

# Nested Sampling: a multi-purpose numerical tool for particle physics and cosmology

Will Handley  
[<wh260@cam.ac.uk>](mailto: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](https://github.com/williamjameshandley/talks)

29<sup>th</sup> July 2022



**The  
Alan Turing  
Institute**



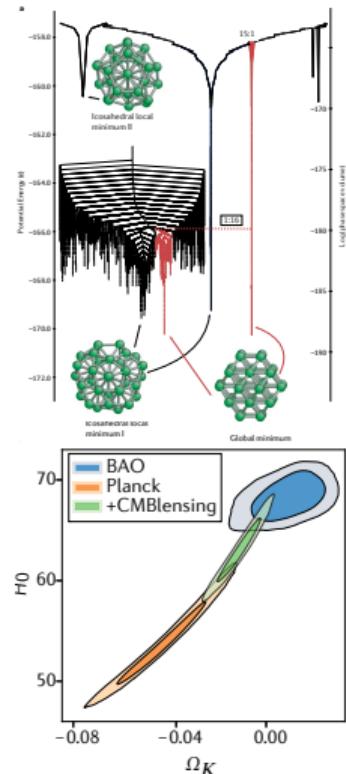
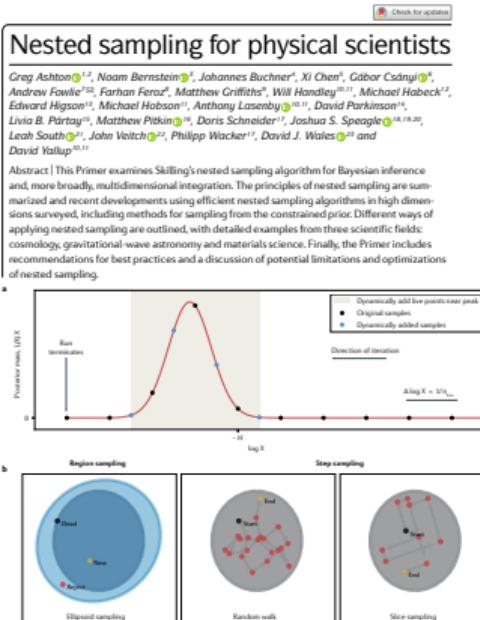
UNIVERSITY OF  
CAMBRIDGE



# 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

## PRIMER



# What is Nested Sampling?

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

Optimisation

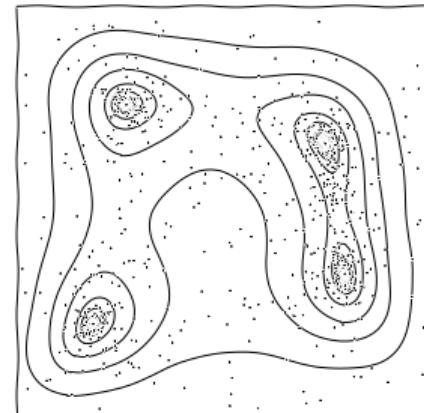
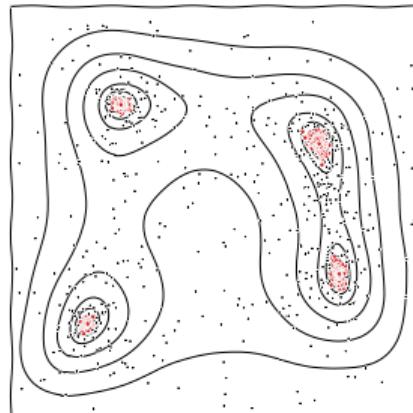
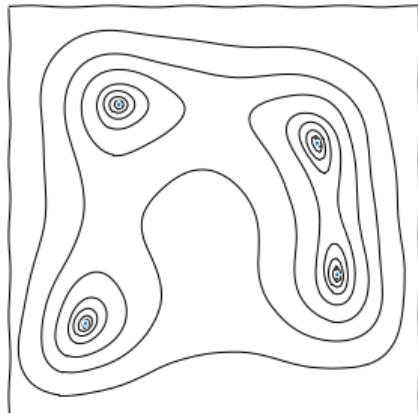
Sampling

Integration

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

draw  $\theta \sim f$

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



# MCMC sampling

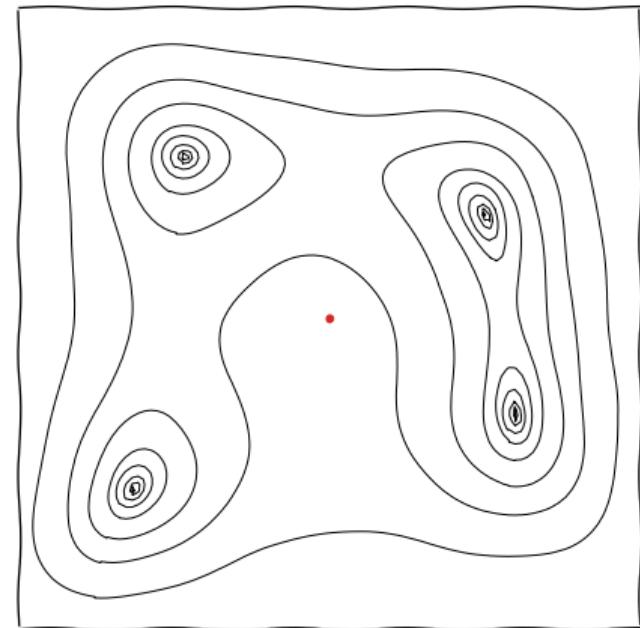
- ▶ Markov chain based methods generate samples from distribution by a stepping procedure.
- ▶ This can get stuck in local peaks.
- ▶ Cannot compute normalisation  $\mathcal{Z}$  of Bayes theorem:

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

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}, \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

- ▶ We generally want the evidence  $\mathcal{Z} = P(D|M)$  for the second stage of inference: model comparison:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \quad \text{Science}(M) = \frac{\mathcal{Z}_M \Pi_M}{\sum_m \mathcal{Z}_m \Pi_m}.$$



# MCMC sampling

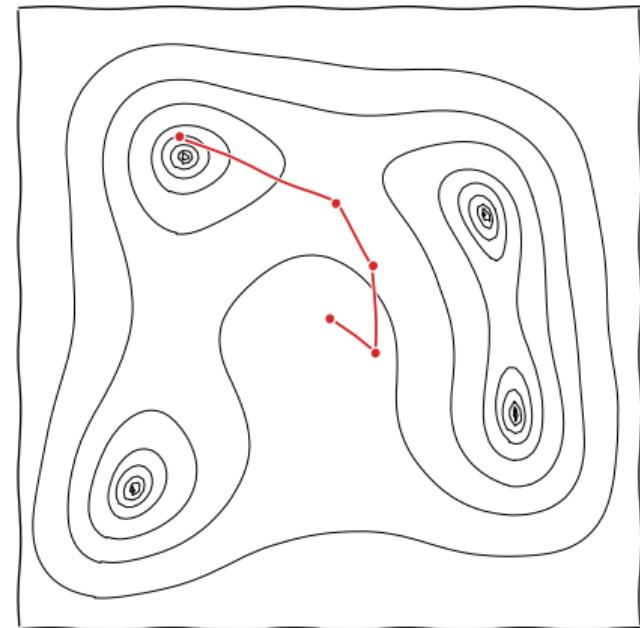
- ▶ Markov chain based methods generate samples from distribution by a stepping procedure.
- ▶ This can get stuck in local peaks.
- ▶ Cannot compute normalisation  $\mathcal{Z}$  of Bayes theorem:

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

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}, \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

- ▶ We generally want the evidence  $\mathcal{Z} = P(D|M)$  for the second stage of inference: model comparison:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \quad \text{Science}(M) = \frac{\mathcal{Z}_M \Pi_M}{\sum_m \mathcal{Z}_m \Pi_m}.$$



# MCMC sampling

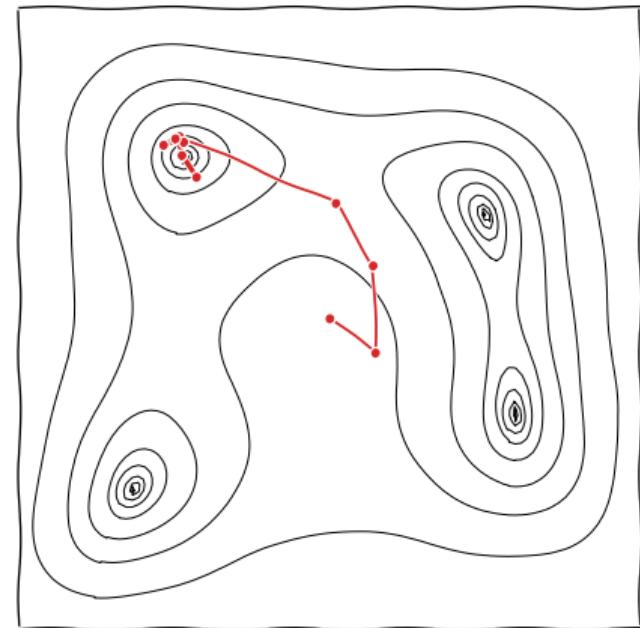
- ▶ Markov chain based methods generate samples from distribution by a stepping procedure.
- ▶ This can get stuck in local peaks.
- ▶ Cannot compute normalisation  $\mathcal{Z}$  of Bayes theorem:

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

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}, \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

- ▶ We generally want the evidence  $\mathcal{Z} = P(D|M)$  for the second stage of inference: model comparison:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \quad \text{Science}(M) = \frac{\mathcal{Z}_M \Pi_M}{\sum_m \mathcal{Z}_m \Pi_m}.$$



# MCMC sampling

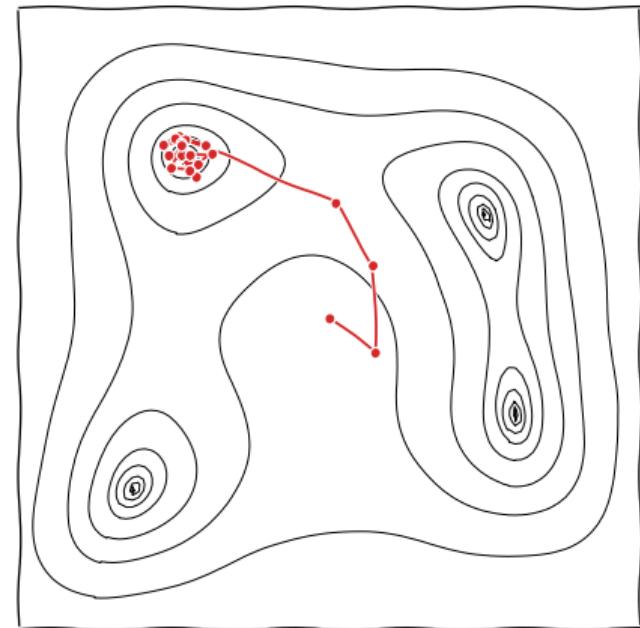
- ▶ Markov chain based methods generate samples from distribution by a stepping procedure.
- ▶ This can get stuck in local peaks.
- ▶ Cannot compute normalisation  $\mathcal{Z}$  of Bayes theorem:

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

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}, \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

- ▶ We generally want the evidence  $\mathcal{Z} = P(D|M)$  for the second stage of inference: model comparison:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \quad \text{Science}(M) = \frac{\mathcal{Z}_M \Pi_M}{\sum_m \mathcal{Z}_m \Pi_m}.$$



# MCMC sampling

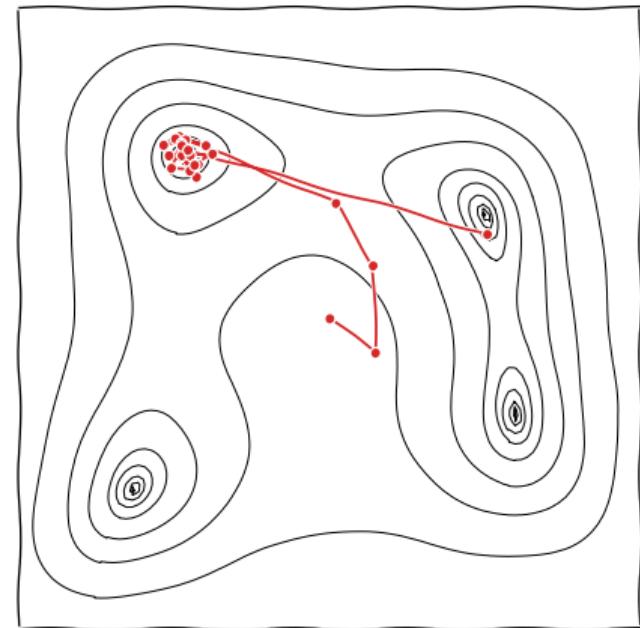
- ▶ Markov chain based methods generate samples from distribution by a stepping procedure.
- ▶ This can get stuck in local peaks.
- ▶ Cannot compute normalisation  $\mathcal{Z}$  of Bayes theorem:

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

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}, \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

- ▶ We generally want the evidence  $\mathcal{Z} = P(D|M)$  for the second stage of inference: model comparison:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \quad \text{Science}(M) = \frac{\mathcal{Z}_M \Pi_M}{\sum_m \mathcal{Z}_m \Pi_m}.$$



