Next generation cosmological analysis with nested sampling

Will Handley
<wh260@cam.ac.uk>

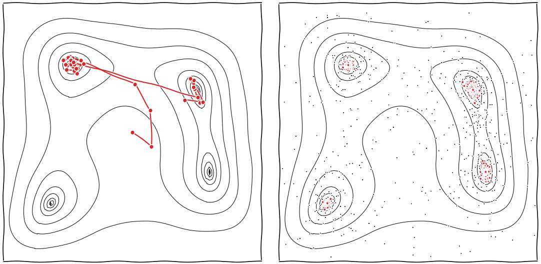
Royal Society University Research Fellow & Turing Fellow
Astrophysics Group, Cavendish Laboratory, University of Cambridge
Kavli Institute for Cosmology, Cambridge
Gonville & Caius College
github.com/williamjameshandley/talks
willhandley.co.uk

8th September 2022

THE ROYAL SOCIETY
The Alan Turing Institute
KICC Kavli Institute for Cosmology Cambridge
UNIVERSITY OF CAMBRIDGE
DiRAC

Overview

- DiRAC 2020 RAC allocation of 30MCPUh
- Main goal: Planck Legacy Archive equivalent
- Parameter estimation $\to$ Model comparison
- MCMC $\to$ Nested sampling
- Planck $\to$ $\{\text{Planck}, \text{DESY1}, \text{BAO}, \ldots\}$
- Pairwise combinations
- Suite of tools for processing these
  - `anesthetic` 2.0
  - `unimpeded` 1.0
  - `zenodo` archive
- MCMC chains also available.
- Work in progress, but beta testers requested (email wh260@cam.ac.uk)



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

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

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

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

Tension quantification

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

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

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

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_*:

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

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

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

- Natural regularisation which penalises models with too many parameters.

Kullback Liebler divergence

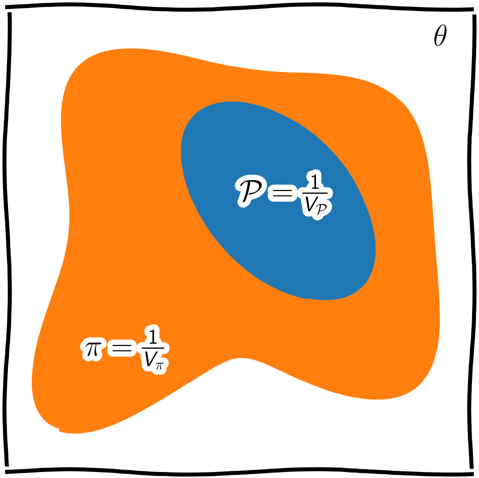
- The KL divergence between prior π and posterior $\mathcal{P}$ is is defined as:

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

- 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":

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

(this is exact for top-hat distributions).

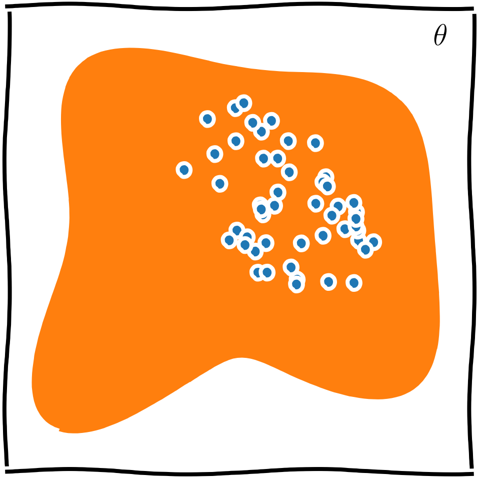


Why do sampling?

- The cornerstone of numerical Bayesian inference is working with **samples**.
- Generate a set of representative parameters drawn in proportion to 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

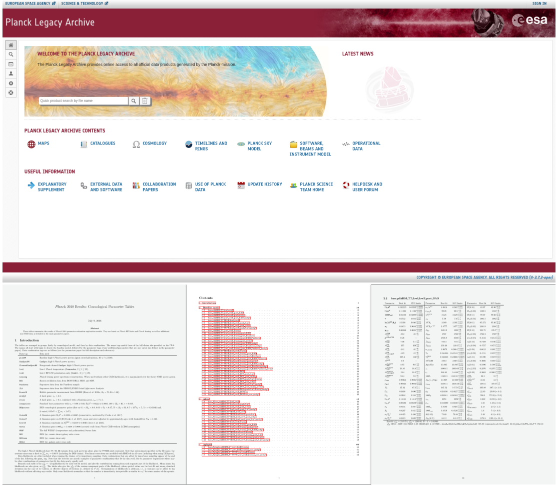
$$\boxed{f(\,\langle X\rangle\,) \ne \langle\, f(X)\,\rangle}$$

