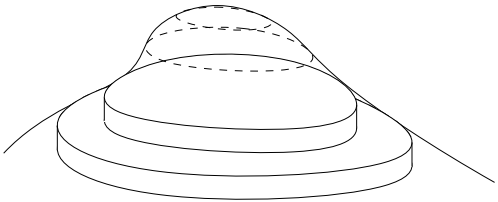
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

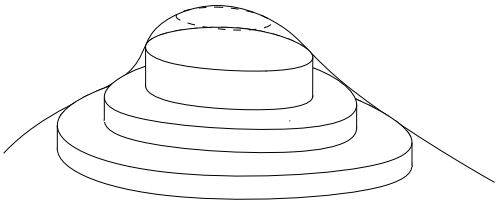
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

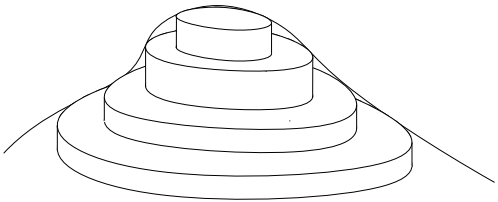
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

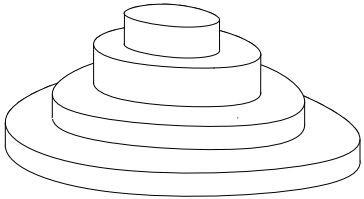
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

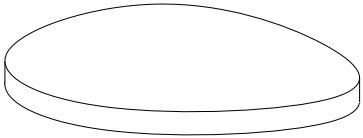
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$



- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

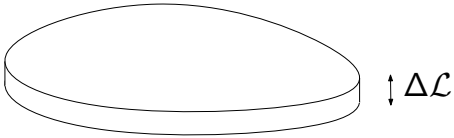
- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$

(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$



- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

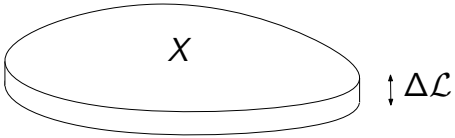
- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$

(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$



- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$

(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

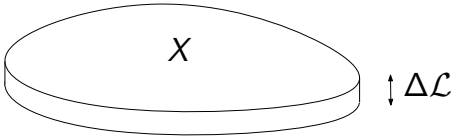
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{n X_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

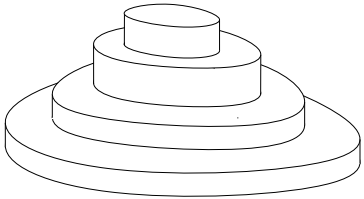
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

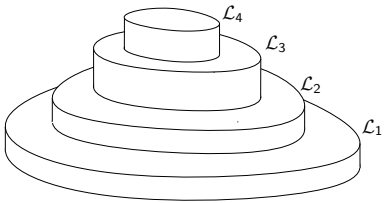
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

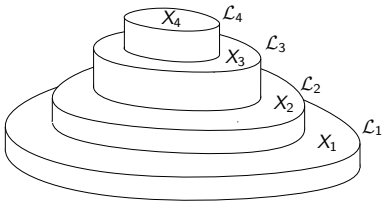
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



(Lesbesgue) Integrating with nested sampling

- At each iteration, the likelihood contour will shrink in volume X by $\approx 1/n$.
- Nested sampling zooms in to the peak of the function $\mathcal{L}$ *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

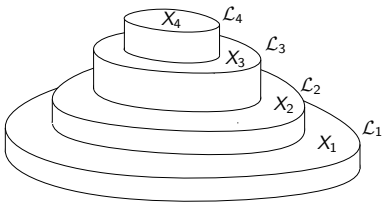
- Although this is only approximate, we can quantify the error

$$P(X_i|X_{i-1}) = \frac{X_i^{n-1}}{nX_{i-1}^n} \times [0 < X_i < X_{i-1}].$$

- Integral can be discretised in several ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$

$$\mathcal{Z} \approx \sum_i X_i \Delta\mathcal{L}_i$$

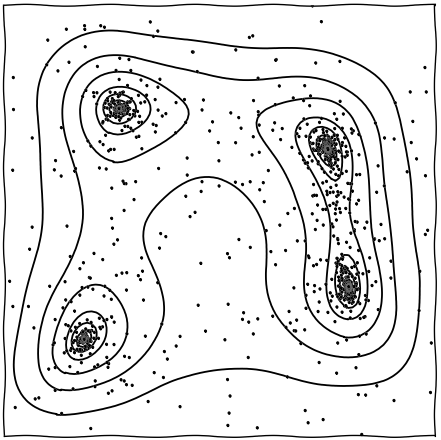


Dead points: posteriors & evidences

- At the end, one is left with a set of discarded points.
- These may be weighted to form weighted posterior samples using $w_i = \mathcal{L}_i \Delta X_i$.
- They can also be used to calculate the integral $\mathcal{Z} = \sum \mathcal{L}_i \Delta X_i$, or more generally $\sum_i f(\mathcal{L}_i) \Delta X_i$.
  - Nested sampling probabilistically estimates the volume of the parameter space

$$X_i \approx \left(\frac{n}{n+1}\right) X_{i-1} \quad \Rightarrow \quad X_i \approx \left(\frac{n}{n+1}\right)^i \approx e^{-i/n},$$

  - Nested sampling thus estimates the density of states,
  - it is therefore a partition function calculator $Z(\beta) = \sum_i \mathcal{L}_i^\beta \Delta X_i$.
- The evolving ensemble of live points allows algorithms to perform self-tuning and mode clustering.

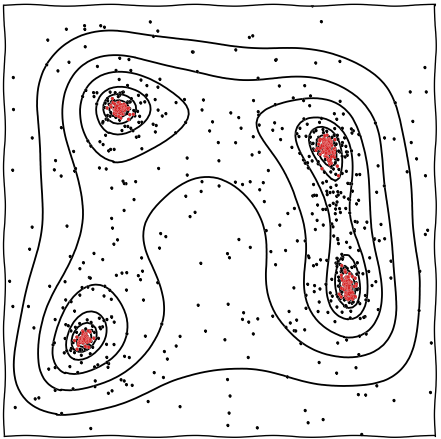


Dead points: posteriors & evidences

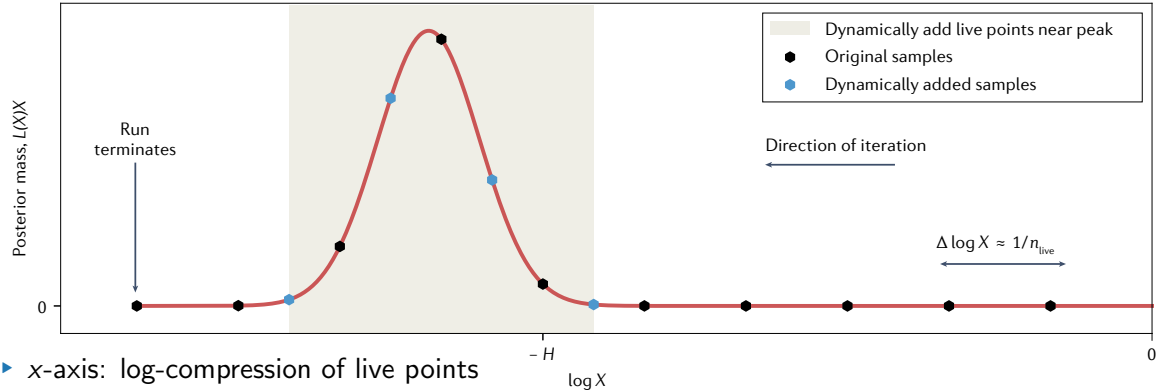
- At the end, one is left with a set of discarded points.
- These may be weighted to form weighted posterior samples using $w_i = \mathcal{L}_i \Delta X_i$.
- They can also be used to calculate the integral $\mathcal{Z} = \sum \mathcal{L}_i \Delta X_i$, or more generally $\sum_i f(\mathcal{L}_i) \Delta X_i$.
  - Nested sampling probabilistically estimates the volume of the parameter space

$$X_i \approx \left(\frac{n}{n+1}\right) X_{i-1} \quad \Rightarrow \quad X_i \approx \left(\frac{n}{n+1}\right)^i \approx e^{-i/n},$$

  - Nested sampling thus estimates the density of states,
  - it is therefore a partition function calculator $Z(\beta) = \sum_i \mathcal{L}_i^\beta \Delta X_i$.
- The evolving ensemble of live points allows algorithms to perform self-tuning and mode clustering.