# MCMC sampling

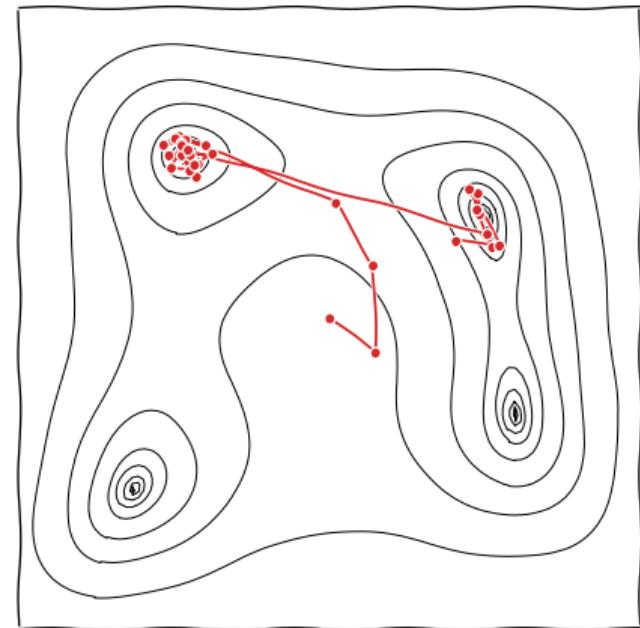
- ▶ Markov chain based methods generate samples from distribution by a stepping procedure.
- ▶ This can get stuck in local peaks.
- ▶ Cannot compute normalisation  $\mathcal{Z}$  of Bayes theorem:

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

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}, \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

- ▶ We generally want the evidence  $\mathcal{Z} = P(D|M)$  for the second stage of inference: model comparison:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \quad \text{Science}(M) = \frac{\mathcal{Z}_M \Pi_M}{\sum_m \mathcal{Z}_m \Pi_m}.$$

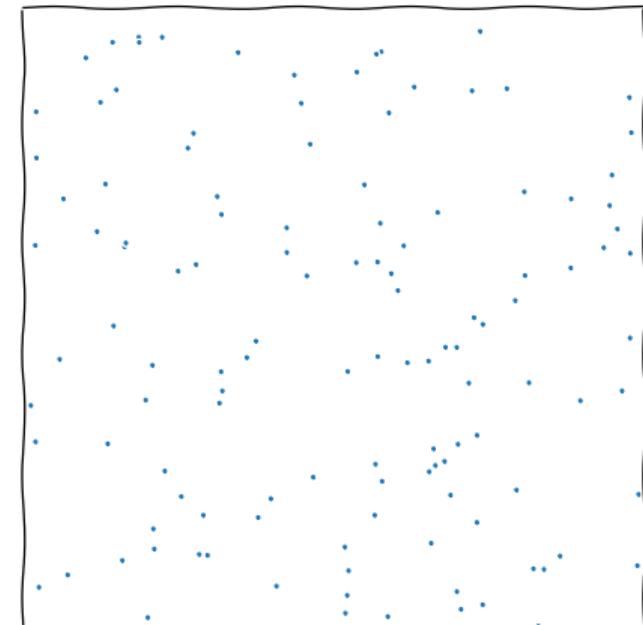


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

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

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.

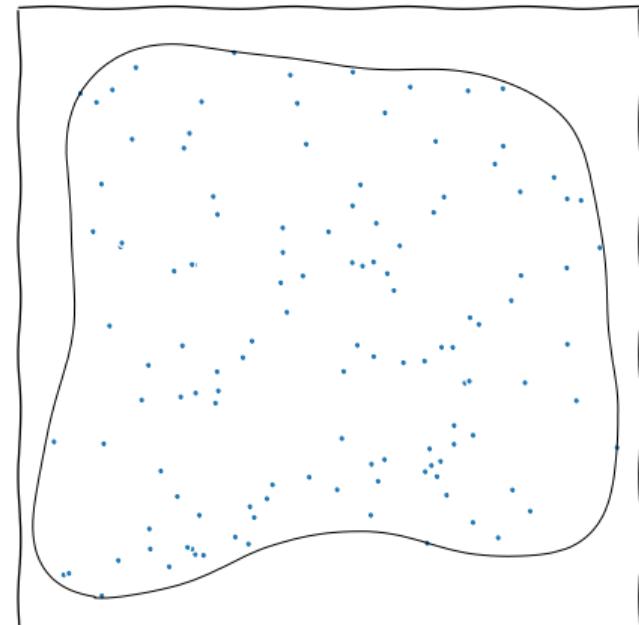


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

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

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.

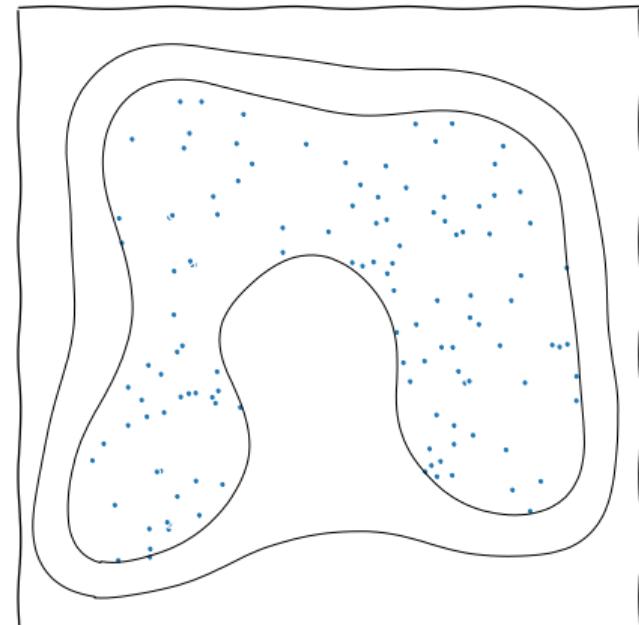


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

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

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.

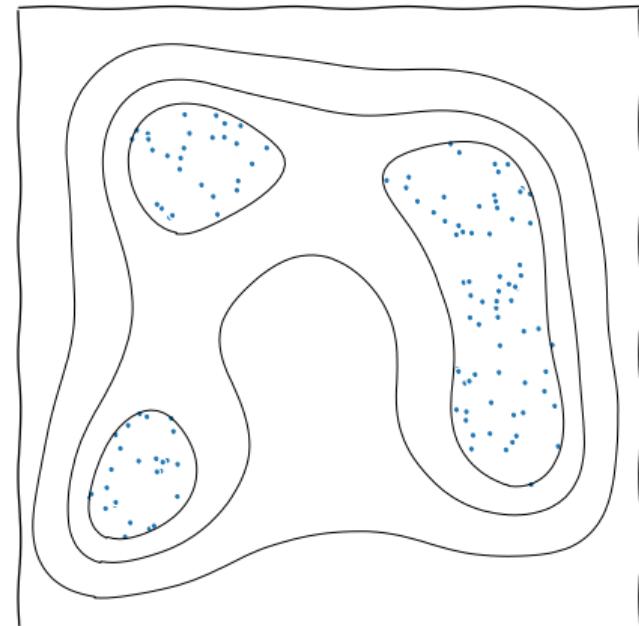


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

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

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.

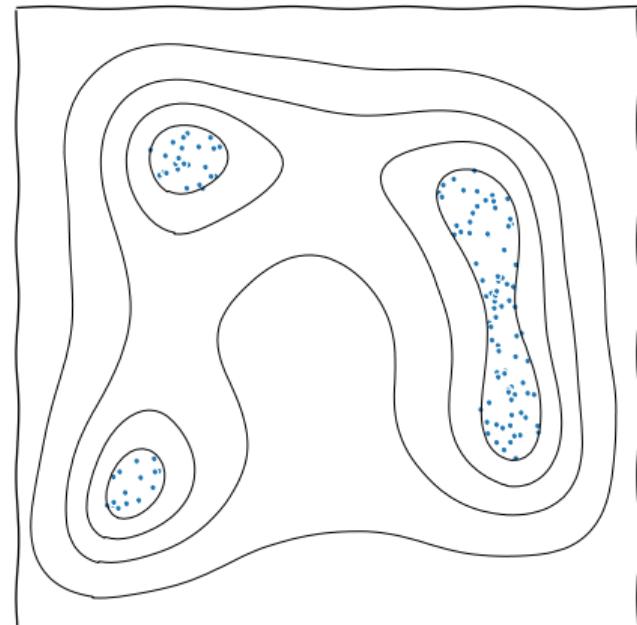


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

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

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.

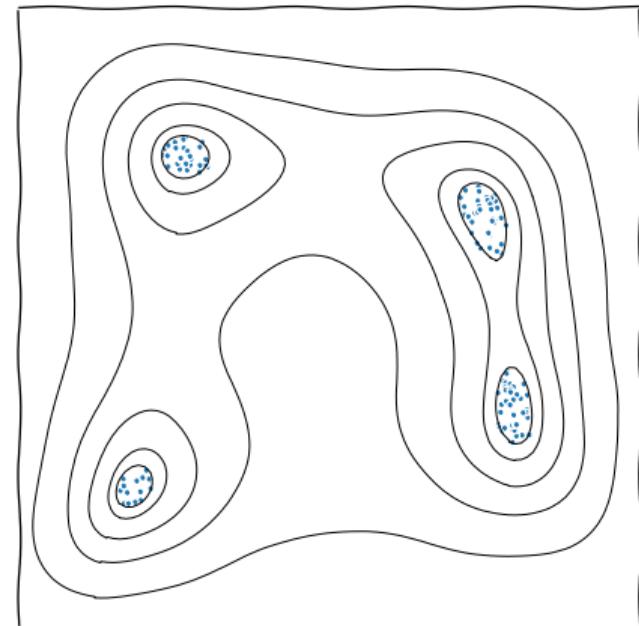


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

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

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.

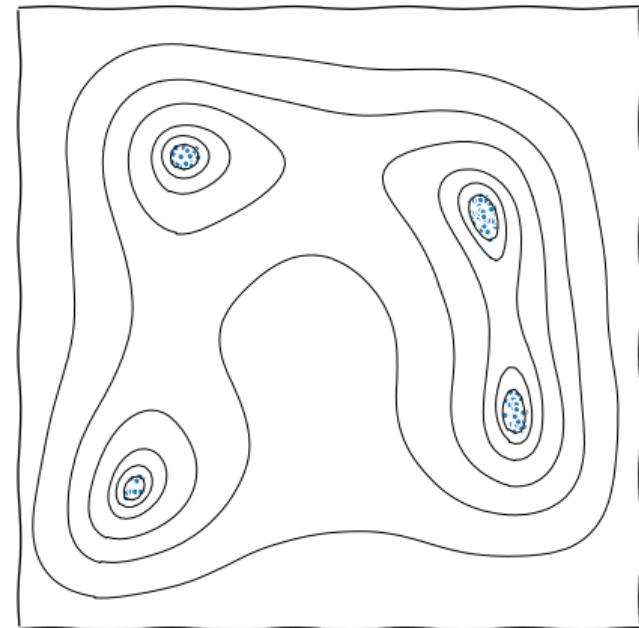


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

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

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.

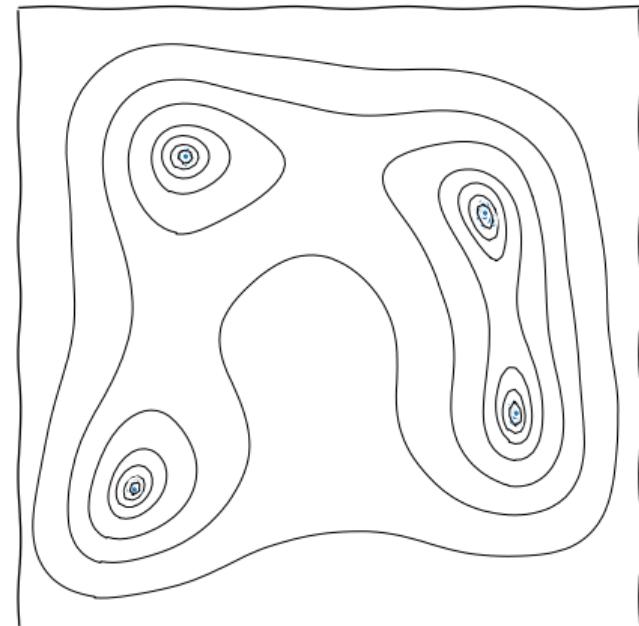


## Nested sampling

- ▶ Nested sampling: completely different way to scan.
- ▶ Ensemble sampling compresses entire space → peak(s).
- ▶ Sequentially update a set  $S$  of  $n$  samples:
  - $S_0$ : Generate  $n$  samples uniformly over the space (from a measure  $\pi$ ).
  - $S_{i+1}$ : 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 constraint*:

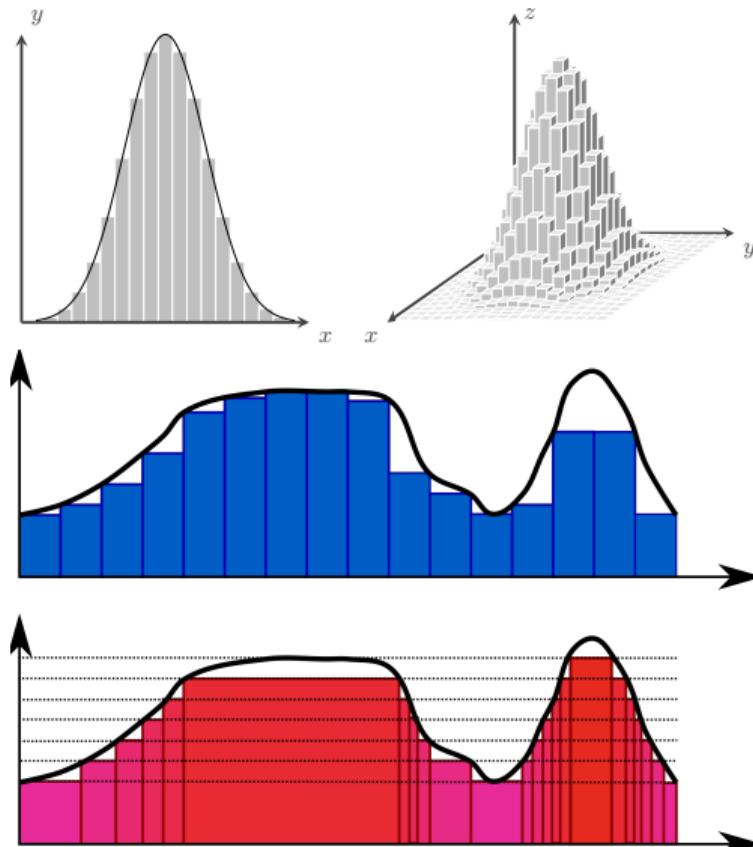
$$\{\theta \sim \pi : \mathcal{L}(\theta) > \mathcal{L}_*\}$$

- ▶ This procedure optimises (multimodally), and can calculate the **evidence**/integral of function & posterior/sample weights.