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

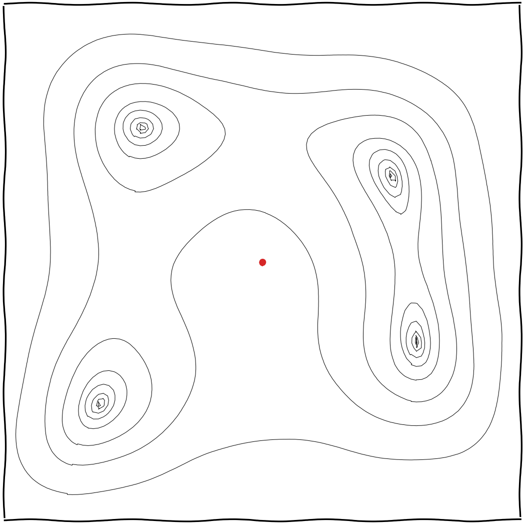


The Planck legacy archive

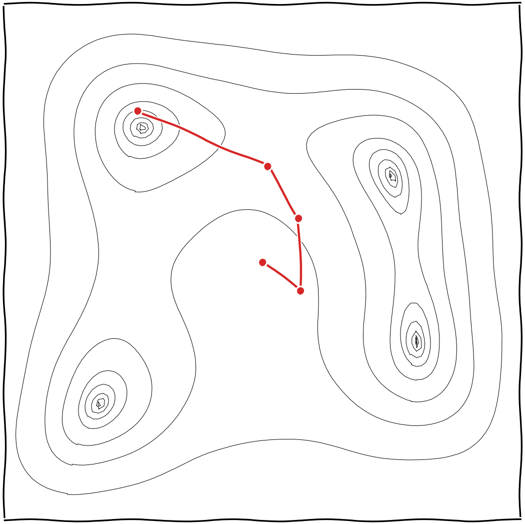
- *Planck* collaboration science products
- distributed cosmology inference results as MCMC chains
- Across a grid of:
  - subsets/combinations of *Planck* data
    - TT, lowl, lowE, lensing
  - ΛCDM extensions
    - base, mnu, nrun, omegak, r
- importance sampling across some other likelihoods (BAO, JLA,. . .)
- Cannot compute evidences in high dimensions from MCMC chains
  - Only parameter estimation
  - no model comparison



