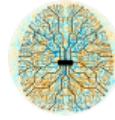


---

# Nested sampling for next generation gravitational-wave inference

Metha Prathaban



# Table of Contents

## Background and context

## Accelerating NS with ML

## BLACKJAX NS



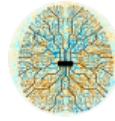
## About Me

- ▶ 3rd year PhD student in Will Handley's group
- ▶ Handley group works on Bayesian statistical inference and artificial intelligence methodologies, with a focus on analyzing complex datasets from next-generation surveys to explore a wide range of physics questions related to dark matter, dark energy, and the early Universe.
- ▶ My work on Bayesian numerical method development in context of GWs

### The Handley group!



+ others!

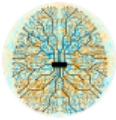


# Table of Contents

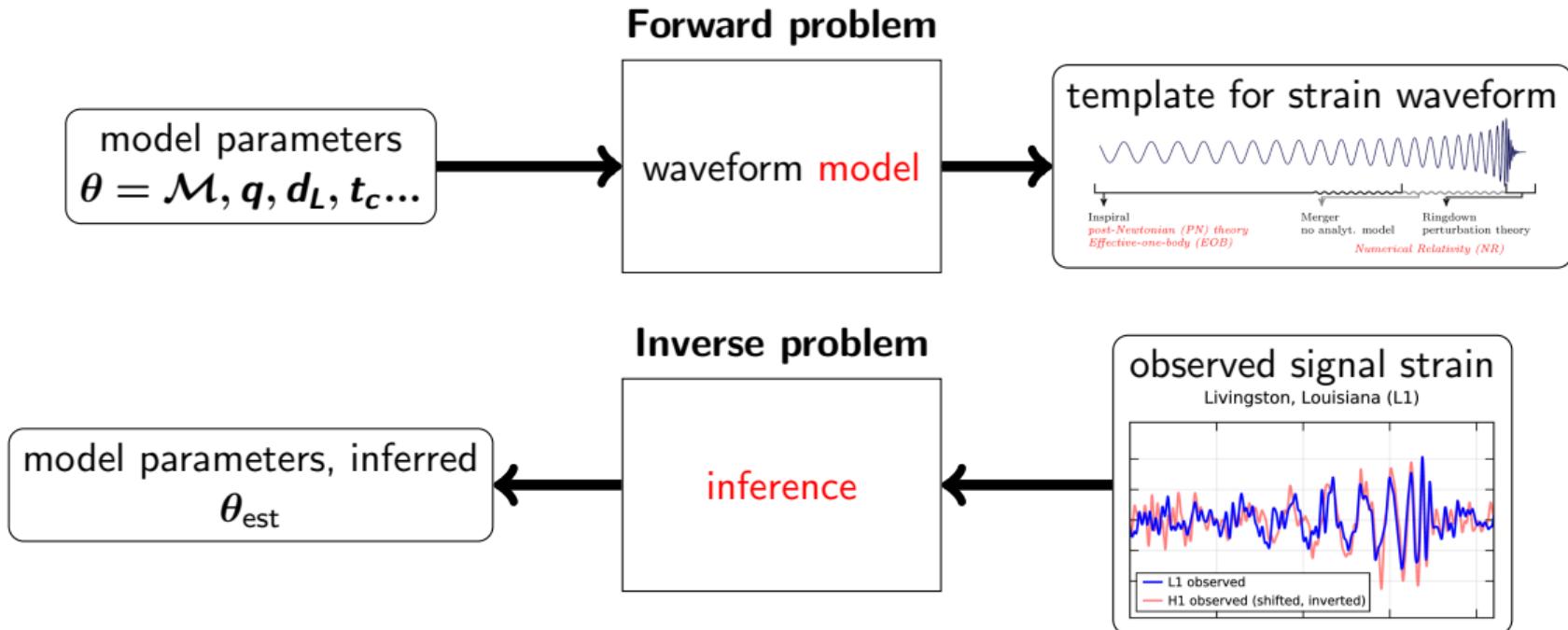
## Background and context

## Accelerating NS with ML

## BLACKJAX NS



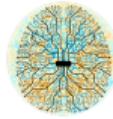
# Inverse problems in GW physics



Given some model  $\mathcal{M}$  and observed signal  $\mathcal{D}$ , Bayes' theorem enables us to relate the **posterior** probability of the set of parameters  $\theta$  which generated the signal to the **likelihood** of the  $\mathcal{D}$  given  $\theta$  and the **prior** probability of  $\theta$  given  $\mathcal{M}$ :