# Integration in high dimensions

- ▶ 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 **Lebesgue integration**.



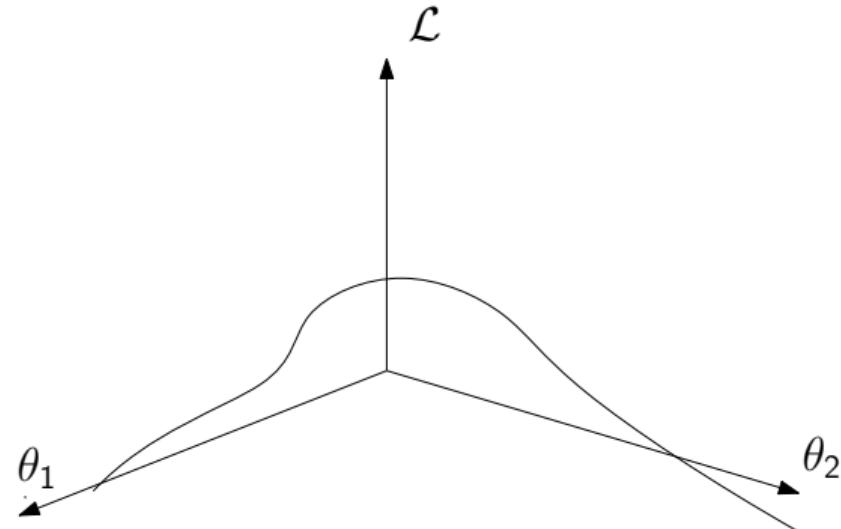
## (Lesbesgue) Integrating with nested sampling

- ▶ 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.

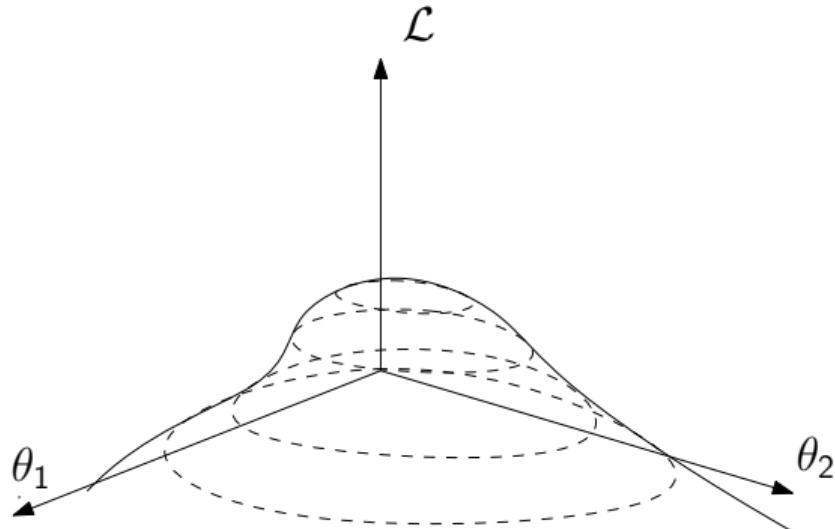
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

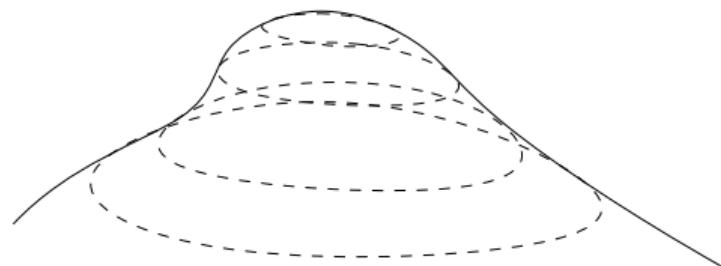
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

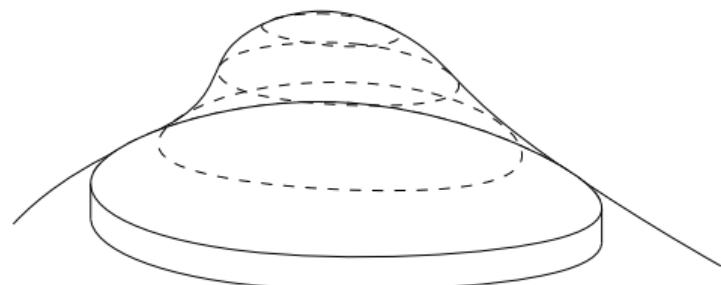
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

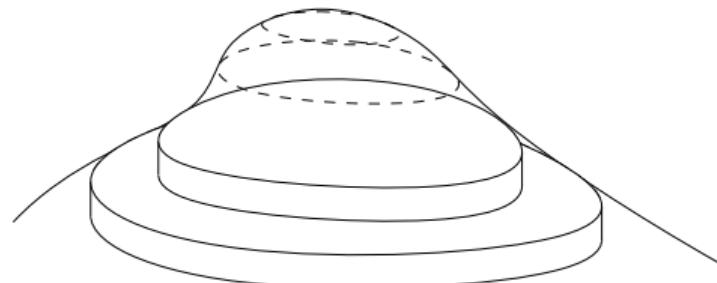
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

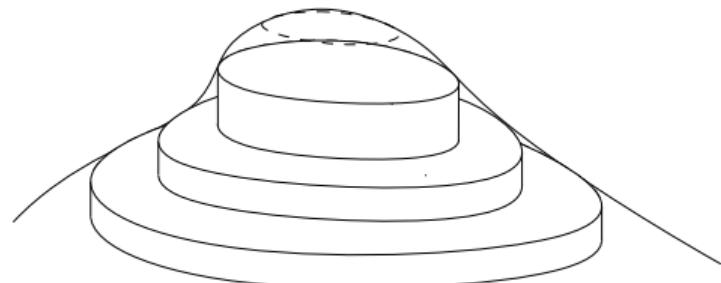
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

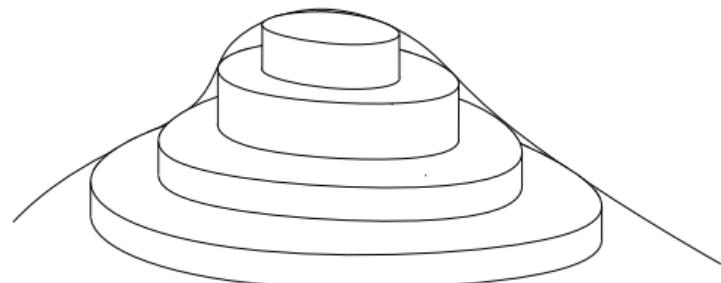
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

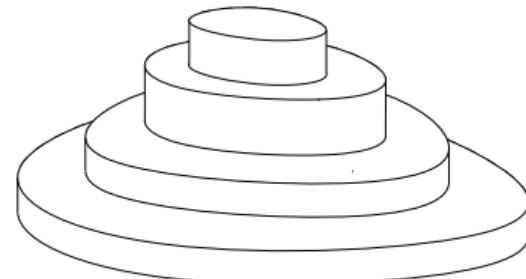
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

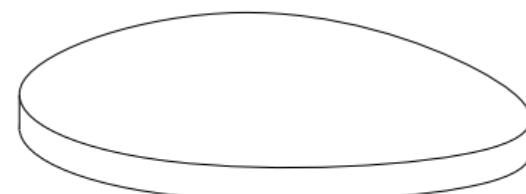
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ Integral can be discretised in several ways

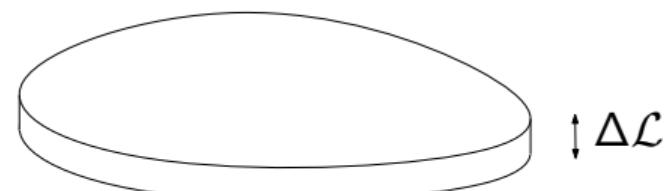
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i = \sum_i \frac{\mathcal{L}_i + \mathcal{L}_{i-1}}{2} (X_{i-1} - X_i).$$



## (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

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

- ▶ 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.

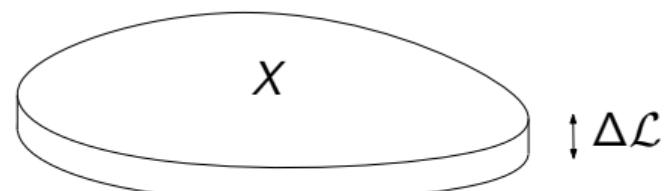
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

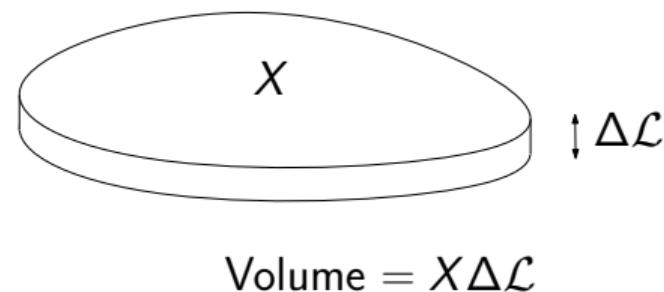
$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- ▶ 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

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



$$\text{Volume} = X\Delta\mathcal{L}$$

## (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.

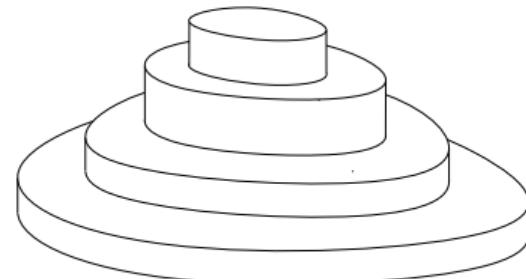
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

- ▶ 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.

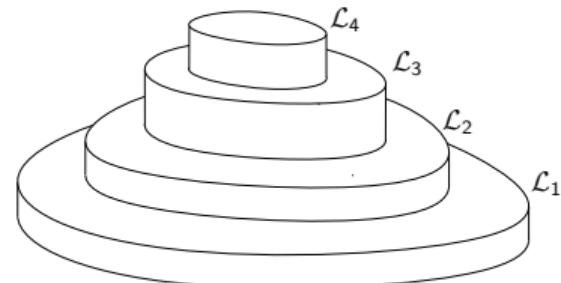
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

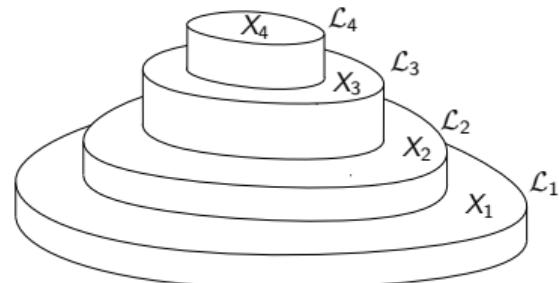
$$\mathcal{Z} \approx \sum_i \Delta \mathcal{L}_i X_i, \quad X_{i+1} \approx \frac{n}{n+1} X_i, \quad X_0 = 1.$$

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

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

- ▶ 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.