Time complexity of nested sampling



- x-axis: log-compression of live points
- Area $\propto$ posterior mass
- Shows Bayesian balance of likelihood vs prior
- Run proceeds right to left
- Run finishes after bump (typical set)

- Time complexity

$$T = n_{\text{live}} \times T_{\mathcal{L}} \times T_{\text{sampler}} \times \mathcal{D}_{\text{KL}}(\mathcal{P}||\pi)$$

- Error complexity $\sigma \propto \sqrt{\mathcal{D}_{\text{KL}}(\mathcal{P}||\pi)/n_{\text{live}}}$

Sampling from a hard likelihood constraint

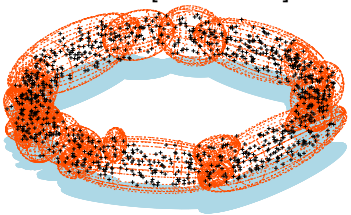
"It is not the purpose of this introductory paper to develop the technology of navigation within such a volume. We merely note that exploring a hard-edged likelihood-constrained domain should prove to be neither more nor less demanding than exploring a likelihood-weighted space."

— John Skilling

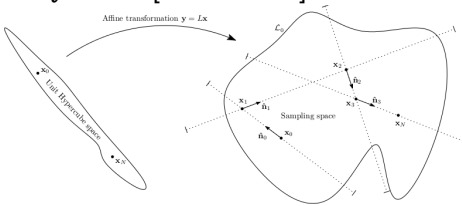
- A large fraction of the work in NS to date has been in attempting to implement a hard-edged sampler in the NS meta-algorithm $\{\theta \sim \pi : \mathcal{L}(\theta) > \mathcal{L}_*\}$.
- `https://projecteuclid.org/euclid.ba/1340370944`.
- There has also been much work beyond this (focus of this talk).

Implementations of Nested Sampling [2205.15570](NatReview)

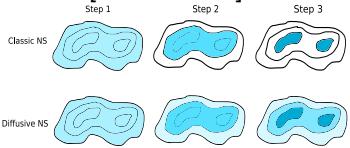
MultiNest [0809.3437]



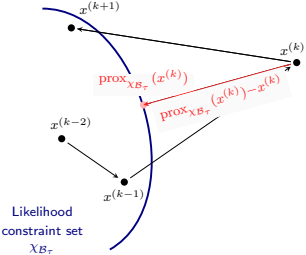
PolyChord [1506.00171]



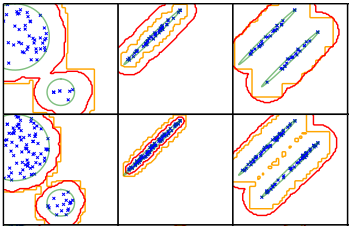
DNest [1606.03757]



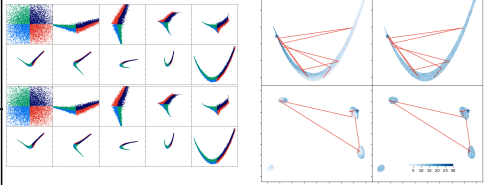
ProxNest [2106.03646]



UltraNest [2101.09604]



NeuralNest [1903.10860]



dynesty [1904.02180]

nessai [2102.11056]

Types of nested sampler

- Broadly, most nested samplers can be split into how they create new live points.
- i.e. how they sample from the hard likelihood constraint $\{\theta \sim \pi : \mathcal{L}(\theta) > \mathcal{L}_*\}$.

Rejection samplers

- e.g. `MultiNest`, `UltraNest`.
- Constructs bounding region and draws many invalid points until $\mathcal{L}(\theta) > \mathcal{L}_*$.
- Efficient in low dimensions, exponentially inefficient $\sim \mathcal{O}(e^{d/d_0})$ in high $d > d_0 \sim 10$.

Chain-based samplers

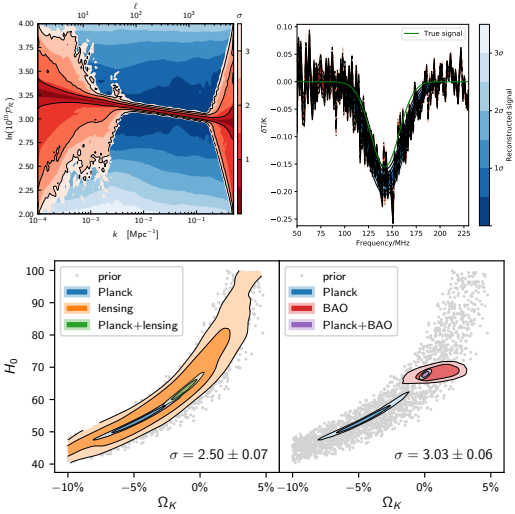
- e.g. `PolyChord`, `ProxNest`.
- Run Markov chain starting at a live point, generating many valid (correlated) points.
- Linear $\sim \mathcal{O}(d)$ penalty in decorrelating new live point from the original seed point.

- Nested samplers usually come with:
  - *resolution* parameter n_{live} (which improve results as $\sim \mathcal{O}(n_{\text{live}}^{-1/2})$.
  - set of *reliability* parameters [2101.04525], which don't improve results if set arbitrarily high, but introduce systematic errors if set too low.
  - e.g. `Multinest` efficiency `eff` or `PolyChord` chain length n_{repeats}.

Applications of nested sampling

Cosmology

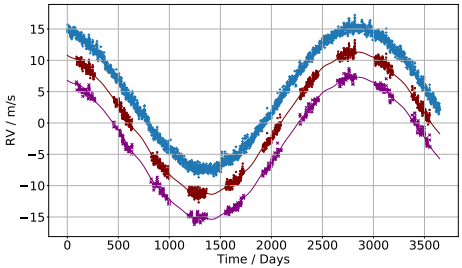
- Battle-tested in Bayesian cosmology on
  - Parameter estimation: multimodal alternative to MCMC samplers.
  - Model comparison: using integration to compute the Bayesian evidence
  - Tension quantification: using deep tail sampling and suspiciousness computations.
- Plays a critical role in major cosmology pipelines: Planck, DES, KiDS, BAO, SNe.
- The default ΛCDM cosmology is well-tuned to have Gaussian-like posteriors for CMB data.
- Less true for alternative cosmologies/models and orthogonal datasets, so nested sampling crucial.