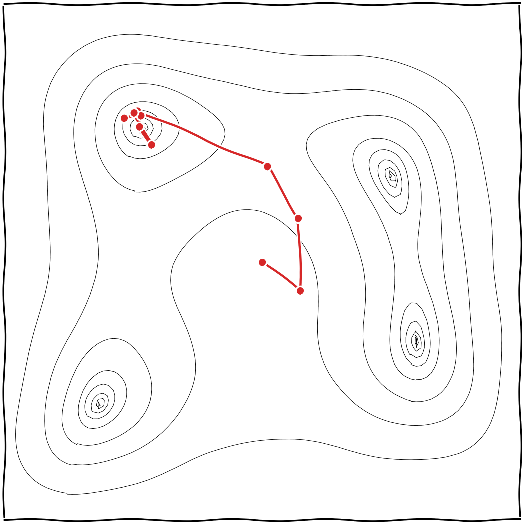
MCMC



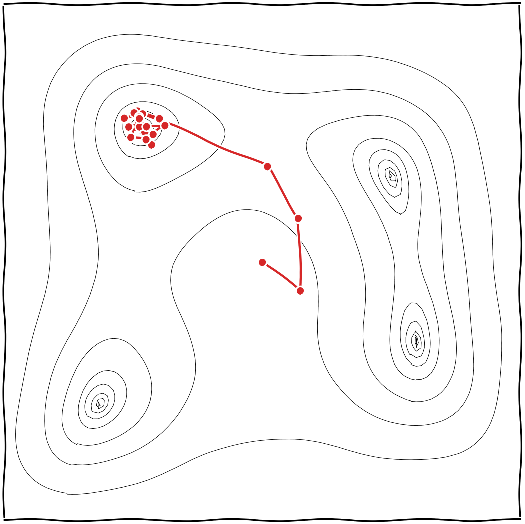
MCMC



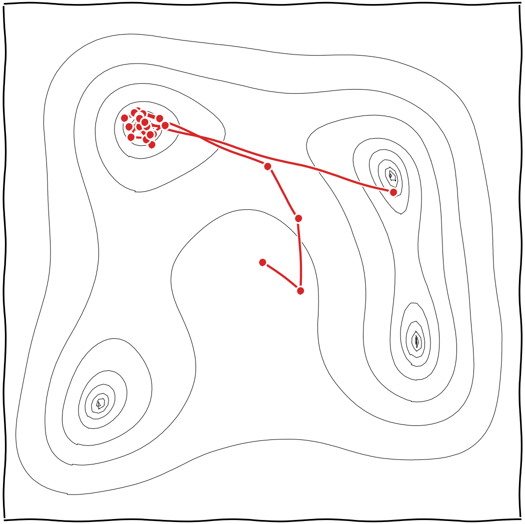
MCMC



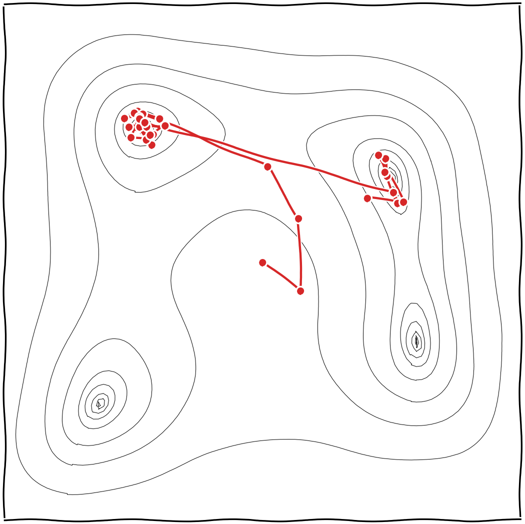
MCMC



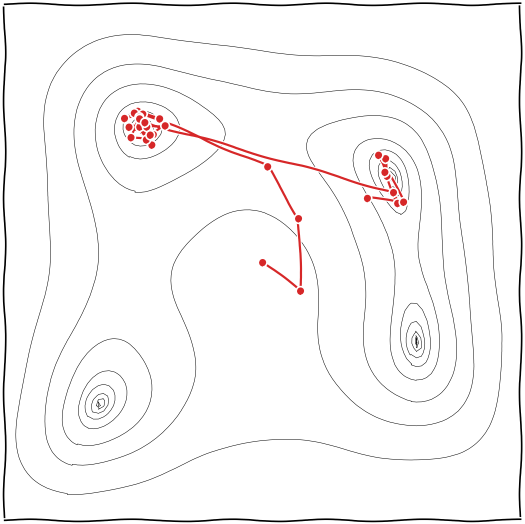
MCMC



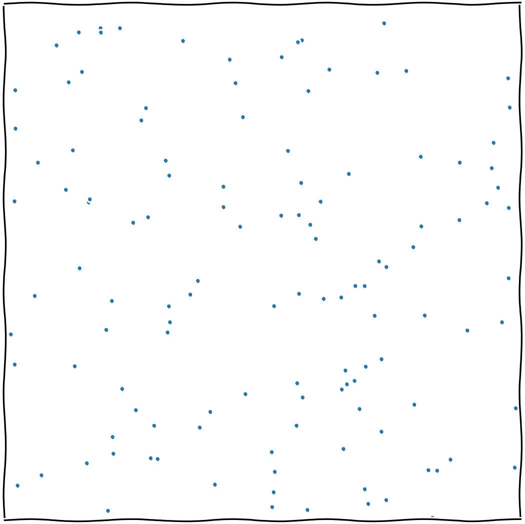
MCMC



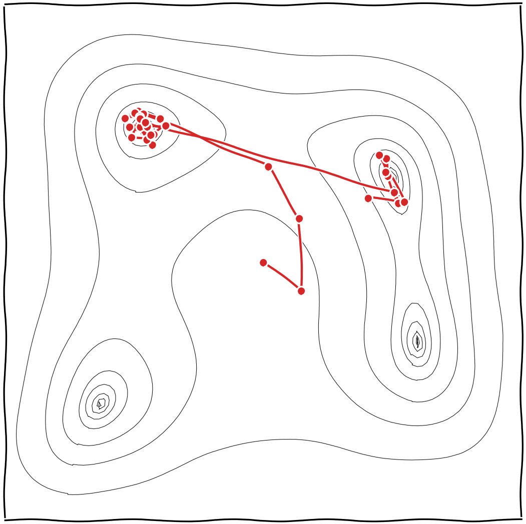
MCMC



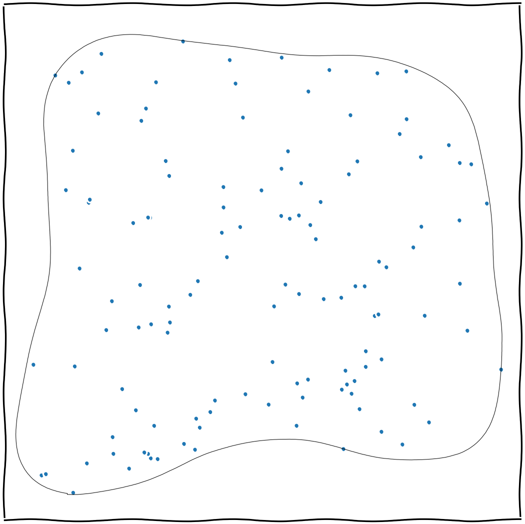
Nested sampling



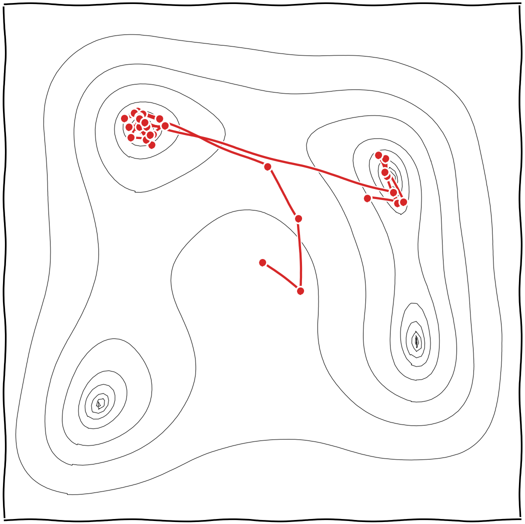
MCMC



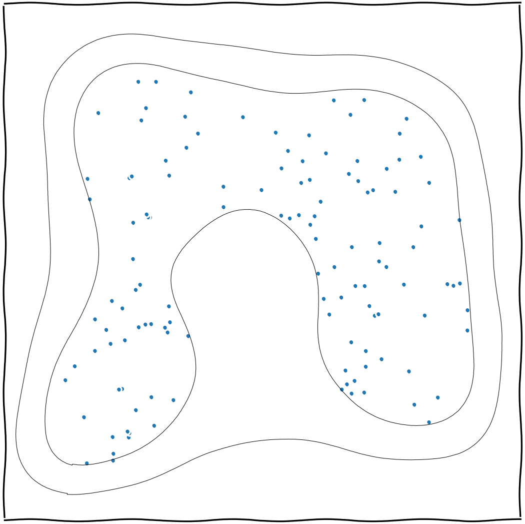
Nested sampling



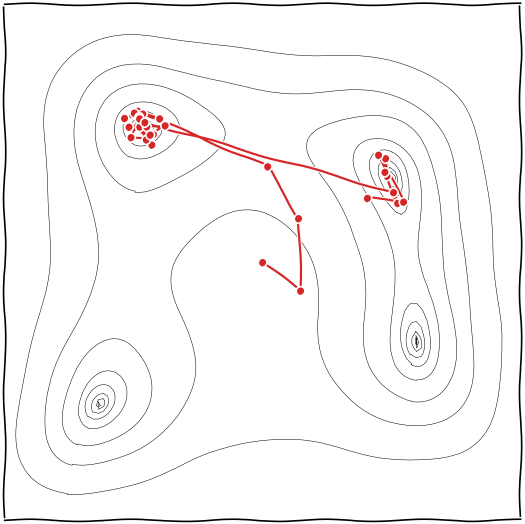
MCMC



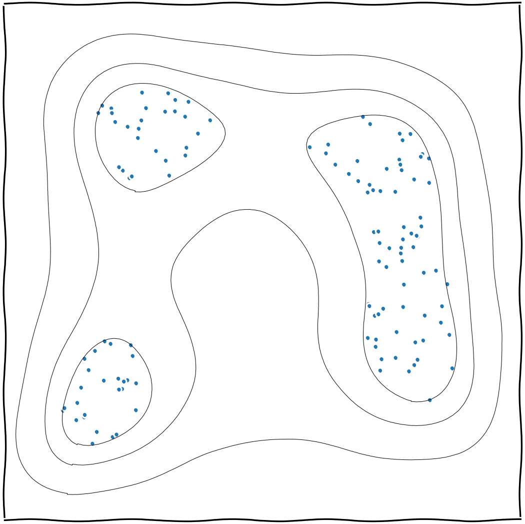
Nested sampling



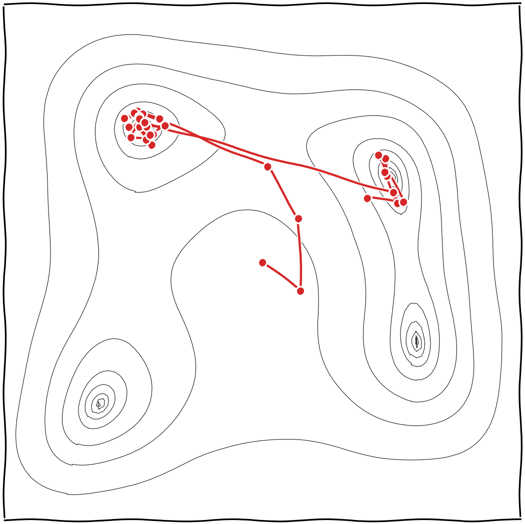
MCMC



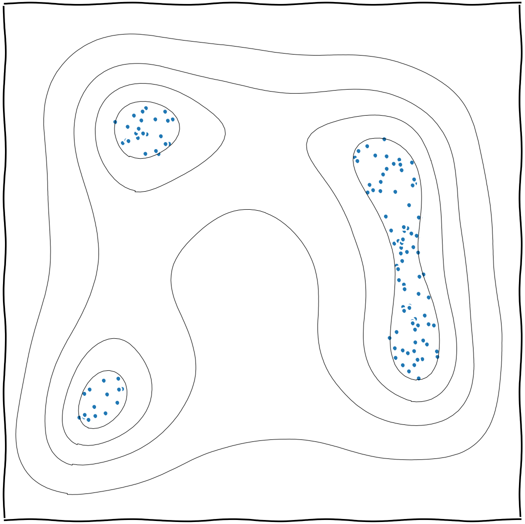
Nested sampling