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

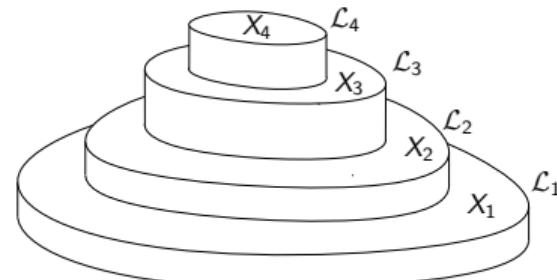
- ▶ Although this is only approximate, we can quantify the error

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

- ▶ Integral can be discretised in several ways

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

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

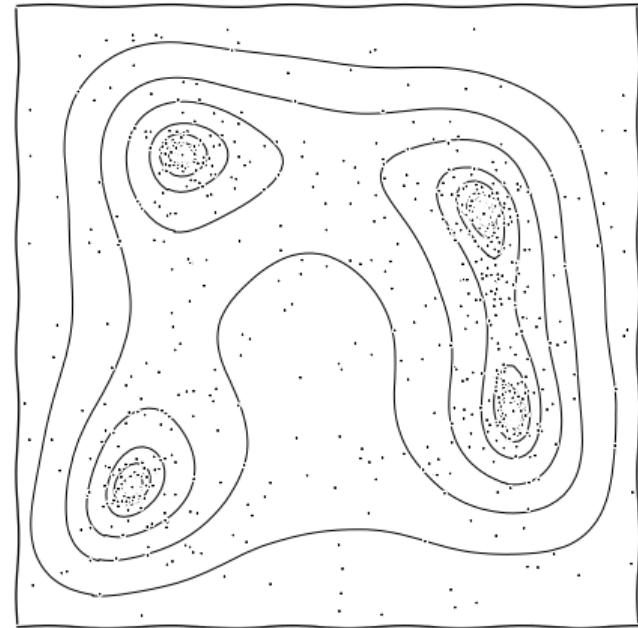


## Dead points: posteriors & evidences

- ▶ 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.

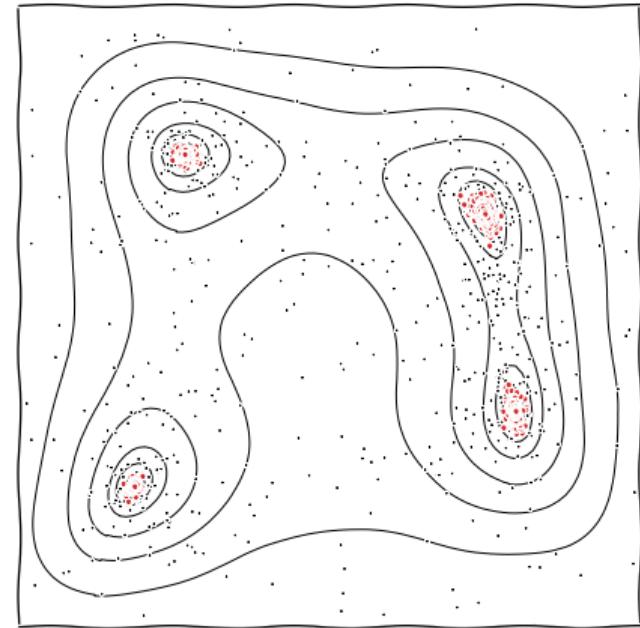


## Dead points: posteriors & evidences

- ▶ 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.



# Sampling from a hard likelihood constraint

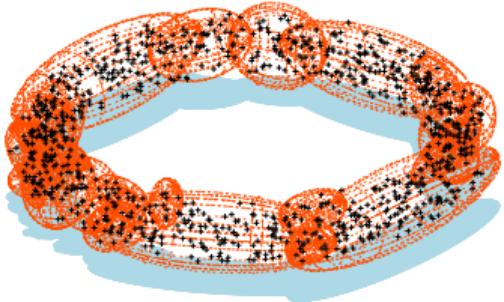
*"It is not the purpose of this introductory paper to develop the technology of navigation within such a volume. We merely note that exploring a hard-edged likelihood-constrained domain should prove to be neither more nor less demanding than exploring a likelihood-weighted space."*

— John Skilling

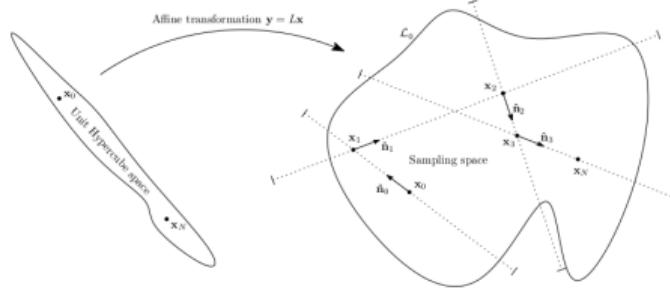
- ▶ A large fraction of the work in NS to date has been in attempting to implement a hard-edged sampler in the NS meta-algorithm  $\{\theta \sim \pi : \mathcal{L}(\theta) > \mathcal{L}_*\}$ .
- ▶ <https://projecteuclid.org/euclid.ba/1340370944>.
- ▶ There has also been much work beyond this (focus of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

# Implementations of Nested Sampling

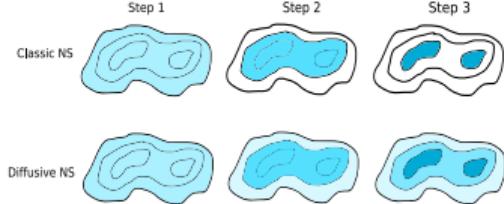
MultiNest [0809.3437]



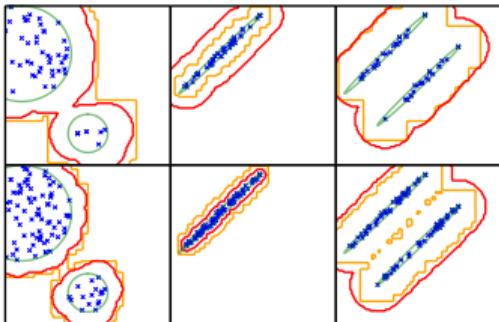
PolyChord [1506.00171]



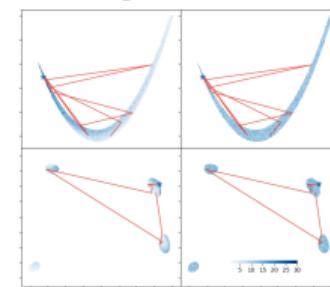
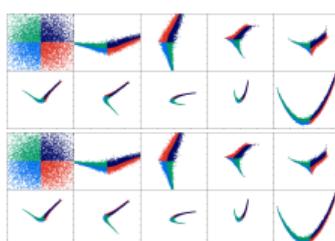
DNest [1606.03757]



UltraNest [2101.09604]

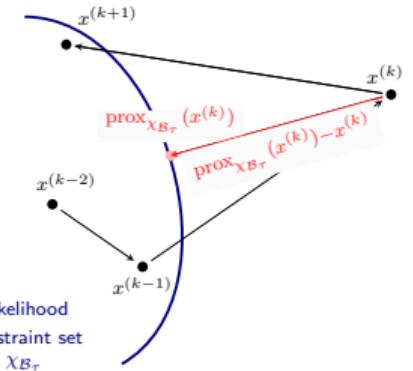


NeuralNest [1903.10860]



dynesty [1904.02180]

ProxNest [2106.03646]



# 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

## Chain-based samplers

- ▶ e.g. MultiNest, UltraNest.
- ▶ Constructs bounding region and draws many invalid points until one is found within  $\mathcal{L}_*$ .
- ▶ Efficient in low dimensions, exponentially inefficient  $\sim \mathcal{O}(e^{d/d_0})$  in high  $d > d_0 \sim 10$ .
- ▶ Nested samplers usually come with:
  - ▶ resolution parameter  $n_{\text{live}}$  (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  $n_{\text{repeats}}$ .
- ▶ 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.

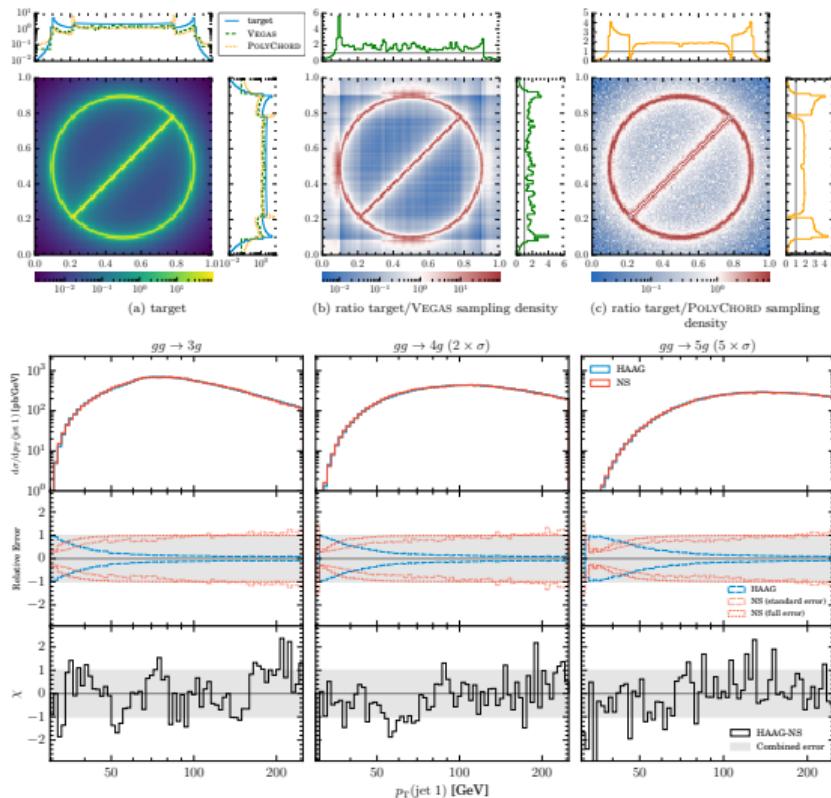
# Exploration of phase space [2106.02056]

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

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

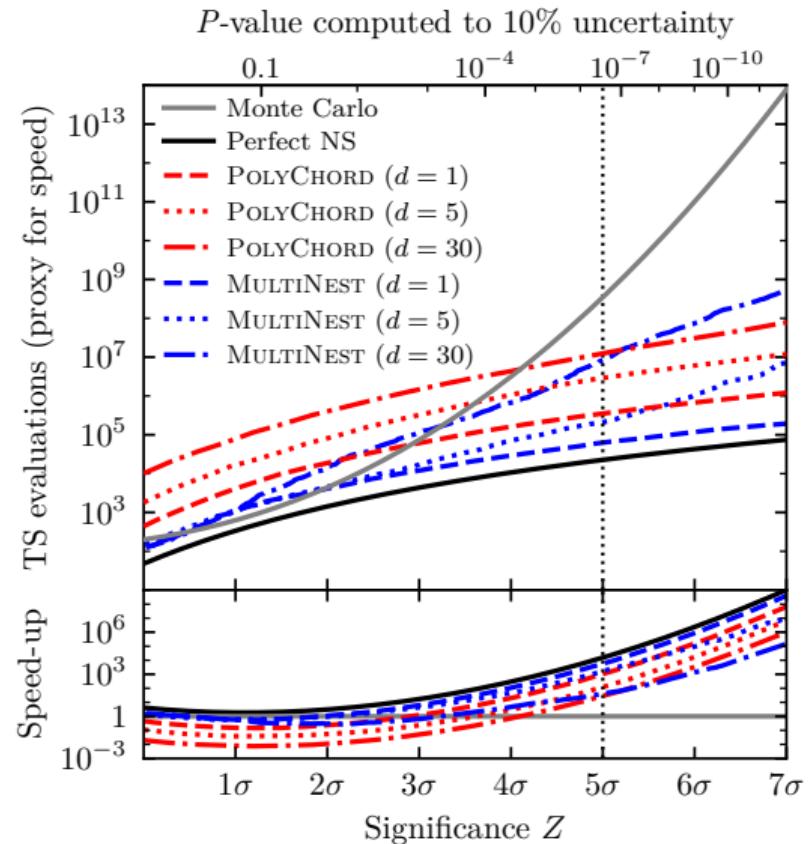
$\Omega$  phase space of kinematic configurations  $\Phi$ , each with matrix element  $\mathcal{M}(\Phi)$ .

- ▶ 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.



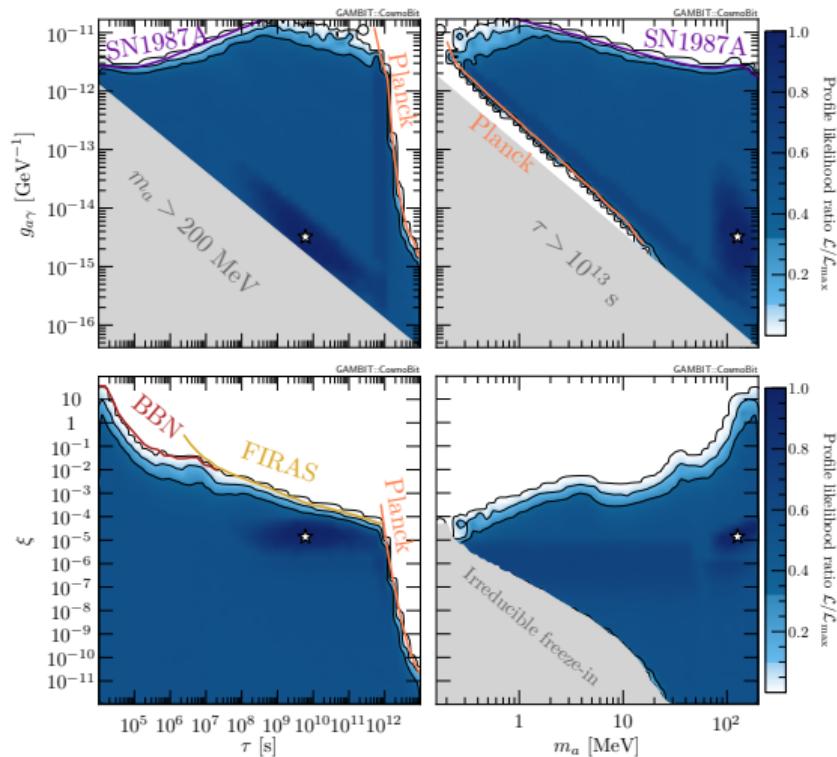
# Statistics: fast estimation of small $p$ -values [2106.02056](PRL)

- ▶ Nested sampling for frequentist computation!?
- ▶  $p$ -value:  $P(\lambda > \lambda^* | H_0)$  – probability that test statistic  $\lambda$  is at least as great as observed  $\lambda^*$ .
- ▶ Computation of a tail probability from sampling distribution of  $\lambda$  under  $H_0$ .
- ▶ For gold-standard  $5\sigma$ , 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.