Applications of nested sampling

Astrophysics

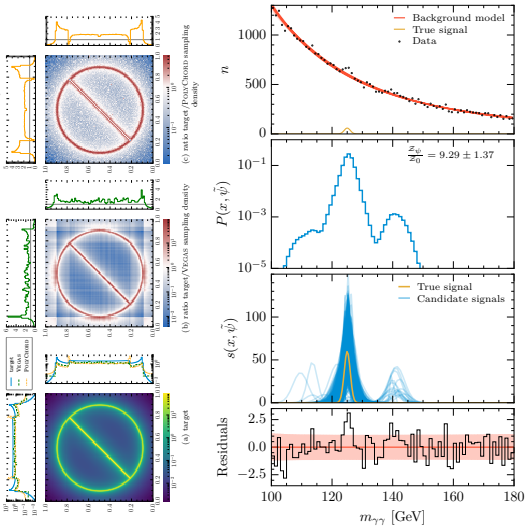
- In exoplanets [1806.00518]
  - Parameter estimation: determining properties of planets.
  - Model comparison: how many planets? Stellar modelling [2007.07278].
  - exoplanet problems regularly have posterior phase transitions [2102.03387]
- In gravitational waves
  - Parameter estimation: Binary merger properties
  - Model comparison: Modified theories of gravity, selecting phenomenological parameterisations [1803.10210]
  - Likelihood reweighting: fast slow properties



Applications of nested sampling

Particle physics

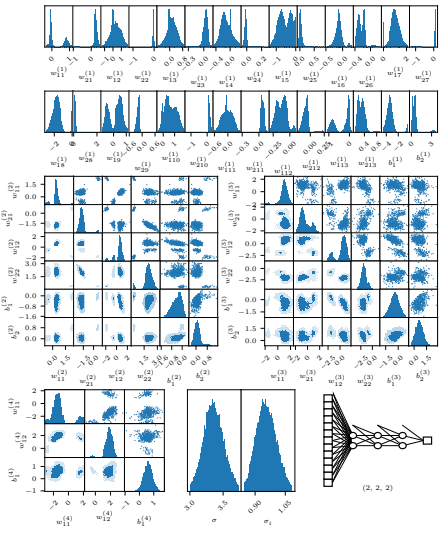
- Nested sampling for cross section computation/event generation $\sigma = \int_\Omega d\Phi |\mathcal{M}|^2.$
- Nested sampling can explore the phase space Ω and compute integral blind with comparable efficiency to HAAG/RAMBO [2205.02030].
- Bayesian sparse reconstruction [1809.04598] applied to bump hunting allows evidence-based detection of signals in phenomenological backgrounds [2211.10391].
- Now applying to lattice field theory, and lattice gravity Lagrangians.
- Fine tuning quantification



Applications of nested sampling

Machine learning

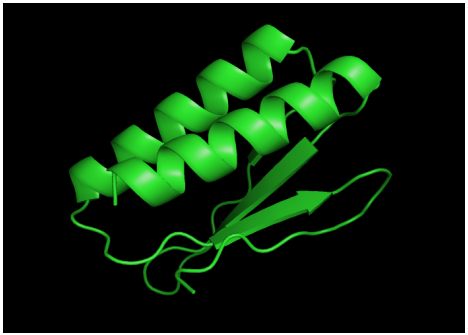
- Machine learning requires:
  - Training to find weights
  - Choice of architecture/topology/hyperparameters
- Bayesian NNs treat training as a model fitting problem
- Compute posterior of weights (parameter estimation), rather than optimisation (gradient descent)
- Use evidence to determine best architecture (model comparison), correlates with out-of-sample performance!
- Solving the full "shallow learning" problem without compromise [2004.12211][2211.10391].
- Promising work ongoing to extend this to transfer learning and deep nets.



Applications of nested sampling

and beyond...

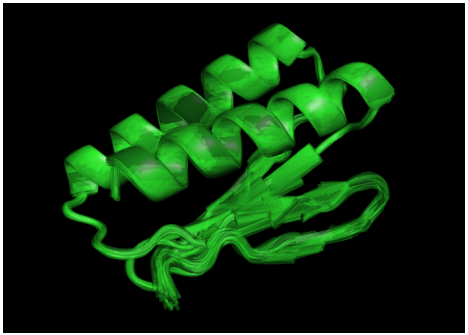
- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.



Applications of nested sampling

and beyond...

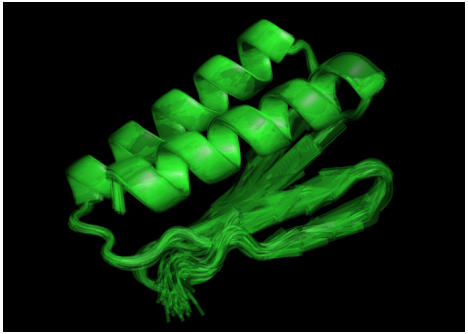
- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.



Applications of nested sampling

and beyond...

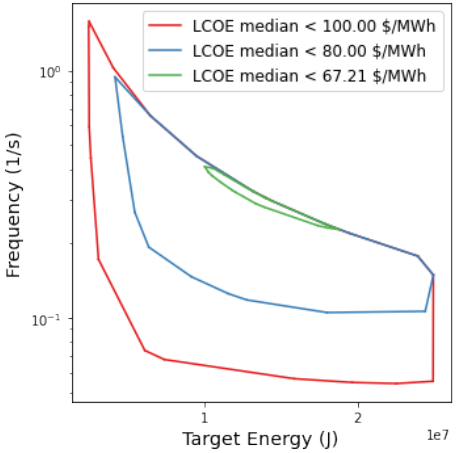
- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.



Applications of nested sampling

and beyond...

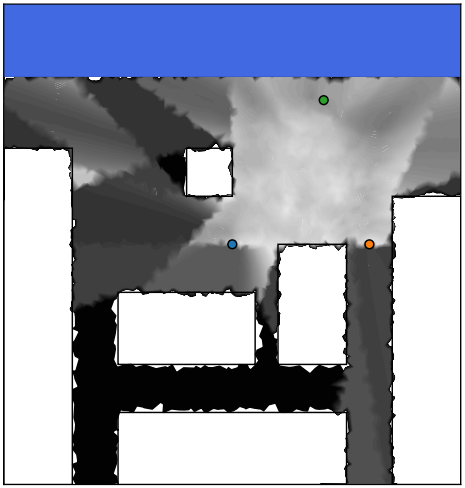
- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.



Applications of nested sampling

and beyond...

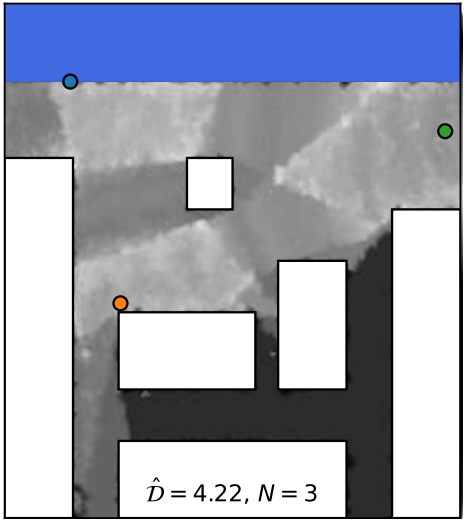
- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.



Applications of nested sampling

and beyond...

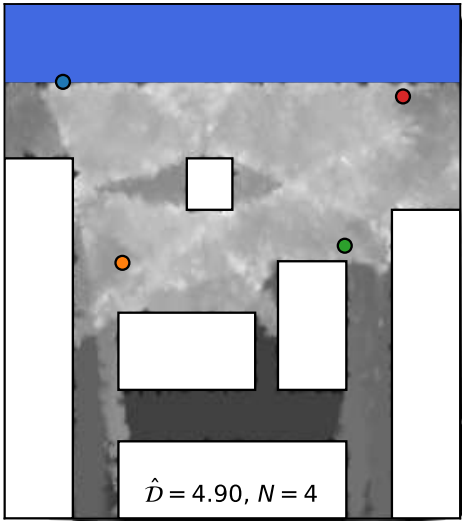
- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.



Applications of nested sampling

and beyond...

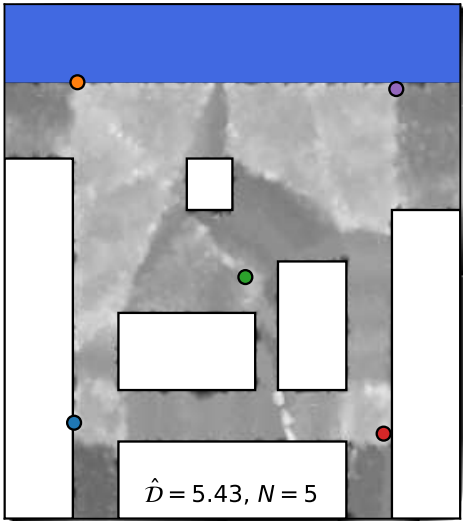
- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.