$$P(\theta|\mathcal{D}, \mathcal{M}) = \frac{P(\mathcal{D}|\theta, \mathcal{M})P(\theta|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})} = \frac{\mathcal{L}(\mathcal{D}|\theta)\pi(\theta)}{\mathcal{Z}} \quad (1)$$

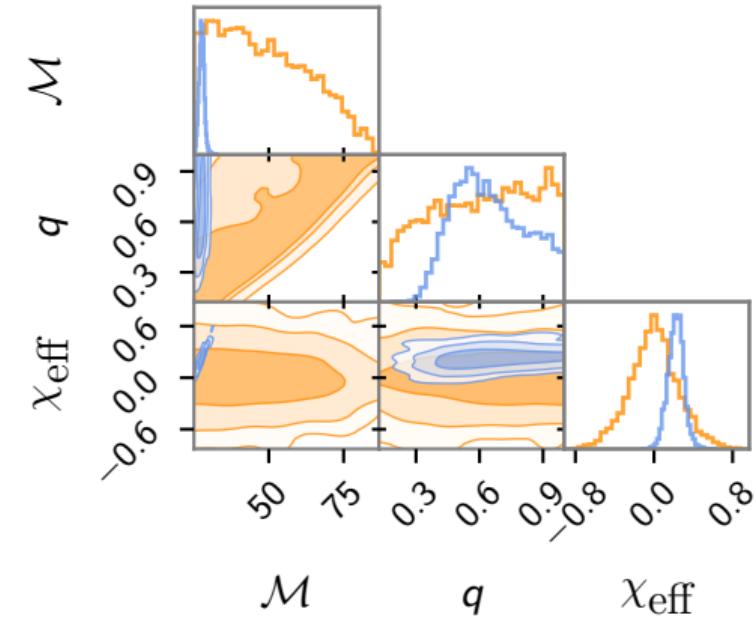
The **evidence**,  $\mathcal{Z}$ , plays a key role in model comparison.



- ▶ Define **prior**, **sample** (unnormalized) **posterior** ( $\mathcal{L}(\mathbf{d}|\theta) \times \pi(\theta)$ ).

Challenges:

- ▶ High-dimensional parameter spaces  $\Rightarrow$  posterior occupies vanishingly small region of prior.
- ▶ Complex waveform models with high costs





Goal is efficient exploration of parameter space, to do GW inference in feasible timescales.

**Posterior** samplers:

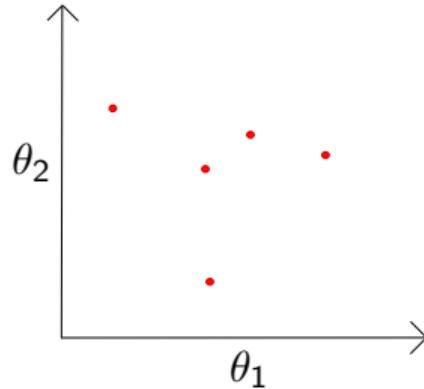
- ▶ Metropolis-Hastings
- ▶ Hamiltonian Monte-Carlo
- ▶ Ensemble samplers

None of these calculate the **evidence**,  $\mathcal{Z}$  - crucial for Bayesian model comparison (e.g. testing for precession vs. no precession)!

## Nested sampling (NS)

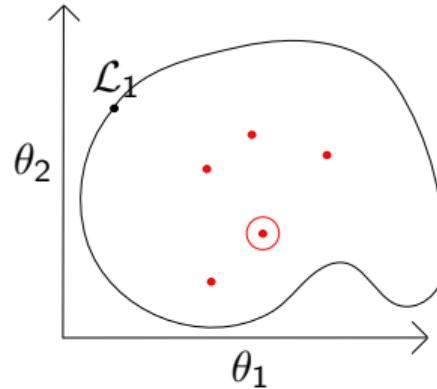
Nested sampling first and foremost calculates **evidence**,  $\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta$ .