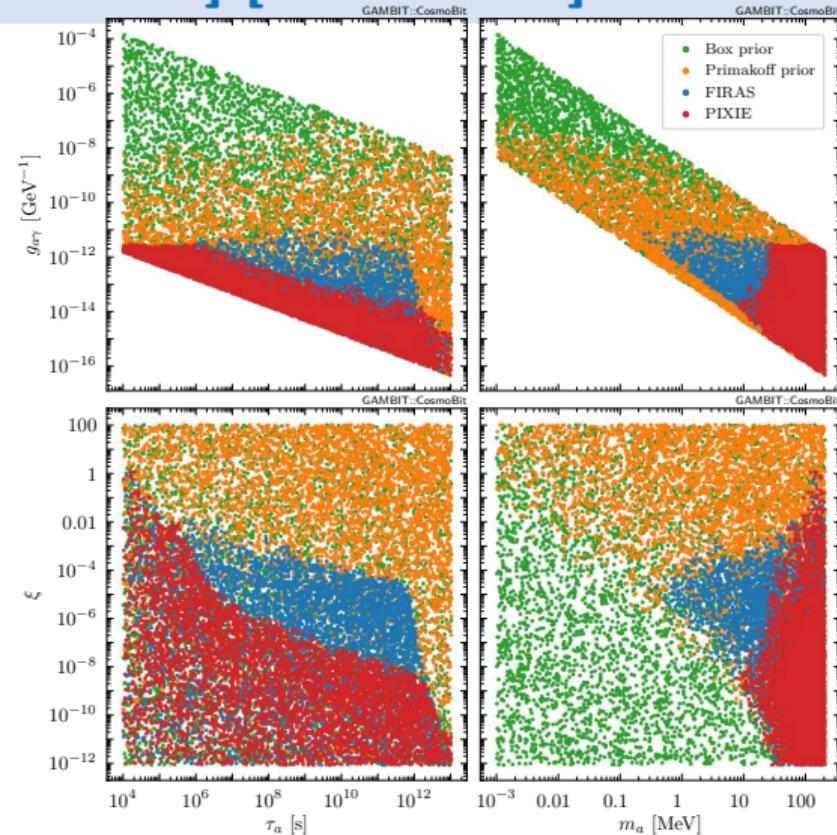
# Quantification of fine tuning [2101.00428] [2205.13549]

- ▶ 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,  $\tau_n$  and nuisance params)
- ▶ Data: CMB, BBN, FIRAS, SMM, BAO.
- ▶ Standard profile likelihood fit shows ruled out regions and best-fit point.



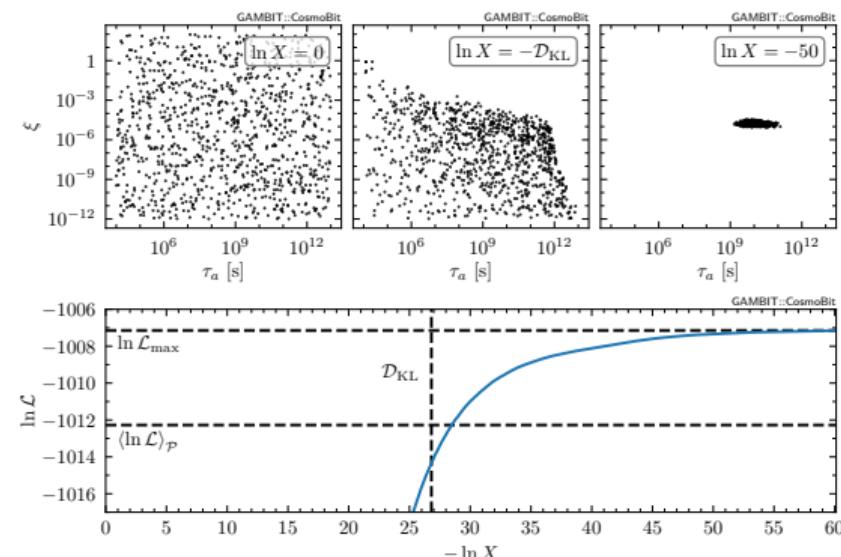
# Quantification of fine tuning [2101.00428] [2205.13549]

- ▶ 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,  $\tau_n$  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  $\mathcal{D}_{KL}$ .
  - ▶ 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,  $\tau_n$  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  $\mathcal{D}_{KL}$ .
  - ▶ 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.

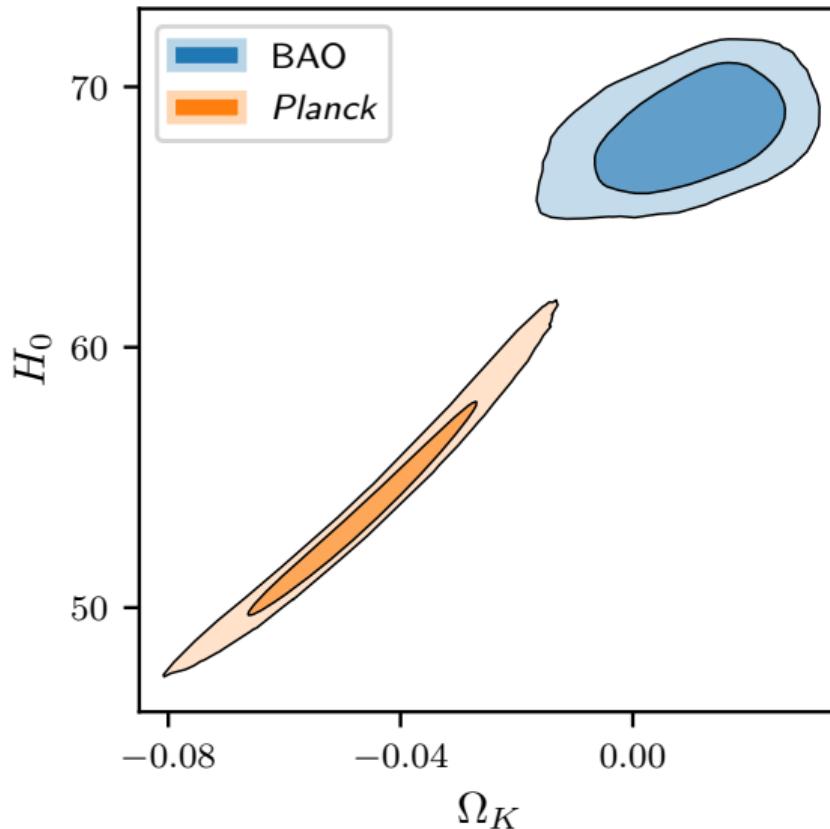


## What is a model?

- ▶ Model comparison in its purest form answers question such as:
  - ▶ “Is the universe  $\Lambda$ CDM?”
  - ▶ “Are neutrinos in a normal or inverted hierarchy?”
  - ▶ “Is there a detectable global signal in this data?”
- ▶ However model  $\mathcal{M}$  is likelihood  $\mathcal{L} = P(D|\theta, \mathcal{M})$  and priors  $\pi = P(\theta|\mathcal{M})$ ,  $\Pi = P(\mathcal{M})$ .
- ▶ 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?”

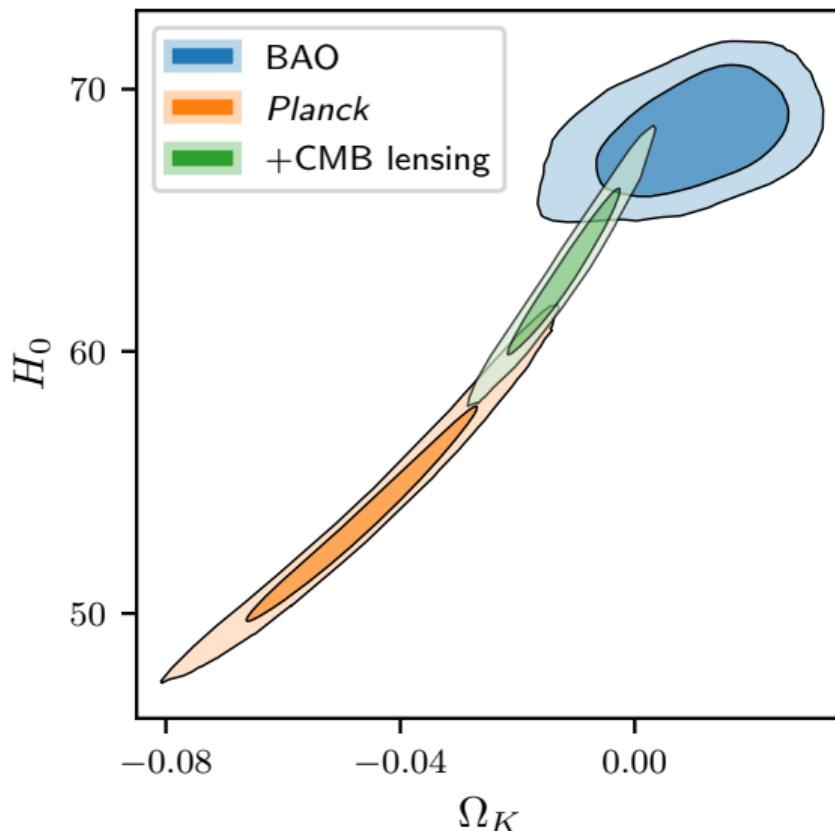
# Model comparison and parameter estimation [1908.09139]

- ▶ 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  $2.5\sigma$  in tension, and *Planck* vs BAO is  $3\sigma$ .
- ▶ Reduced if plik  $\rightarrow$  camspec [2002.06892]
- ▶ BAO and lensing summary assume  $\Lambda$ CDM.
- ▶ 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].



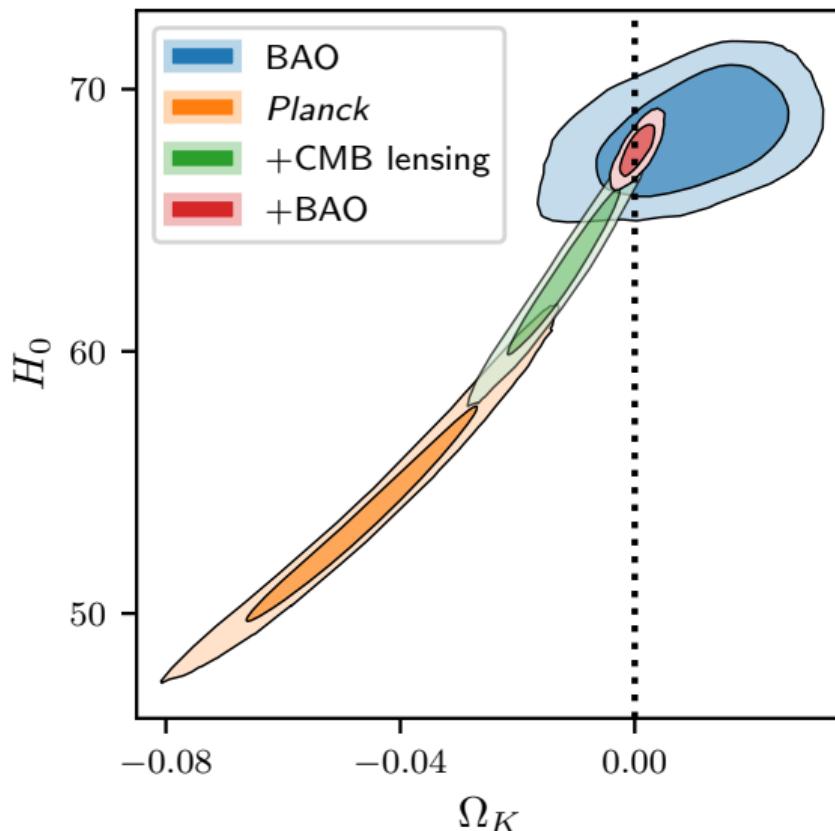
# Model comparison and parameter estimation [1908.09139]

- ▶ 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  $2.5\sigma$  in tension, and *Planck* vs BAO is  $3\sigma$ .
- ▶ Reduced if plik  $\rightarrow$  camspec [2002.06892]
- ▶ BAO and lensing summary assume  $\Lambda$ CDM.
- ▶ 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  $2.5\sigma$  in tension, and *Planck* vs BAO is  $3\sigma$ .
- ▶ Reduced if plik  $\rightarrow$  camspec [2002.06892]
- ▶ BAO and lensing summary assume  $\Lambda$ CDM.
- ▶ 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].