Applications of nested sampling

and beyond...

- Techniques have been spun-out (PolyChord Ltd) to:
- Protein folding
  - Navigating free energy surface.
  - Computing misfolds.
  - Thermal motion.
- Nuclear fusion reactor optimisation
  - multi-objective.
  - uncertainty propagation.
- Telecoms & DSTL research (MIDAS)
  - Optimising placement of transmitters/sensors.
  - Maximum information data acquisition strategies.

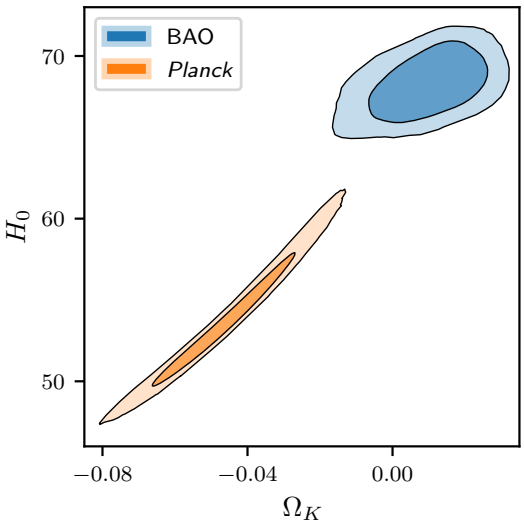


What is a model?

- Model comparison in its purest form answers question such as:
  - "Is the universe ΛCDM?"
  - "Are neutrinos in a normal or inverted hierarchy?"
  - "Is there a detectable global signal in this data?"
- However model $\mathcal{M}$ is likelihood $\mathcal{L} = P(D|\theta, M)$ and priors $\pi = P(\theta|M)$, $\Pi = P(M)$.
- Can use the evidence $\mathcal{Z}$ to decide on which out of a set of likelihoods best describe data (e.g. Gaussian, Cauchy, Poisson, radiometric).
- Can also use it for antenna selection [2106.10193] [2109.10098].
- In principle can use it to decide between theoretically motivated priors (care needed).
- It can also be used for non-parametric reconstruction:
  - "How many polynomial terms best describe the data?"
  - "How complicated a sky model do I need?"
  - "Which is the best sky model?"

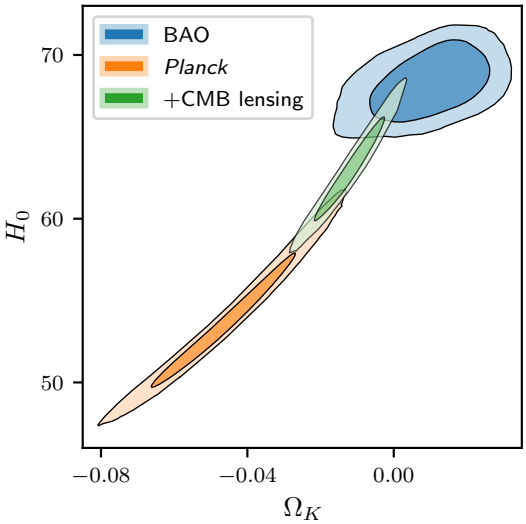
Model comparison and parameter estimation [1908.09139]

- If you allow $\Omega_K \neq 0$, *Planck* (`plikTTTEEE`) has a moderate preference for closed universes (50:1 betting odds on), $\Omega_K = -4.5 \pm 1.5\%$
- *Planck*+lens+BAO strongly prefer $\Omega_K = 0$.
- But, *Planck* vs lensing is 2.5σ in tension, and Planck vs BAO is 3σ.
- Reduced if `plik` $\rightarrow$ `camspec` [2002.06892]
- BAO and lensing summary assume ΛCDM.
- Doing this properly with BAO retains preference for closed universe (though closer to flat $\Omega_K = -0.4 \pm 0.2\%$) [2205.05892].
- Present-day curvature has profound consequences for inflation [2205.07374].



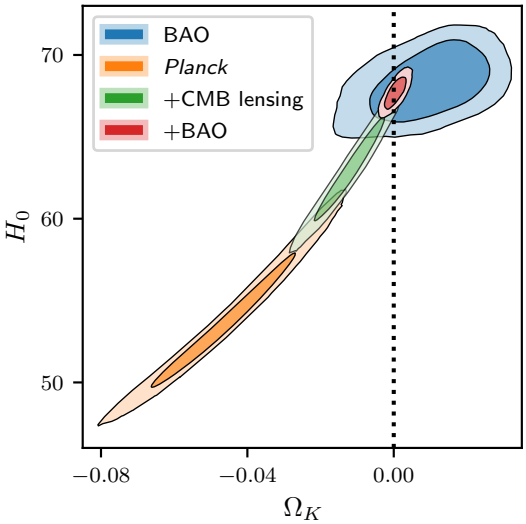
Model comparison and parameter estimation [1908.09139]

- If you allow $\Omega_K \neq 0$, *Planck* (`plikTTTEEE`) has a moderate preference for closed universes (50:1 betting odds on), $\Omega_K = -4.5 \pm 1.5\%$
- *Planck*+lens+BAO strongly prefer $\Omega_K = 0$.
- But, *Planck* vs lensing is 2.5σ in tension, and Planck vs BAO is 3σ.
- Reduced if `plik` $\rightarrow$ `camspec` [2002.06892]
- BAO and lensing summary assume ΛCDM.
- Doing this properly with BAO retains preference for closed universe (though closer to flat $\Omega_K = -0.4 \pm 0.2\%$) [2205.05892].
- Present-day curvature has profound consequences for inflation [2205.07374].



Model comparison and parameter estimation [1908.09139]

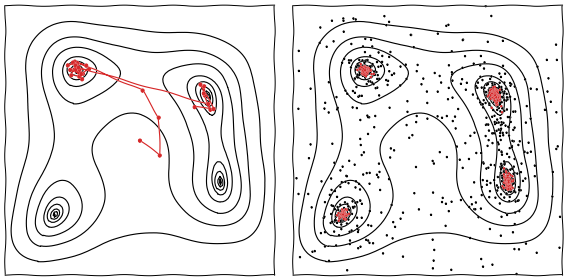
- If you allow $\Omega_K \neq 0$, *Planck* (`plikTTTEEE`) has a moderate preference for closed universes (50:1 betting odds on), $\Omega_K = -4.5 \pm 1.5\%$
- *Planck*+lens+BAO strongly prefer $\Omega_K = 0$.
- But, *Planck* vs lensing is 2.5σ in tension, and Planck vs BAO is 3σ.
- Reduced if `plik` → `camspec` [2002.06892]
- BAO and lensing summary assume ΛCDM.
- Doing this properly with BAO retains preference for closed universe (though closer to flat $\Omega_K = -0.4 \pm 0.2\%$) [2205.05892].
- Present-day curvature has profound consequences for inflation [2205.07374].