- ▶ Prior is populated with set of ‘live points’.

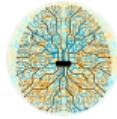


## Nested sampling (NS)

Nested sampling first and foremost calculates **evidence**,  $\mathcal{Z} = \int \mathcal{L}(\theta) \pi(\theta) d\theta$ .

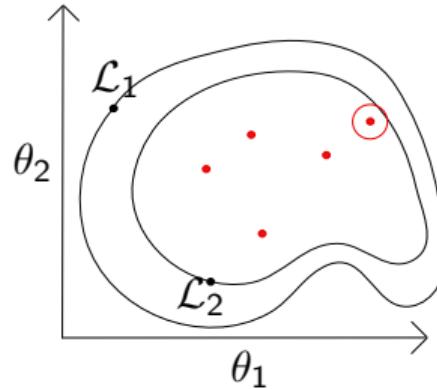


- ▶ Prior is populated with set of ‘live points’.
- ▶ At each iteration  $i$ , point is lowest likelihood is deleted and new live point is drawn, which must have a likelihood higher than that of the deleted point.



## Nested sampling (NS)

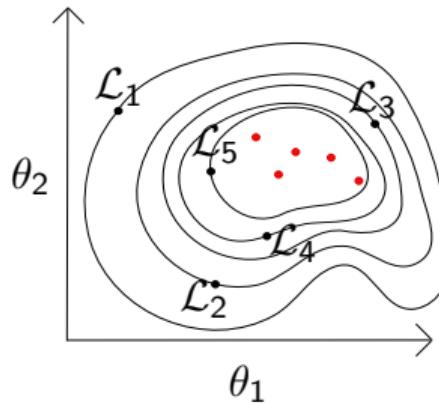
Nested sampling first and foremost calculates **evidence**,  $\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta$ .



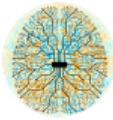
- ▶ Prior is populated with set of 'live points'.
- ▶ At each iteration  $i$ , point is lowest likelihood is deleted and new live point is drawn, which must have a likelihood higher than that of the deleted point.

# Nested sampling (NS)

Nested sampling first and foremost calculates **evidence**,  $\mathcal{Z} = \int \mathcal{L}(\theta) \pi(\theta) d\theta$ .

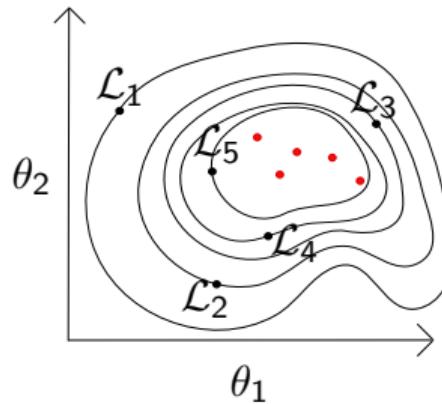


- ▶ Prior is populated with set of ‘live points’.
- ▶ At each iteration  $i$ , point is lowest likelihood is deleted and new live point is drawn, which must have a likelihood higher than that of the deleted point.
- ▶ Live points compress exponentially towards peak of likelihood.



# Nested sampling (NS)

Nested sampling first and foremost calculates **evidence**,  $\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta$ .



- ▶ Prior is populated with set of ‘live points’.
- ▶ At each iteration  $i$ , point is lowest likelihood is deleted and new live point is drawn, which must have a likelihood higher than that of the deleted point.
- ▶ Live points compress exponentially towards peak of likelihood.
- ▶ **Evidence** is calculated as weighted sum over deleted (‘dead’) points.