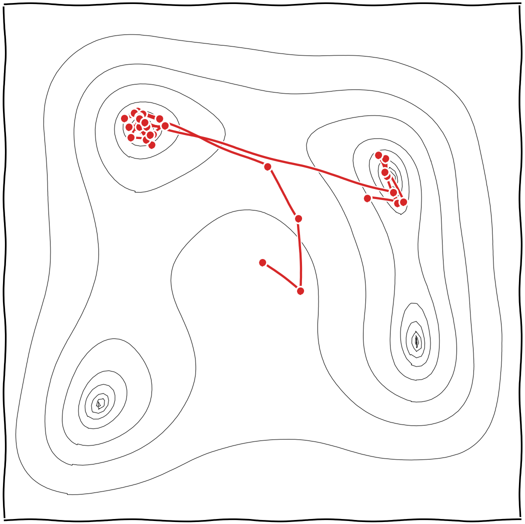
MCMC



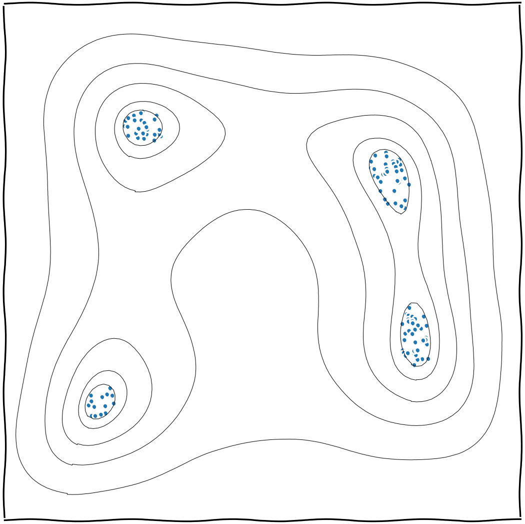
Nested sampling



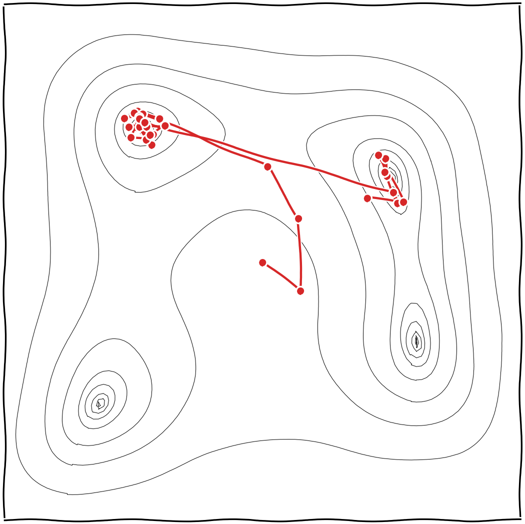
MCMC



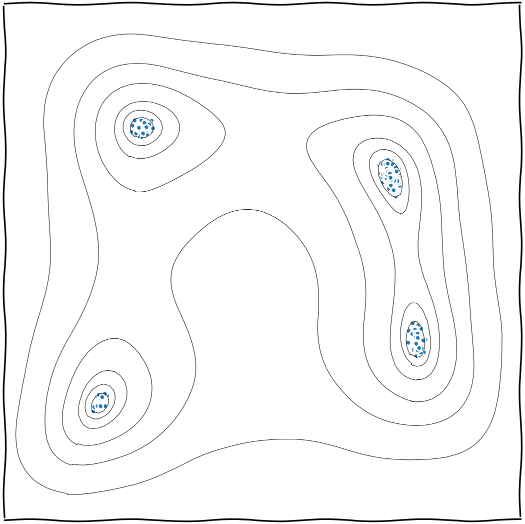
Nested sampling



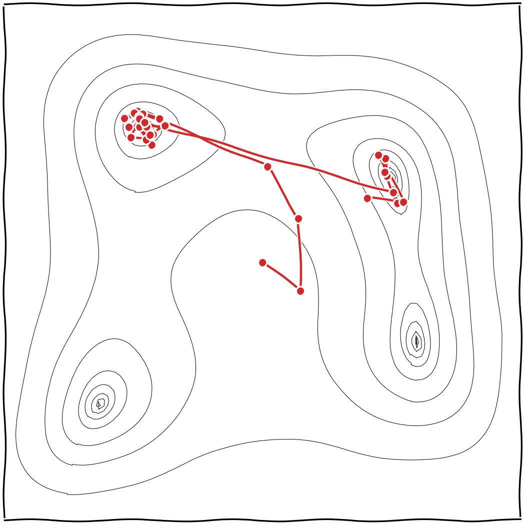
MCMC



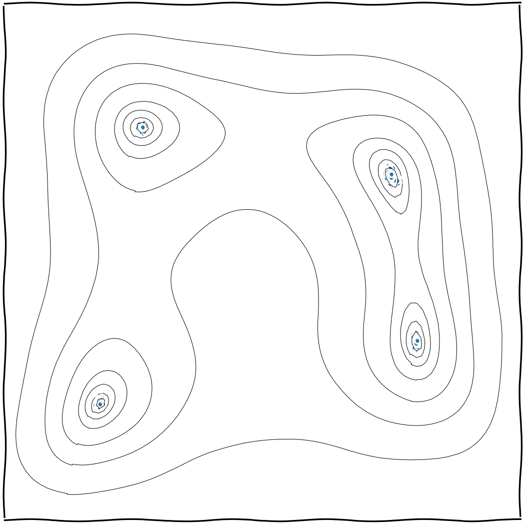
Nested sampling



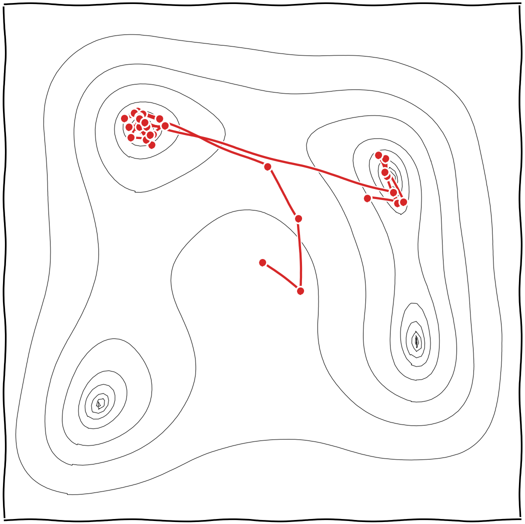
MCMC



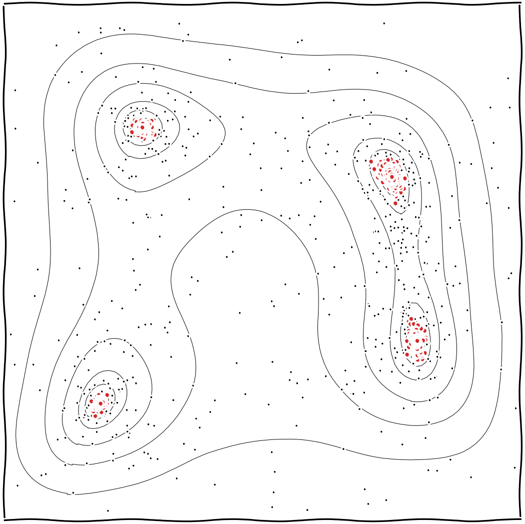
Nested sampling



MCMC

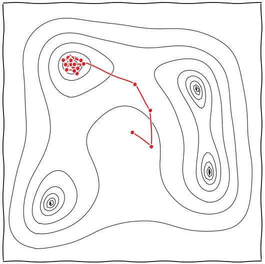


Nested sampling



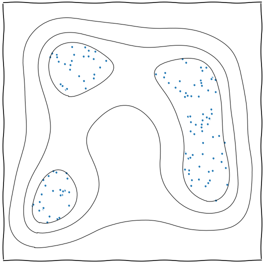
MCMC

- 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



The grid (so far)

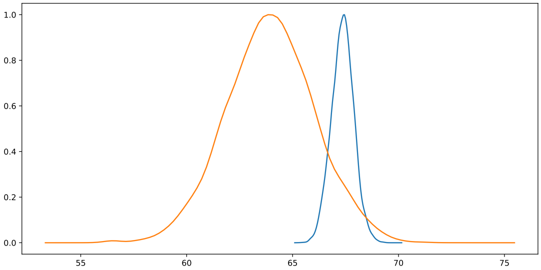
- Models: $[\Lambda\text{CDM}, \Omega_K, \nu, r, w, w(a)]$