`unimpeded`: legacy suites for the next generation

- DiRAC 2020 RAC allocation of 30MCPUh
- Main goal: Planck Legacy Archive equivalent
- Parameter estimation → Model comparison
- MCMC → Nested sampling
- Planck → {Planck, DESY1, BAO, ...}
- Pairwise combinations
- Suite of tools for processing these
  - `anesthetic` 2.0
  - `unimpeded` 1.0
  - `zenodo` archive
  - `margarine`
- MCMC chains also available.
- Library of bijectors emulators for fast re-use

DiRAC

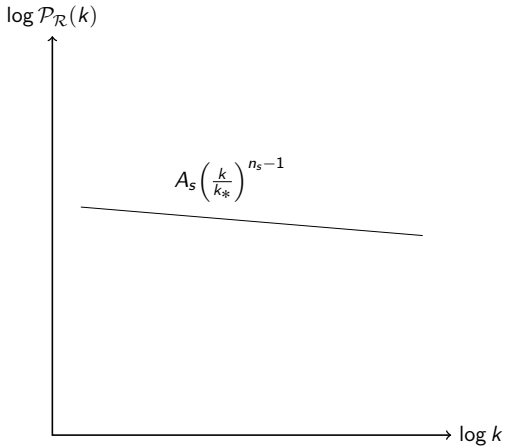


Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction [1908.00906]

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

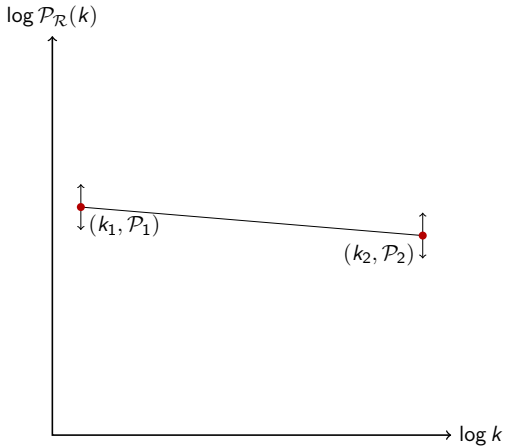


Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction [1908.00906]

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_\mathcal{R}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

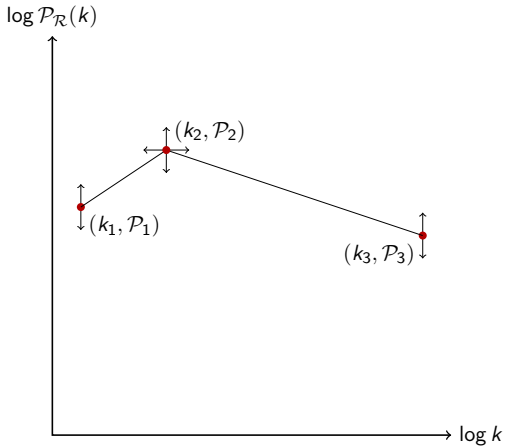


Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction [1908.00906]

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_\mathcal{R}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

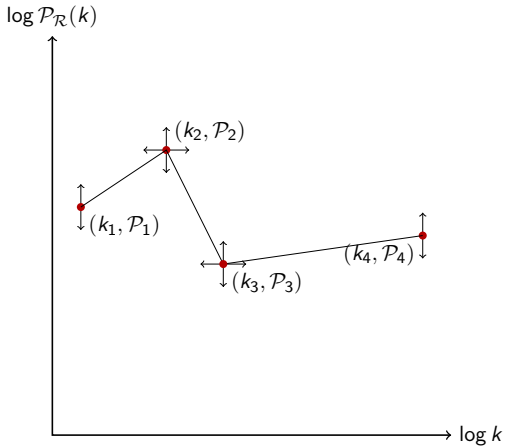


Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction [1908.00906]

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_\mathcal{R}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

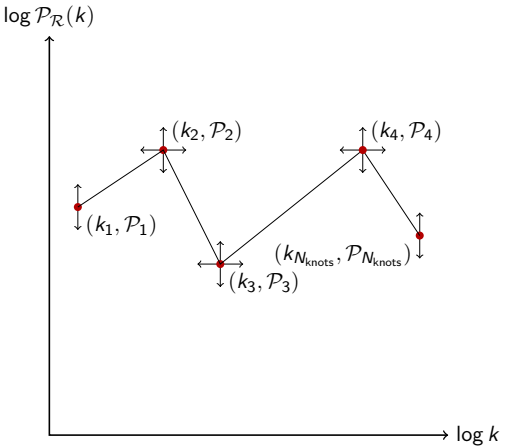


Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction [1908.00906]