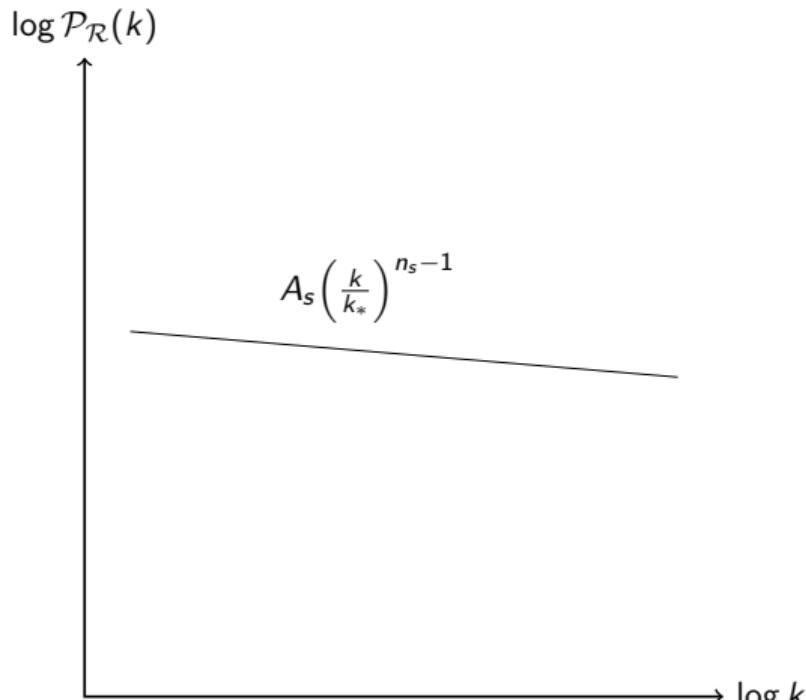


# Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction [1908.00906]

- ▶ 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  $n_{\text{run}}$  (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

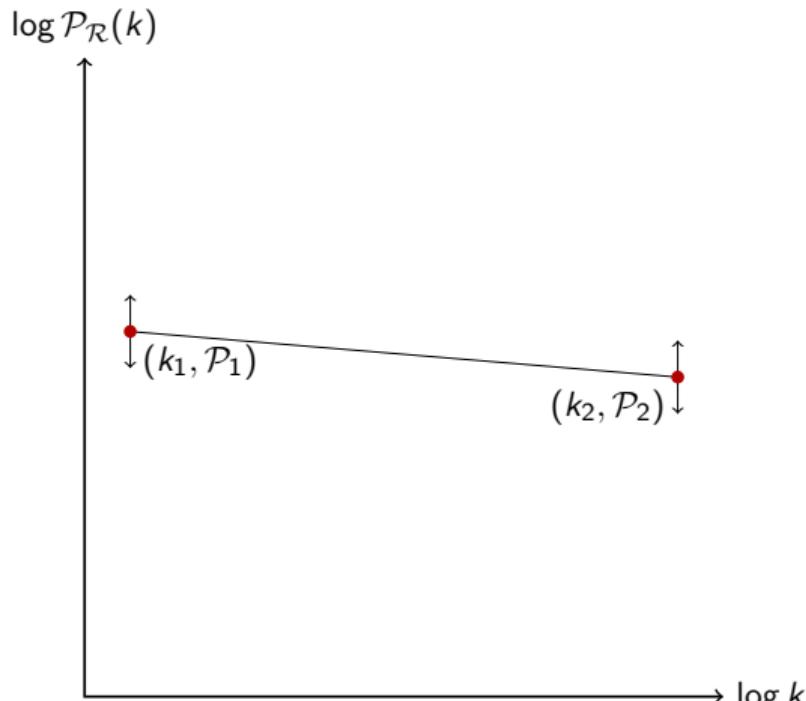


# Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction [1908.00906]

- ▶ 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  $n_{\text{run}}$  (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

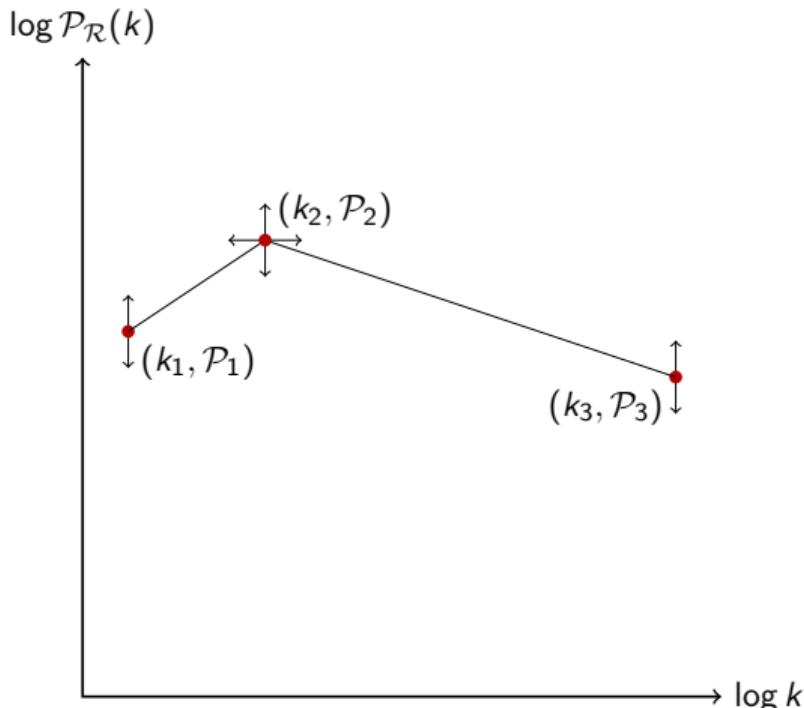


# Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction [1908.00906]

- ▶ 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  $n_{\text{run}}$  (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

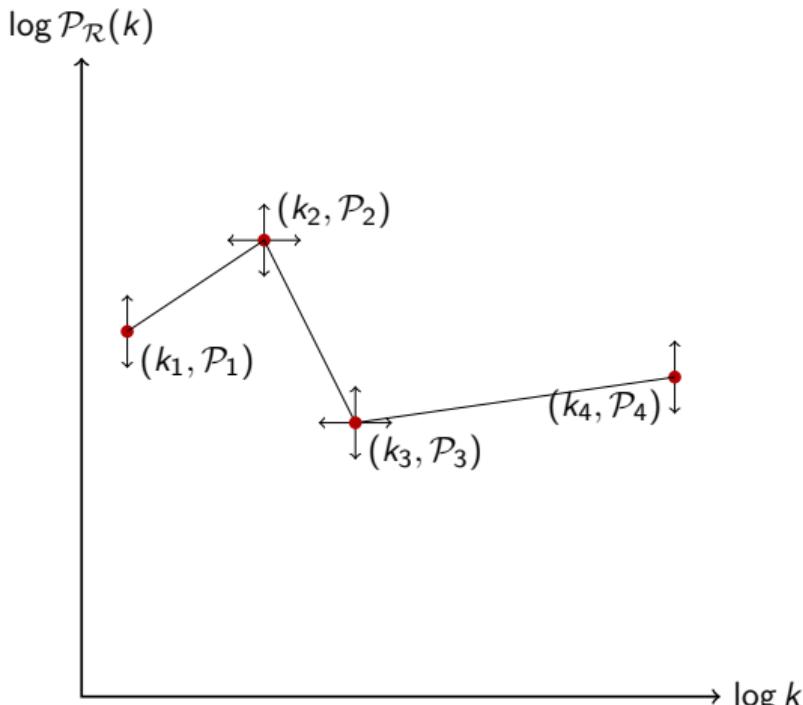


# Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction [1908.00906]

- ▶ 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  $n_{\text{run}}$  (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

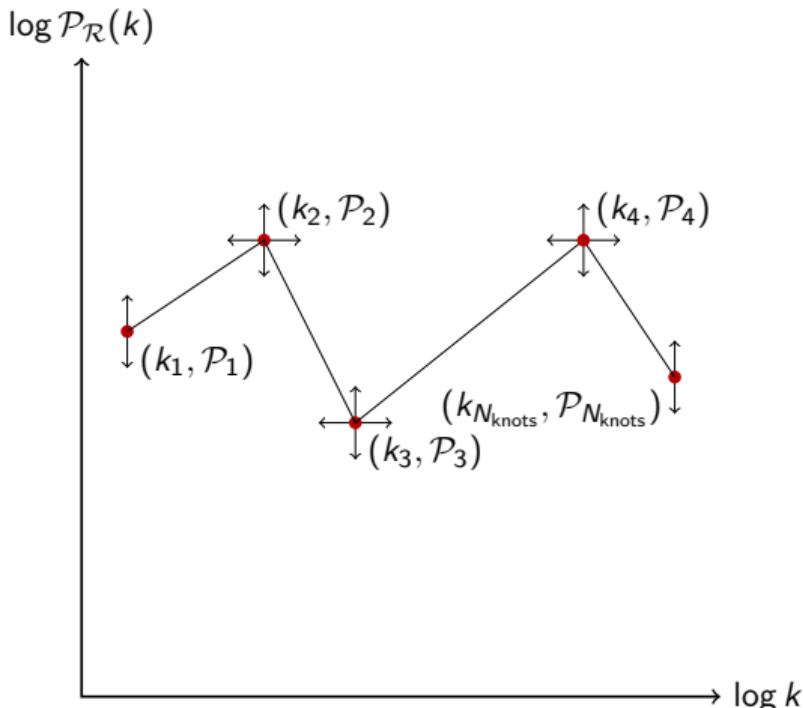


# Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction [1908.00906]

- ▶ 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  $n_{\text{run}}$  (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