- Data: [`plik`, `camspec`, `DESY1`, `bicep+keck`, `BAO(DR16)`, `pantheon`]
- Pairwise combinations of datasets
- Breakdown of Planck & BAO data
- Samplers: [Metropolis Hastings MCMC, Nested Sampling]
- These exhaust what is currently available by default in `cobaya`
- Wide priors to allow for importance readjustment as desired
- roughly halfway through computational allocation.
- Feedback desirable as to what extensions to the grid would be of community interest (email wh260@cam.ac.uk).
- Further checking needed before first release by end of this year.

unimpeded

Universal Model comparison and Parameter Estimation Distributed over Every Dataset

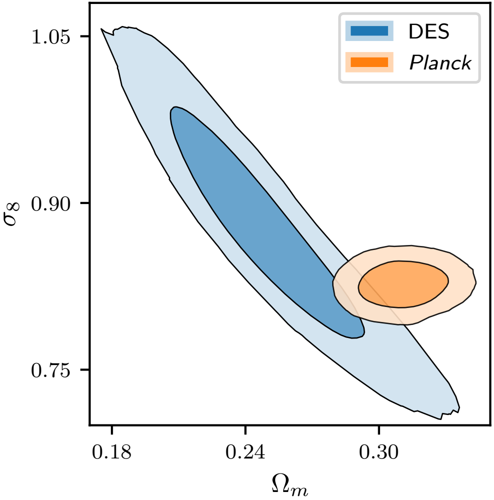