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

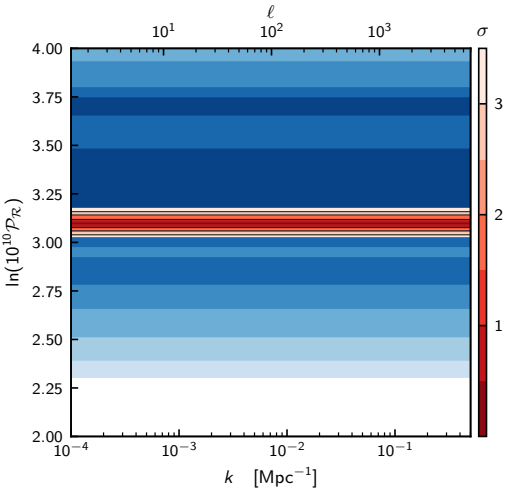


0 internal knots

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

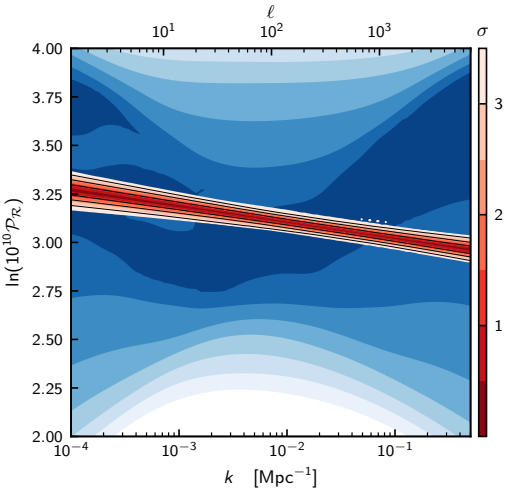


1 internal knot

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

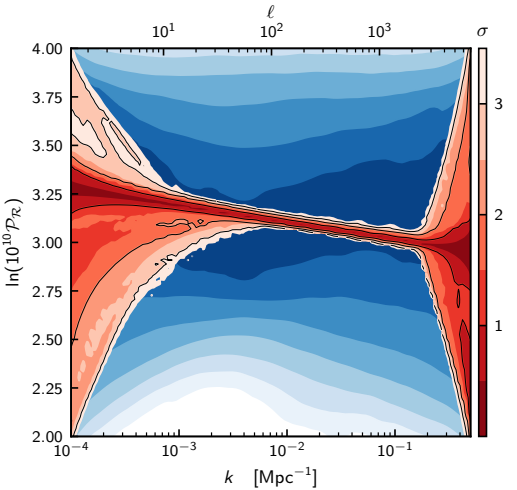


2 internal knots

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

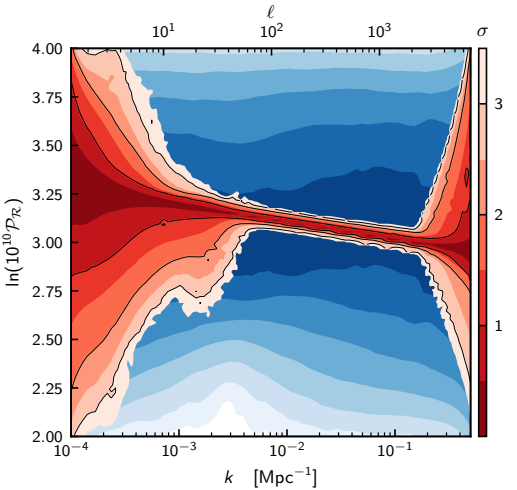


3 internal knots

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

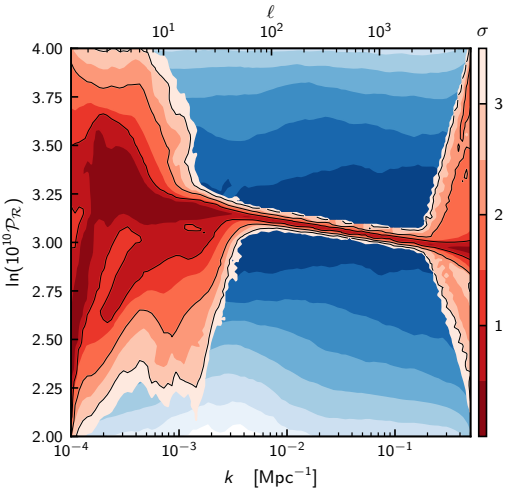


4 internal knots

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

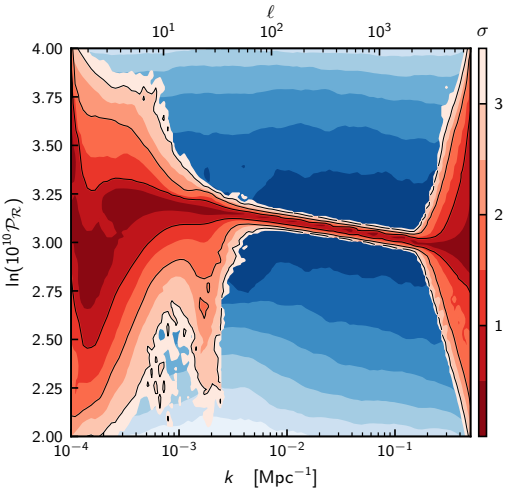


5 internal knots

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

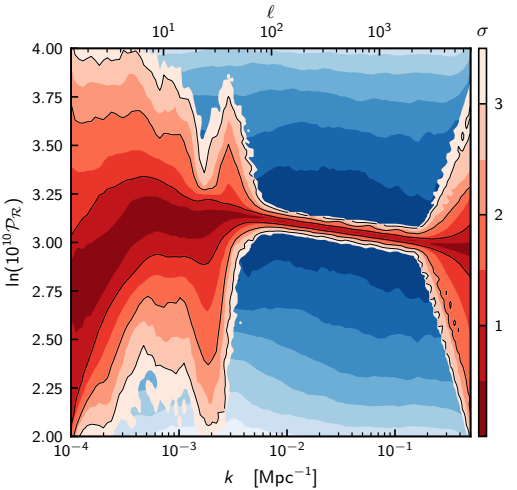


6 internal knots

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

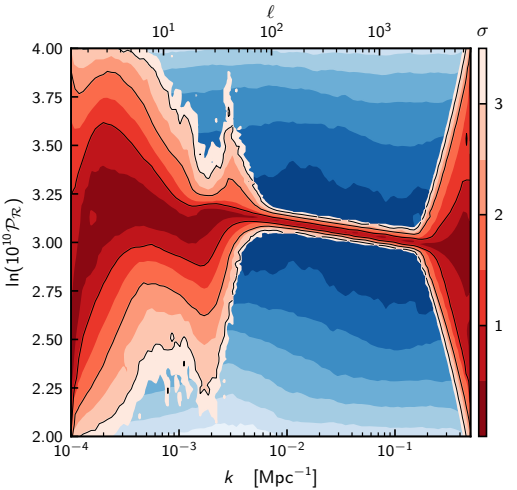


7 internal knots

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

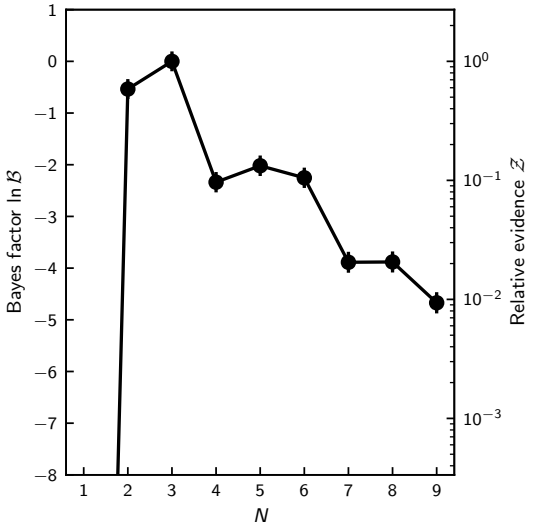


Bayes Factors

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_\mathcal{R}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

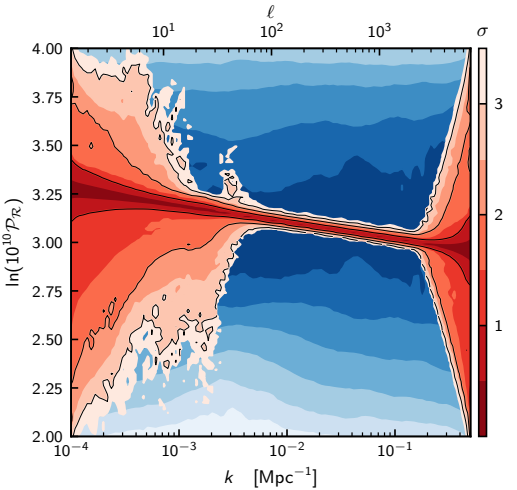


Marginalised plot

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

$$\mathcal{P}_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*} \right)^{n_s - 1}$$

- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

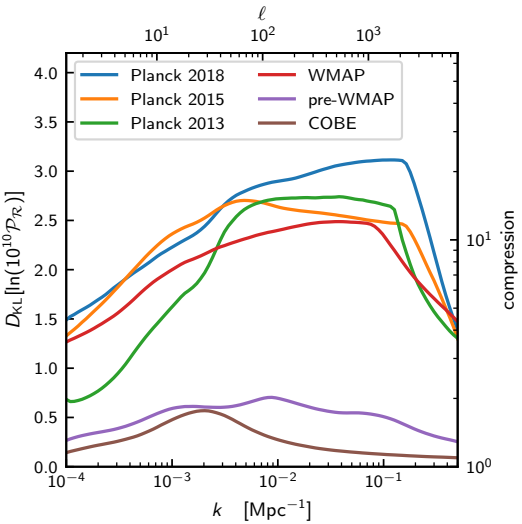


Kullback-Liebler divergences

- Traditionally parameterise the primordial power spectrum with (A_s, n_s)

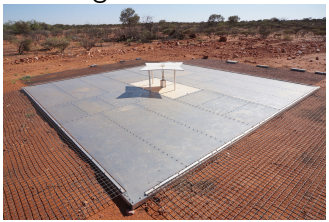
$$\mathcal{P}_\mathcal{R}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1}$$

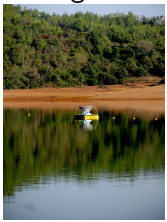
- To add more degrees of freedom, can add "running" parameters n_{run} (higher order polynomial in index)
- Alternative non-parametric technique introduces a more flexible phenomenological parameterisation: "FlexKnots"
- Let the Bayesian evidence decide when you've introduced too many parameters

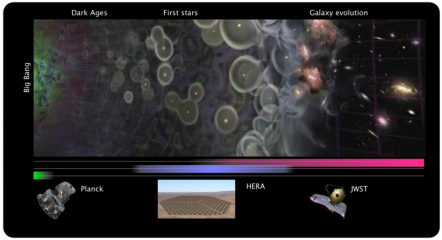


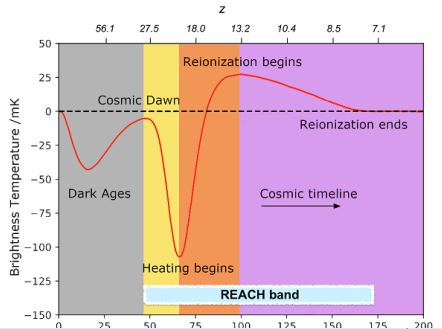
REACH: Global 21cm cosmology [NatAstro]

- Imaging the universal dark ages using CMB backlight.
- 21cm hyperfine line emission from neutral hydrogen.
- Global experiments measure monopole across frequency.
- Gives a specific absorption trough, which if detected allows constraints on the physics of the dark ages decade(s) before SKA.
- Challenge: science hidden in foregrounds $\sim 10^4 \times$signal.