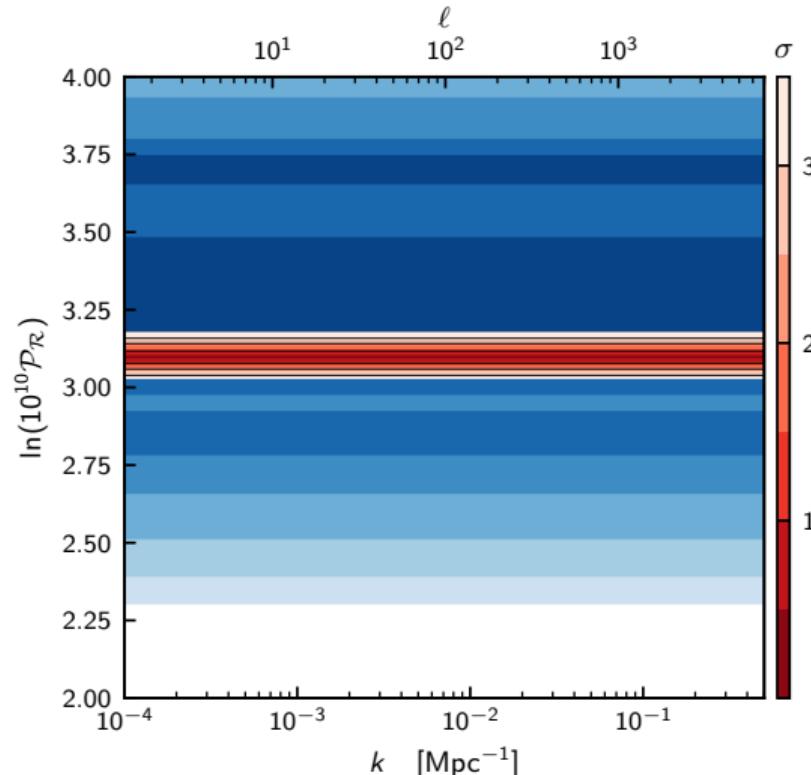


## 0 internal knots

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

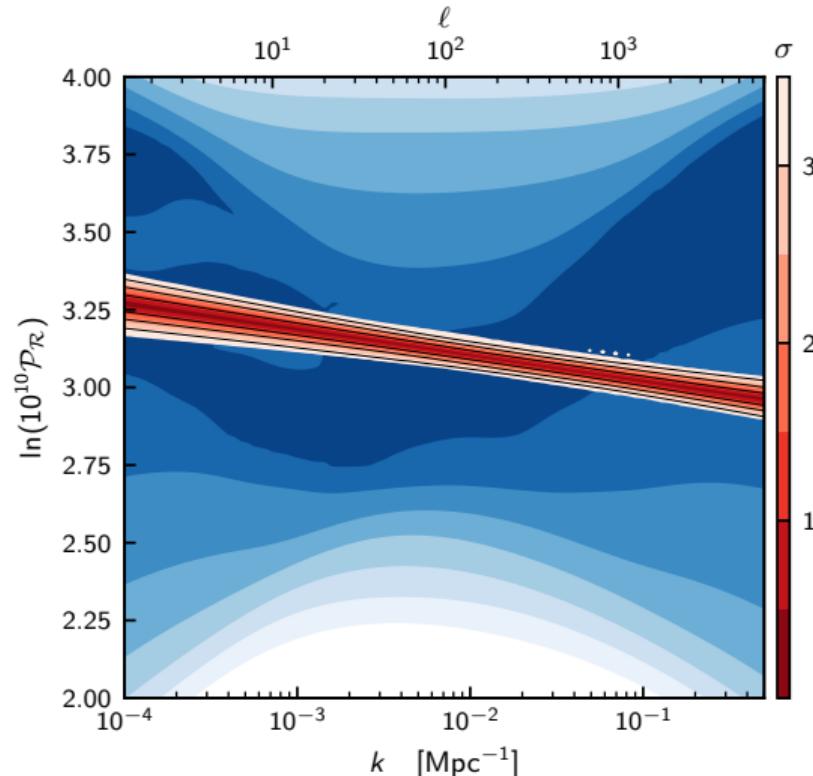


# 1 internal knot

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

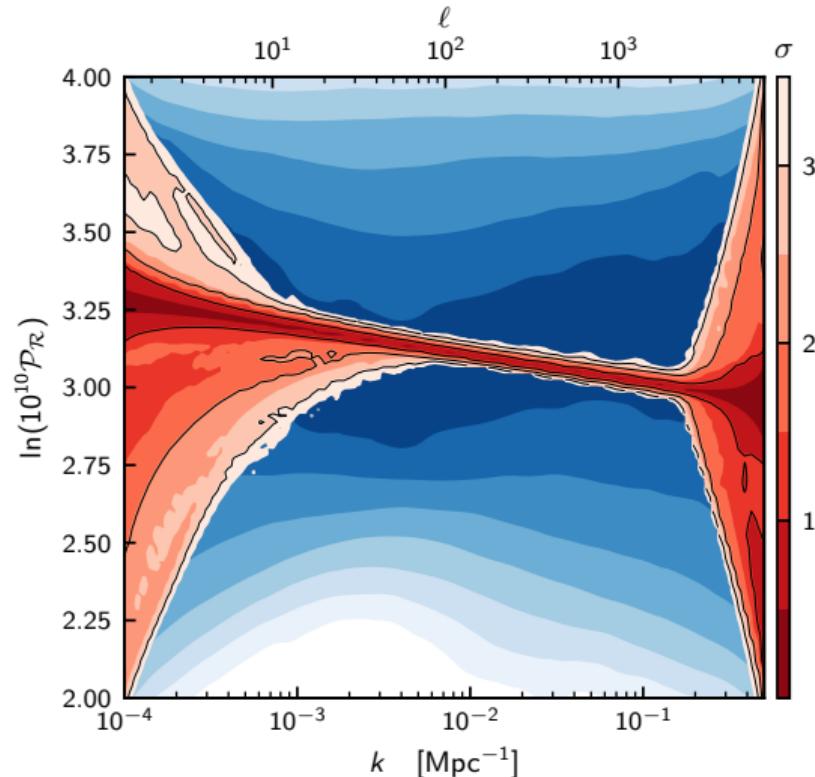


## 2 internal knots

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

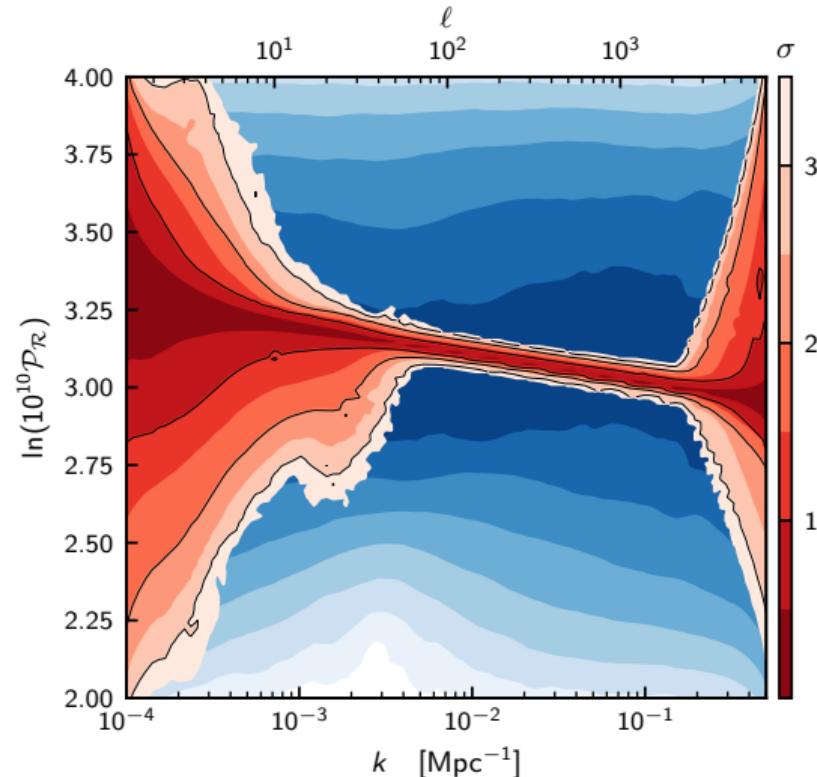


## 3 internal knots

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

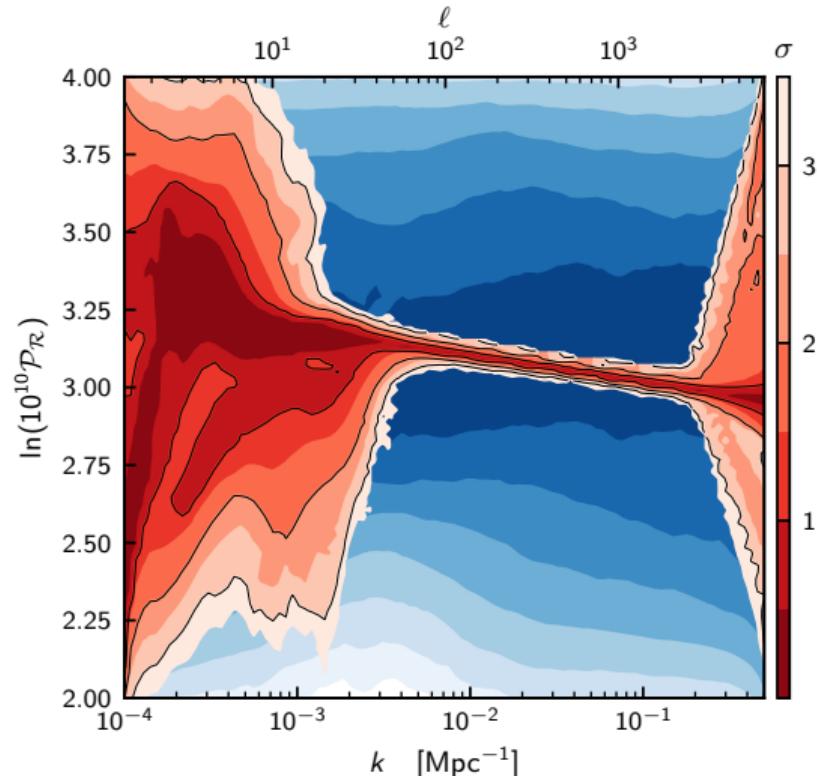


## 4 internal knots

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

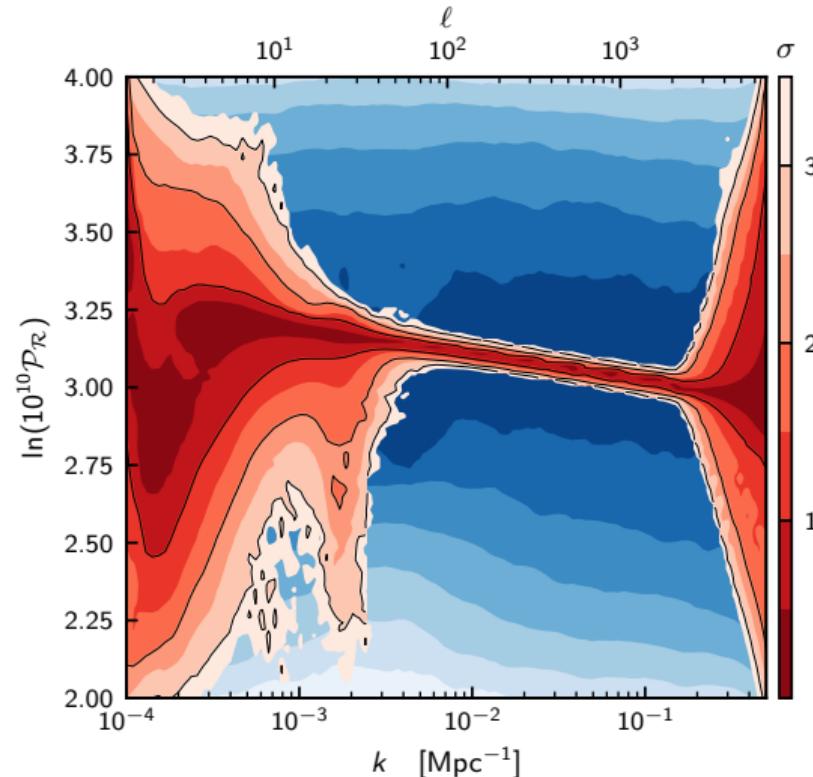


## 5 internal knots

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

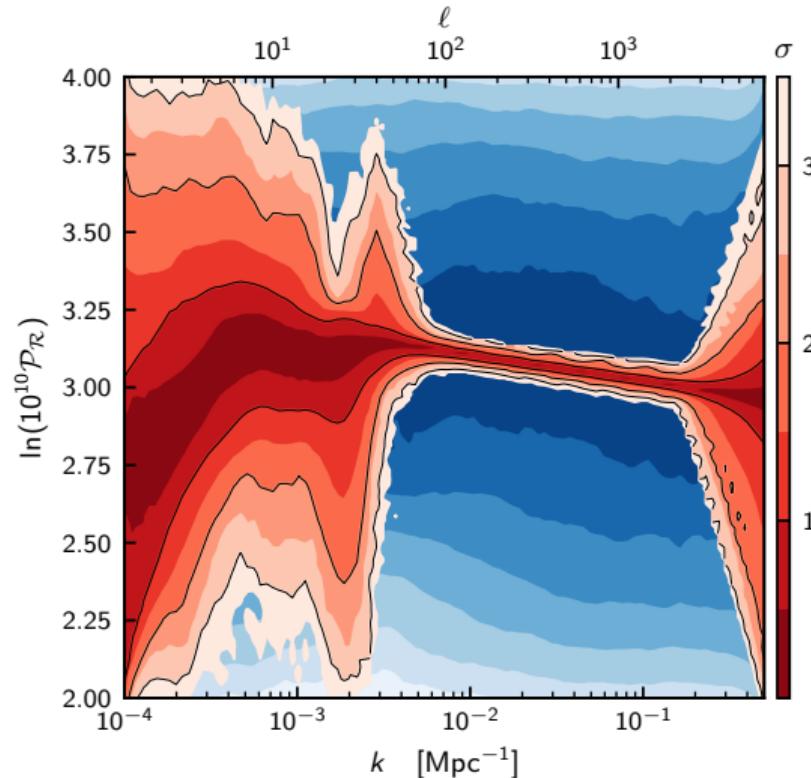


## 6 internal knots

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

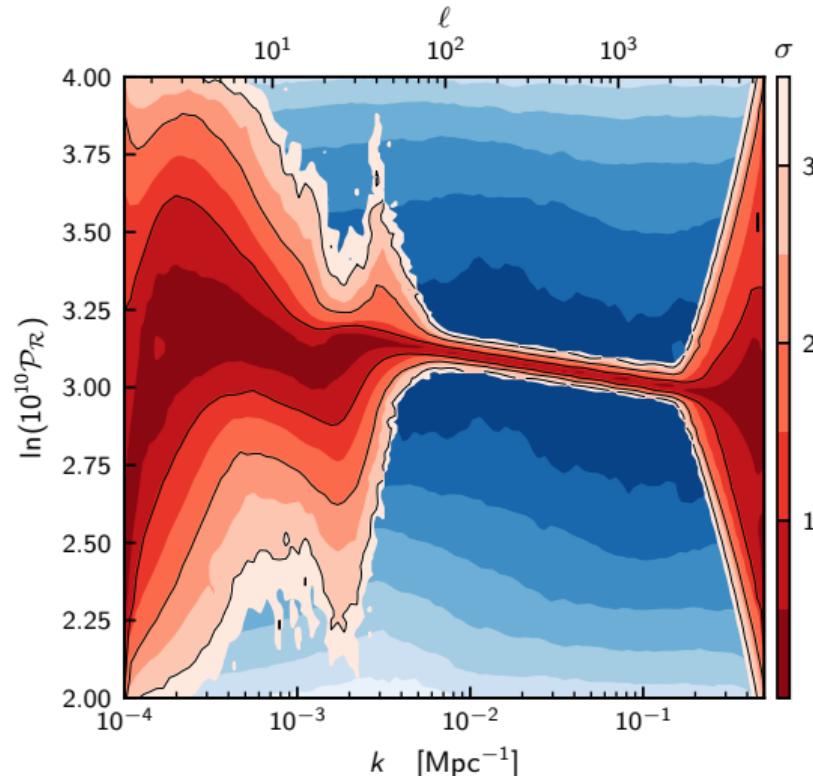


## 7 internal knots

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

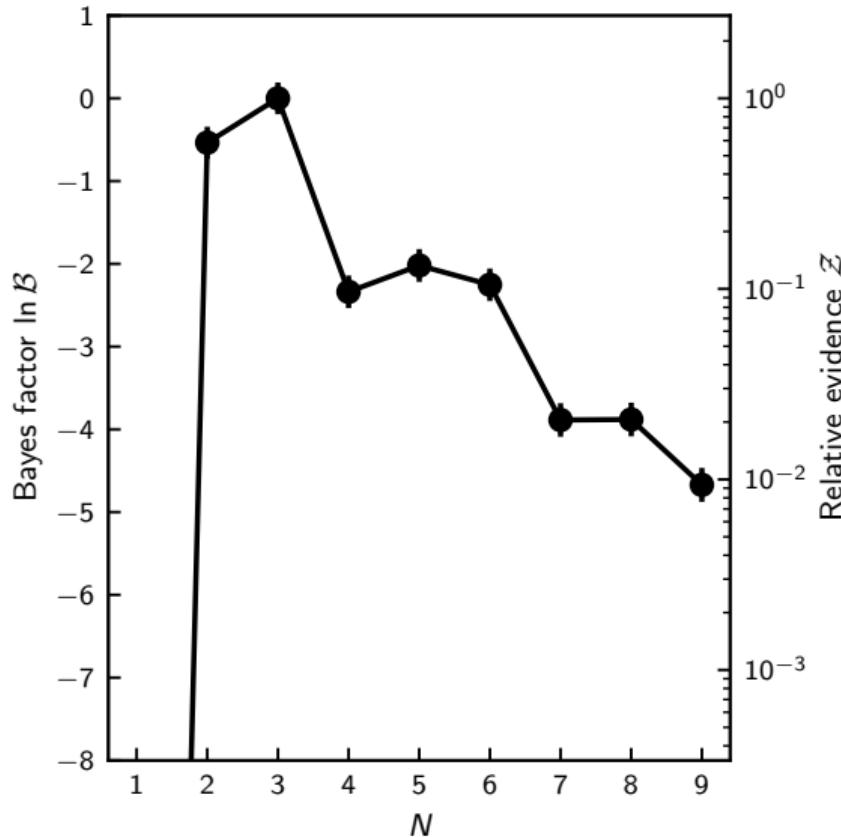


# Bayes Factors

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

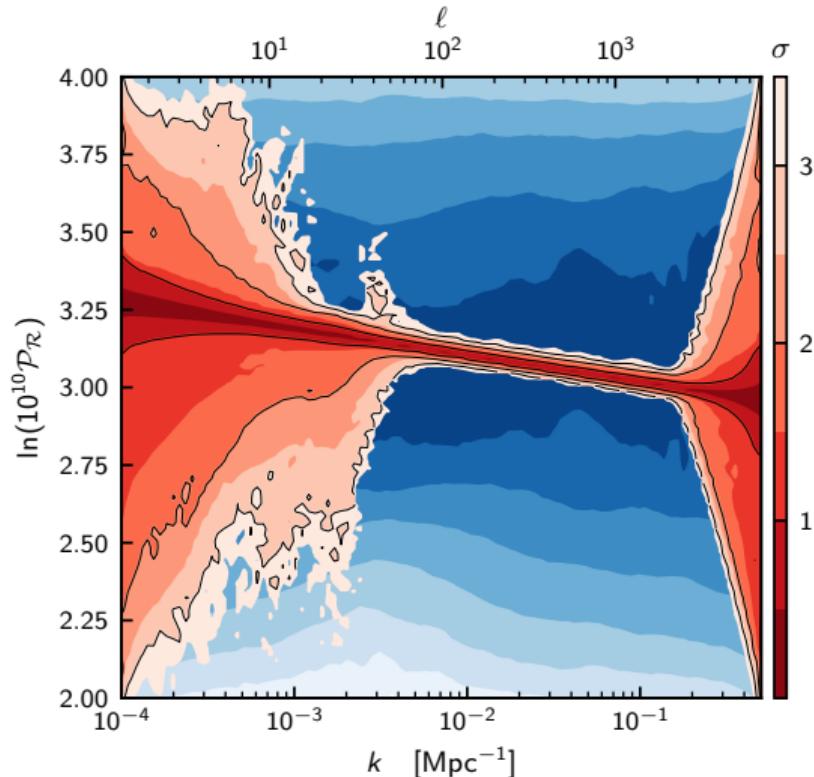


# Marginalised plot

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

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

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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

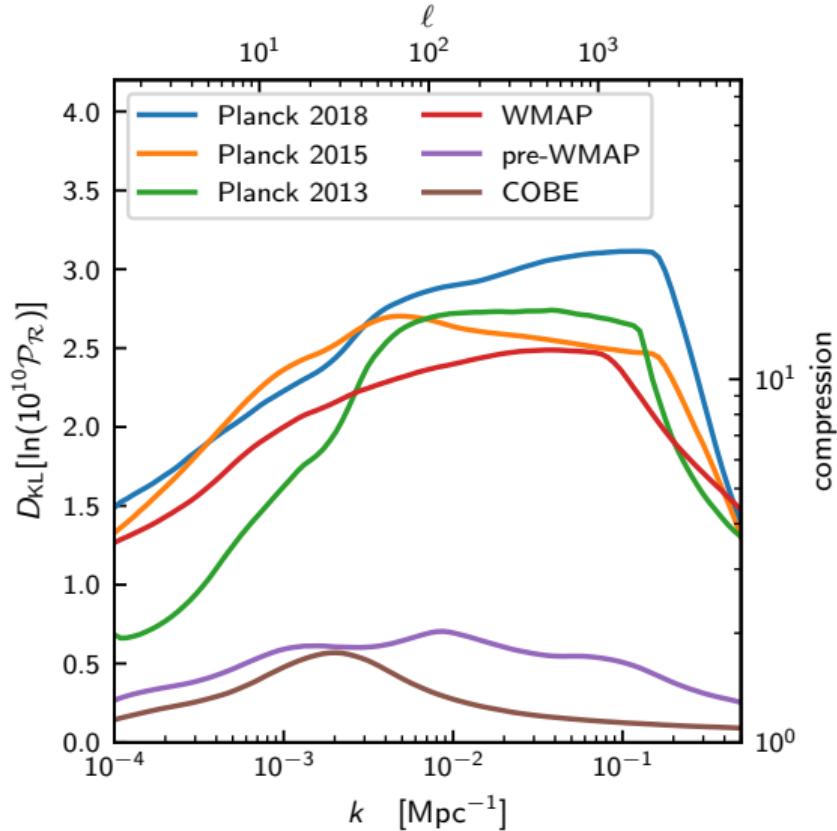


# Kullback-Liebler divergences

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

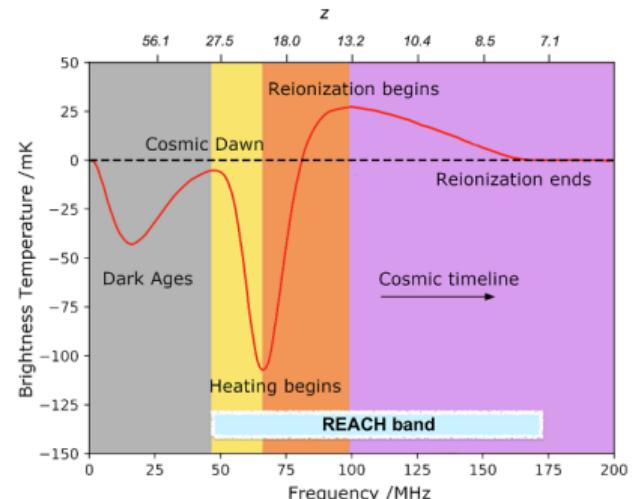
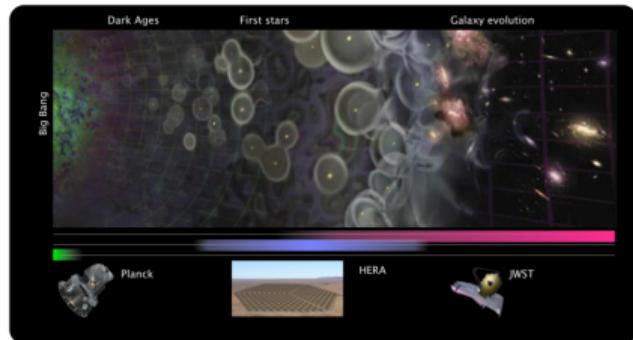
$$\mathcal{P}_R(k) = A_s \left( \frac{k}{k_*} \right)^{n_s - 1}$$

- ▶ To add more degrees of freedom, can add “running” parameters  $n_{\text{run}}$  (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