- Python tool for seamlessly downloading and cacheing chains
- Data stored on zenodo
- hdf5 storage for fast & reliable download & storage
- Library of trained bijectors to be used as priors/emulators [2102.12478]/nuisance marginalised likelihoods [2207.11457]
- anesthetic compatible for processing of chains [1905.04768]
- α-testers wanted! (email wh260@cam.ac.uk)
- End goal – community library which everyone contributes to so expensive runs reusable.

```python
from unimpeded import Unimpeded
store = Unimpeded(cache='data.hdf5')
samps = store('planck')
samps.H0.plot.kde_1d()
samps = store('planck', model='klcdm')
samps.H0.plot.kde_1d()
```



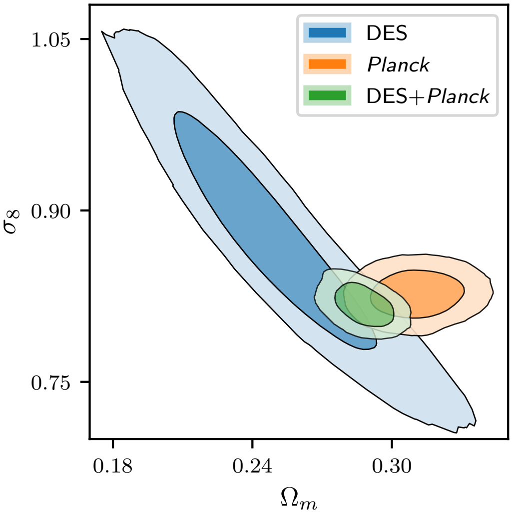
The importance of global measures of tension