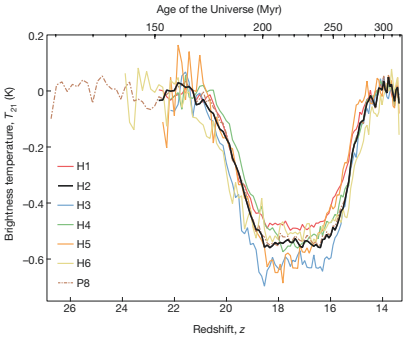






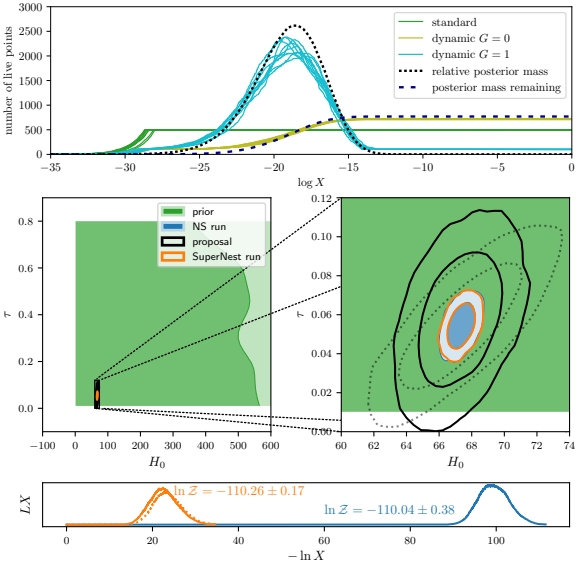
REACH: Global 21cm cosmology [NatAstro]

- EDGES [Nat] claimed a controversial 2019 detection.
- SARAS3 [NatAstro] would have detected this by 2021.
- REACH [NatAstro] aims to settle the debate.
  - Broader band,
  - Honesty about systematic modelling,
  - State of the art inference.
- Create parameterised models of sky, beam and signal, breaking degeneracy with a time-dependent likelihood to measure all three simultaneously.
- Use model comparison based reconstruction to determine complexity of parameterisation.
- Use model comparison to select likelihoods.
- A collaboration powered by nested sampling.



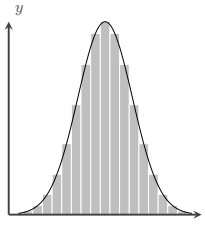
Beyond the meta-algorithm

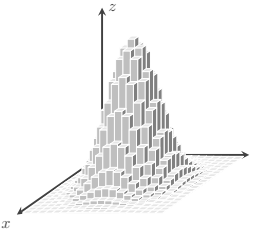
- Dynamic nested sampling [1704.03459]
- Unwoven nested sampling [1703.09701]
- Accelerated nested sampling [2212.01760]
- Precision nested sampling [2006.03371]
- Multiobjective nested sampling
- Nested sampling with gradients?
- Reversible nested sampling?
- Transdimensional nested sampling?
- postprocessing: `anesthetic` [1905.04768]
- crosschecking: `nestcheck` [1804.06406]
- See "Frontiers of nested sampling" talk from last year: willhandley.co.uk/talks

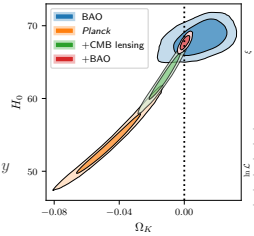


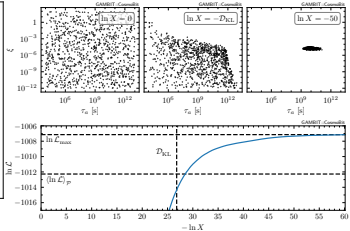
Conclusions

- Nested sampling is a multi-purpose numerical tool for:
  - Numerical integration $\int f(x)dV$,
  - Exploring/scanning/optimising *a priori* unknown functions,
  - Performing Bayesian inference and model comparison.
- It is applied widely across cosmology and particle physics.
- It can be applied to both Bayesian and Frequentist inference.









How does Nested Sampling compare to other approaches?

- In all cases:
  - \+ NS can handle multimodal functions
  - \+ NS computes evidences, partition functions and integrals
  - \+ NS is self-tuning/black-box

Modern Nested Sampling algorithms can do this in $\sim \mathcal{O}(100s)$ dimensions

Optimisation

- Gradient descent
  - – NS cannot use gradients
  - \+ NS does not require gradients
- Genetic algorithms
  - \+ NS discarded points have statistical meaning

Sampling

- Metropolis-Hastings?
  - – Nothing beats well-tuned customised MH
  - \+ NS is self tuning
- Hamiltonian Monte Carlo?
  - – In millions of dimensions, HMC is king
  - \+ NS does not require gradients

Integration

- Thermodynamic integration
  - \+ protective against phase trasitions
  - \+ No annealing schedule tuning
- Sequential Monte Carlo
  - – SMC experts classify NS as a kind of SMC
  - \+ NS is athermal

Nested Sampling: a user's guide

1. Nested sampling is a likelihood scanner, rather than posterior explorer.
   - This means typically most of its time is spent on burn-in rather than posterior sampling.
   - Changing the stopping criterion from 10^{-3} to 0.5 does little to speed up the run, but can make results very unreliable.
2. The number of live points n_{live} is a resolution parameter.
   - Run time is linear in n_{live}, posterior and evidence accuracy goes as $\frac{1}{\sqrt{n_{\text{live}}}}$.
   - Set low for exploratory runs $\sim \mathcal{O}(10)$ and increased to $\sim \mathcal{O}(1000)$ for production standard.
3. Most algorithms come with additional reliability parameter(s).
   - e.g. `MultiNest`: `eff`, `PolyChord`: n_{repeats}.
   - These are parameters which have no gain if set too conservatively, but increase the reliability.
   - Check that results do not degrade if you reduce them from defaults, otherwise increase.

Occam's Razor [2102.11511]

- Bayesian inference quantifies Occam's Razor:
  - *"Entities are not to be multiplied without necessity"* — William of Occam
  - *"Everything should be kept as simple as possible, but not simpler"* — "Albert Einstein"
- Properties of the evidence: rearrange Bayes' theorem for parameter estimation

$$\mathcal{P}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{\mathcal{Z}} \qquad \Rightarrow \qquad \log \mathcal{Z} = \log \mathcal{L}(\theta) - \log \frac{\mathcal{P}(\theta)}{\pi(\theta)}.$$

- Evidence is composed of a "goodness of fit" term and "Occam Penalty".

- RHS true for all θ. Take max likelihood value θ_*:

$$\log \mathcal{Z} = -\chi^2_{\text{min}} - \text{Mackay penalty}.$$

- Be more Bayesian and take posterior average to get the "Occam's razor equation"

$$\boxed{\log \mathcal{Z} = \langle \log \mathcal{L} \rangle_{\mathcal{P}} - \mathcal{D}_{\text{KL}}.}$$

- Natural regularisation which penalises models with too many parameters.