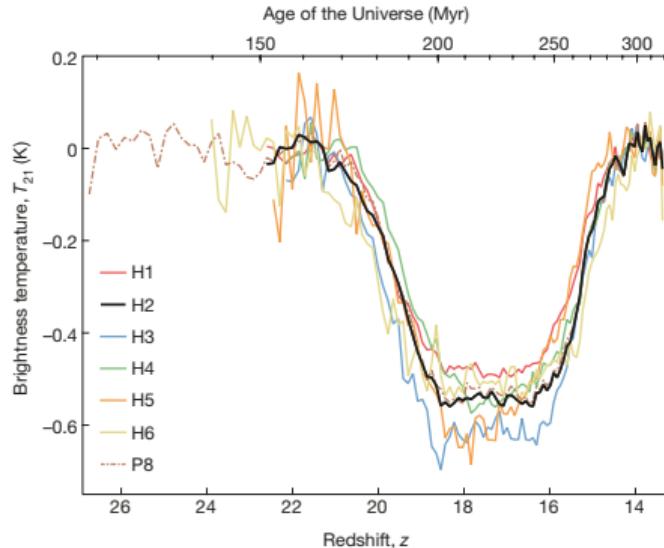
# REACH: Global 21cm cosmology [NatAstro]

- ▶ 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.



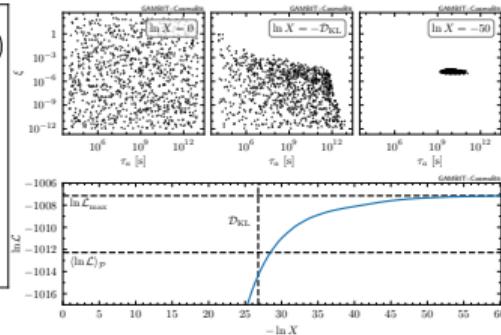
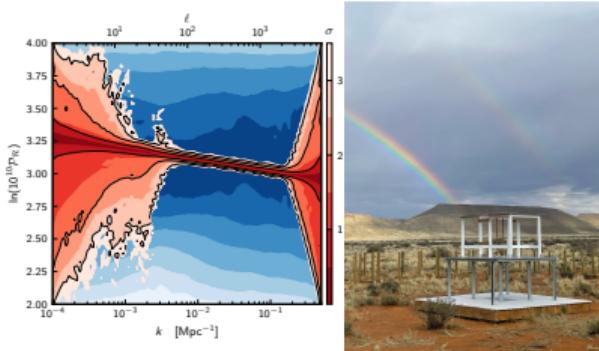
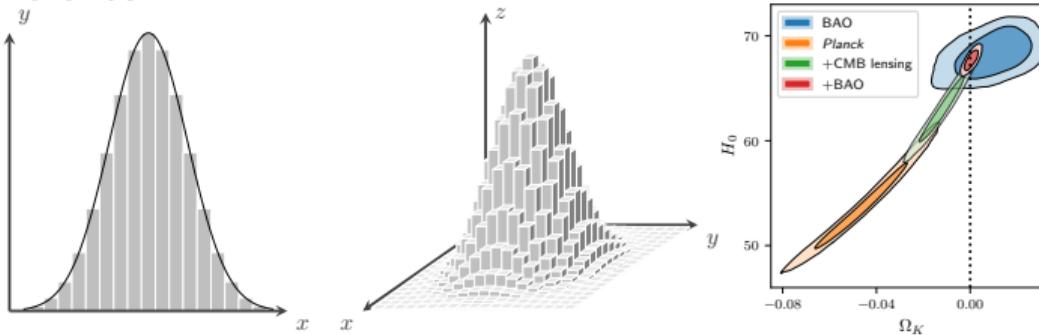
# REACH: Global 21cm cosmology [NatAstro]

- ▶ 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.



# 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 does not require gradients

- ▶ Genetic algorithms
  - + NS discarded points have statistical meaning

## Sampling

- ▶ Metropolis-Hastings?
  - Very little beats a 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 transitions
  - + No annealing schedule tuning
- ▶ Sequential Monte Carlo
  - Some people (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  $n_{\text{live}}$  is a resolution parameter.
  - ▶ Run time is linear in  $n_{\text{live}}$ , 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:  $n_{\text{repeats}}$ .
  - ▶ 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

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

- ▶ Evidence is composed of a “goodness of fit” term and “Occam Penalty” .
- ▶ RHS true for all  $\theta$ . Take max likelihood value  $\theta_*$ :
- ▶ Be more Bayesian and take posterior average to get the “Occam's razor equation”

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

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

- ▶ Natural regularisation which penalises models with too many parameters.

# Kullback Liebler divergence

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

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

- ▶ 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”:

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

(this is exact for top-hat distributions).