- Hubble tension [1907.10625]
  - *Planck*: $H_0 = 67.4 \pm 0.5$
  - SH_0ES: $H_0 = 74.0 \pm 1.4$
- In other situations the discrepancy doesn't exist in a single interpretable parameter
- For example: DES+*Planck* [1902.04029]
- Are these two datasets in tension?
- There are a lot more parameters – are we sure that tensions aren't hiding? Are we sure we've chosen the best ones to reveal the tension?
- Should use "Suspiciousness" statistic $\mathcal{S}$, or Bayes ratio $\mathcal{R}$ to determine global tension.



The importance of global measures of tension

- Hubble tension [1907.10625]
  - *Planck*: $H_0 = 67.4 \pm 0.5$
  - SH_0ES: $H_0 = 74.0 \pm 1.4$
- In other situations the discrepancy doesn't exist in a single interpretable parameter
- For example: DES+*Planck* [1902.04029]
- Are these two datasets in tension?
- There are a lot more parameters – are we sure that tensions aren't hiding? Are we sure we've chosen the best ones to reveal the tension?
- Should use "Suspiciousness" statistic $\mathcal{S}$, or Bayes ratio $\mathcal{R}$ to determine global tension.



The DES evidence ratio R

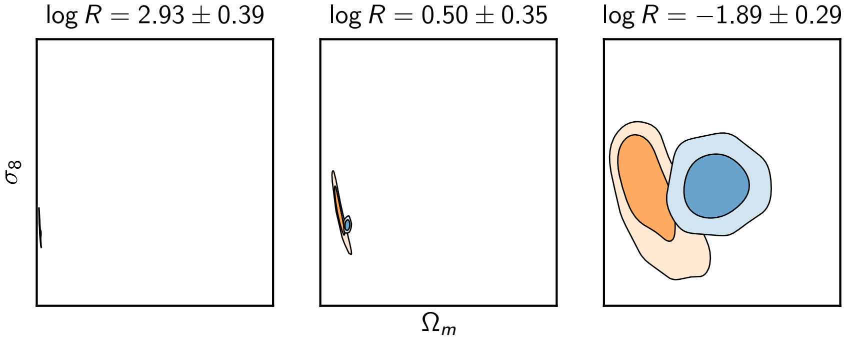
- The Dark Energy Survey [1708.01530] quantifies tension between two datasets A and B using the Bayes ratio:

$$R = \frac{\mathcal{Z}_{AB}}{\mathcal{Z}_A \mathcal{Z}_B} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)}$$

  where $\mathcal{Z}$ is the Bayesian evidence.

- Many attractive properties:
  - Symmetry
  - Parameterisation independence
  - Dimensional consistency
  - Use of well-defined Bayesian quantities
- R gives the relative change in our confidence in data A in light of having seen B (and vice-versa).
- $R > 1$ implies we have more confidence in A having received B.
- Like evidences, it is prior-dependent from $\mathcal{D}$ in $\log \mathcal{Z} = \langle \log \mathcal{L} \rangle_{\mathcal{P}} - \mathcal{D}$
- Increasing prior widths $\Rightarrow$ decreasing evidence.
- Increasing prior widths $\Rightarrow$ increasing confidence.

The DES evidence ratio R: Prior dependency



- What does it mean if increasing prior widths $\Rightarrow$ increasing confidence?
- Wide priors mean *a-priori* the parameters could land anywhere.
- We should be proportionally more reassured when they land close to one another if the priors are wide

How do we deal with the prior dependency in R?

Option 1 Take the Bayesian route, accept the prior dependency, and spend time trying to justify why a given set of priors are "physical".

Option 2 Try to find a principled way of removing this prior dependency

- Decompose ratio using Occam's Razor equation $\log\mathcal{Z} = \langle\log\mathcal{L}\rangle_\mathcal{P} - \mathcal{D}$

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

  where we have defined the suspiciousness S, which is prior independent, and the information $\mathcal{I}$, which depends on the parameter compression of the shared space
- Focussing on the prior-independent portion $\mathcal{S}$ gives R for the "Narrowest reasonable priors" which do not impinge on the posterior
- One of the critical observations is that one can only hide tension by widening priors. Narrowing them will only ever show tension if it is present.