Kullback Liebler divergence

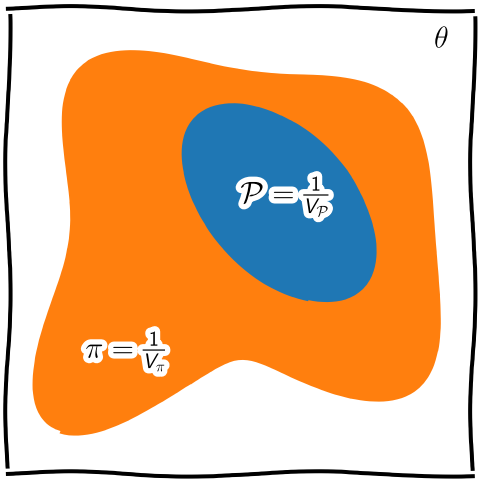
- The KL divergence between prior π and posterior $\mathcal{P}$ is is defined as:

$$\mathcal{D}_{\mathrm{KL}} = \left\langle \log \frac{\mathcal{P}}{\pi} \right\rangle_{\mathcal{P}} = \int \mathcal{P}(\theta) \log \frac{\mathcal{P}(\theta)}{\pi(\theta)} d\theta.$$

- Whilst not a distance, $\mathcal{D} = 0$ when $\mathcal{P} = \pi$.
- Occurs in the context of machine learning as an objective function for training functions.
- In Bayesian inference it can be understood as a log-ratio of "volumes":

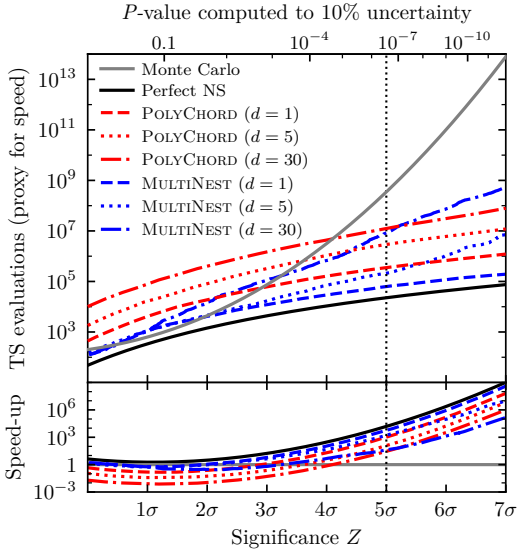
$$\mathcal{D}_{\mathrm{KL}} \approx \log \frac{V_\pi}{V_\mathrm{P}}.$$

(this is exact for top-hat distributions).



Statistics: fast estimation of small p-values [2106.02056](PRL)

- Nested sampling for frequentist computation!?
- p-value: $P(\lambda > \lambda^*|H_0)$ – probability that test statistic λ is at least as great as observed λ^*.
- Computation of a tail probability from sampling distribution of λ under H_0.
- For gold-standard 5σ, this is very expensive to simulate directly ($\sim 10^9$ by definition).
- Need insight/approximation to make efficient.
- Nested sampling is tailor-made for this, just make switch: $X \leftrightarrow p$, $\mathcal{L} \leftrightarrow \lambda$, $\theta \leftrightarrow x$.
- The only real conceptual shift is switching the integrator from parameter- to data-space.

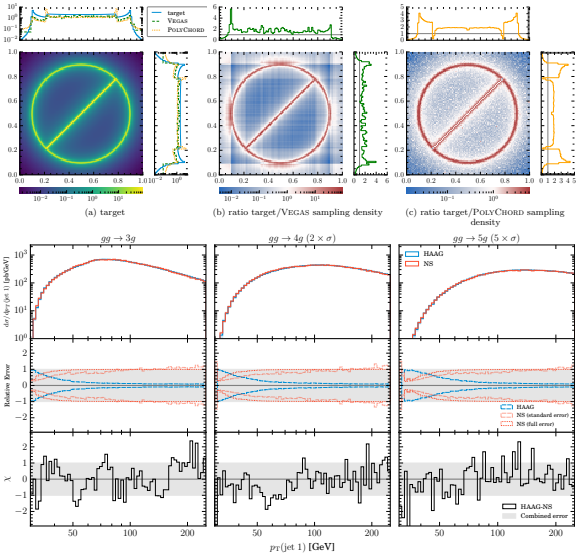


Exploration of phase space [2106.02056]

- Nested sampling for cross section computation/event generation.
- Numerically compute collisional cross section

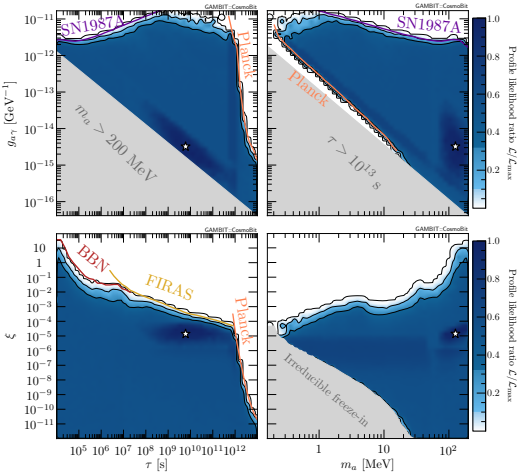
$$\sigma = \int_\Omega d\Phi |\mathcal{M}|^2,$$

  Ω phase space of kinematic configurations Φ, each with matrix element $\mathcal{M}(\Phi)$.
- Current state of the art e.g. HAAG (improvement on RAMBO) requires knowledge of $\mathcal{M}(\Phi)$.
- Nested sampling can explore the phase space and compute integral blind with comparable efficiency.



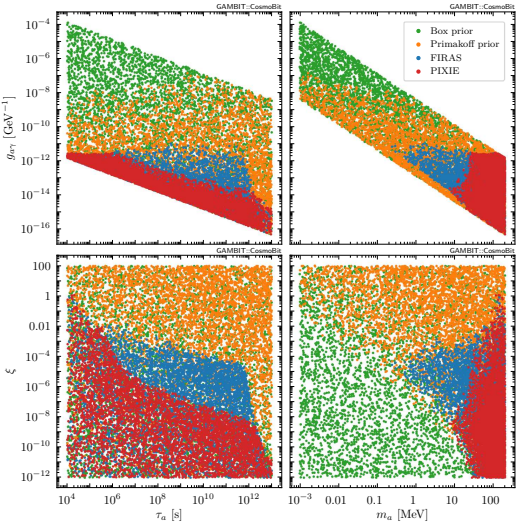
Quantification of fine tuning [2101.00428] [2205.13549]

- Example: Cosmological constraints on decaying axion-like particles [2205.13549].
- Subset of parameters $\xi, m_a, \tau, g_{a\gamma}$: ALP fraction, mass, lifetime and photon coupling. (Also vary cosmology, τ_n and nuisance params)
- Data: CMB, BBN, FIRAS, SMM, BAO.
- Standard profile likelihood fit shows ruled out regions and best-fit point.



Quantification of fine tuning [2101.00428] [2205.13549]

- Example: Cosmological constraints on decaying axion-like particles [2205.13549].
- Subset of parameters $\xi, m_a, \tau, g_{a\gamma}$: ALP fraction, mass, lifetime and photon coupling. (Also vary cosmology, τ_n and nuisance params)
- Data: CMB, BBN, FIRAS, SMM, BAO.
- Standard profile likelihood fit shows ruled out regions and best-fit point.
- Nested sampling scan:
  - Quantifies amount of parameter space ruled out with Kullback-Liebler divergence $\mathcal{D}_{\mathrm{KL}}$.
  - Identifies best fit region as statistically irrelevant from information theory/Bayesian.
  - No evidence for decaying ALPs. Fit the data equally well: but more constrained parameters create Occam penalty.



Quantification of fine tuning [2101.00428] [2205.13549]

- Example: Cosmological constraints on decaying axion-like particles [2205.13549].
- Subset of parameters $\xi, m_a, \tau, g_{a\gamma}$: ALP fraction, mass, lifetime and photon coupling. (Also vary cosmology, τ_n and nuisance params)
- Data: CMB, BBN, FIRAS, SMM, BAO.
- Standard profile likelihood fit shows ruled out regions and best-fit point.
- Nested sampling scan:
  - Quantifies amount of parameter space ruled out with Kullback-Liebler divergence $\mathcal{D}_{\mathrm{KL}}$.
  - Identifies best fit region as statistically irrelevant from information theory/Bayesian.
  - No evidence for decaying ALPs. Fit the data equally well: but more constrained parameters create Occam penalty.