Suspiciousness S

- For a Gaussian set of posteriors:

$$\log\mathcal{S} = \frac{d}{2} - \frac{1}{2}(\mu_A-\mu_B)(\Sigma_A+\Sigma_B)^{-1}(\mu_A-\mu_B).$$

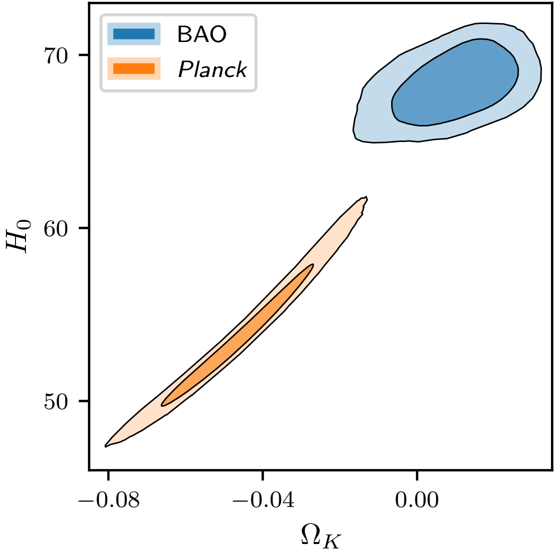
- The Malhanobis term is suggestive, so we can use this to calibrate a "sigma" level of tension using a χ^2 distribution for $\chi^2_d = d - 2\log\mathcal{S}$, or a tension probability.
- S is composed of evidences $\mathcal{Z}$ and KL divergences $\mathcal{D}$, which are Gaussian-independent concepts, so the only thing to determine is d, the "number of shared parameters".
- Can do this with Gaussian dimensionality $\frac{d}{2} = \mathrm{var}_{\mathcal{P}}(\log\mathcal{L})$ [1903.06682]

$$\begin{aligned} \text{Planck vs BAO}: &\quad p = 42\pm 4\% \\ \text{Planck vs DESY1}: &\quad p = 3.2\pm 1.0\% \\ \text{Planck vs S}H_0\text{ES}: &\quad p = 0.25\pm 0.17\% \end{aligned}$$

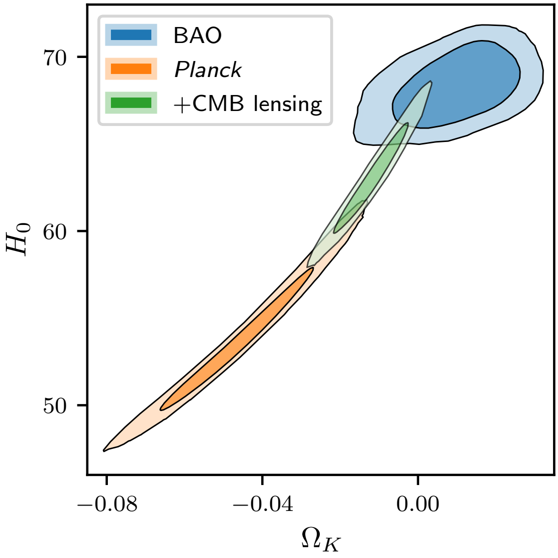
- Under this metric, SH_0ES is unambiguously inconsistent, although not quite as brutal as $>4\sigma$. BAO is consistent, and DESY1 is inconsistent, but only just. This is pleasingly similar to ones intuition.

Curvature tension? [1908.09139]

- If you allow $\Omega_K \neq 0$, *Planck* (`plikTTTEEE`) has a moderate preference for closed universes (50:1 betting odds on), $\Omega_K = -4.5 \pm 1.5\%$ [1911.02087]
- *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` $\to$ `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]

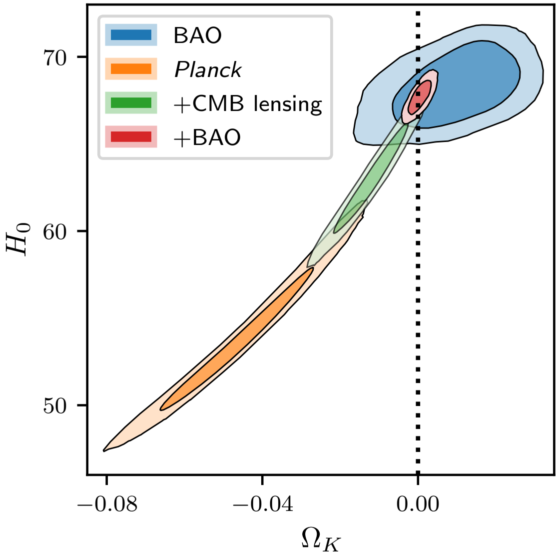


- If you allow $\Omega_K \neq 0$, *Planck* (`plikTTTEEE`) has a moderate preference for closed universes (50:1 betting odds on), $\Omega_K = -4.5 \pm 1.5\%$ [1911.02087]
- *Planck*+lens+BAO strongly prefer $\Omega_K = 0$.
- But, *Planck* vs lensing is 2.5σ in tension, and Planck vs BAO is 3σ.
- Reduced if `plik` $\to$ `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]



Curvature tension? [1908.09139]

- If you allow $\Omega_K \neq 0$, *Planck* (`plikTTTEEE`) has a moderate preference for closed universes (50:1 betting odds on), $\Omega_K = -4.5 \pm 1.5\%$ [1911.02087]
- *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]



Conclusions

- DiRAC RAC allocation for building a legacy grid of
  - MCMC & Nested sampling chains
  - gridded over (pairwise) up-to-date datasets
  - gridded over extensions to ΛCDM
  - Bijectors & emulators for fast re-use
  - Importance sampling toolkit via `anesthetic` for (re)processing
  - Long-term goal: community repository of chains to share model comparison compute resource
- Looking for:
  - α-testers for `unimpeded`
  - Suggestions for more datasets (and their incorporation into `cobaya`)