Time of convergence of NS: 2212.01760

likelihood evaluation time

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (2)$$



Time of convergence of NS: 2212.01760

likelihood evaluation time

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}}$$

resolution

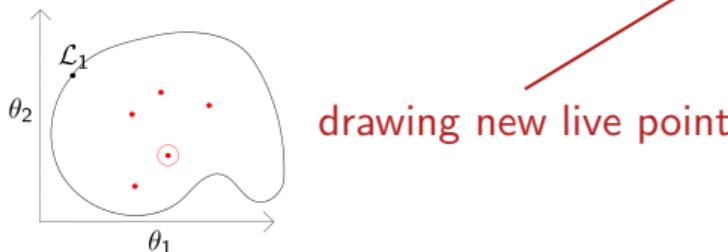
(2)

Time of convergence of NS: 2212.01760

likelihood evaluation time

resolution

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (2)$$



drawing new live point

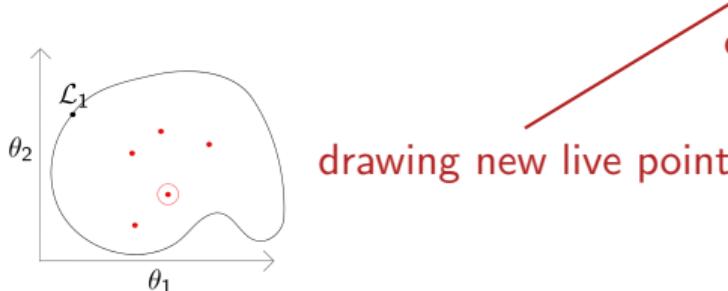
Time of convergence of NS: 2212.01760

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (2)$$

likelihood evaluation time

resolution

compression from prior to posterior ( $\approx \ln \frac{V_{\pi}}{V_{\mathcal{P}}}$ )



drawing new live point



## Current vs. future data

### LVK data so far

- ▶ Short-duration signals from compact binary coalescences (CBCs)
- ▶ End of O4a:  $\approx 200$  confident detections
- ▶ Detection rate:  $\approx 1$  every few days
  - ▶ Signals not overlapping
- ▶ BBH mergers last seconds, BNS mergers last minutes
- ▶ Computationally intensive to analyse, but doable with current pipelines.

### ET + CE

- ▶ Ten-fold improvement in sensitivity compared to 2G detectors
- ▶ Detection rate:  $\approx 10^5 - 10^6$  BBH and  $\approx 7 \times 10^4$  BNS mergers per year
- ▶ Wider frequency band  $\Rightarrow$  more intermediate mass BHs
- ▶ Improved localization
- ▶ Significantly higher SNR
- ▶ Longer signal durations
  - ▶ Overlapping signals expected
- ▶ New GW sources



## Scaling of traditional methods

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

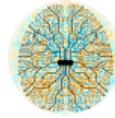


## Scaling of traditional methods

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

higher SNR + longer signal  $\Rightarrow$  narrower posterior



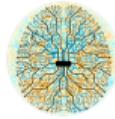
# Scaling of traditional methods

longer signal duration  $\Rightarrow$  slower waveform generation

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

higher SNR + longer signal  $\Rightarrow$  narrower posterior



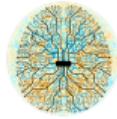
## Scaling of traditional methods

longer signal duration  $\Rightarrow$  slower waveform generation

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

higher SNR + longer signal  $\Rightarrow$  narrower posterior  
overlapping signals  $\Rightarrow$  joint PE



# Scaling of traditional methods

overlapping signals  $\Rightarrow$  joint likelihood

longer signal duration  $\Rightarrow$  slower waveform generation

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

higher SNR + longer signal  $\Rightarrow$  narrower posterior  
overlapping signals  $\Rightarrow$  joint PE

# Scaling of traditional methods

overlapping signals  $\Rightarrow$  joint likelihood

longer signal duration  $\Rightarrow$  slower waveform generation

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

higher SNR + longer signal  $\Rightarrow$  narrower posterior

overlapping signals  $\Rightarrow$  joint PE

duration  $\gtrsim 10\text{mins} \Rightarrow$  extra modeling

# Scaling of traditional methods

non-stationary noise

overlapping signals  $\Rightarrow$  joint likelihood

longer signal duration  $\Rightarrow$  slower waveform generation

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

higher SNR + longer signal  $\Rightarrow$  narrower posterior

overlapping signals  $\Rightarrow$  joint PE

duration  $\gtrsim 10\text{mins} \Rightarrow$  extra modeling

non-stationary noise

# Scaling of traditional methods

non-stationary noise

overlapping signals  $\Rightarrow$  joint likelihood

longer signal duration  $\Rightarrow$  slower waveform generation

$$T_{\text{total}} \propto N_{\text{signals}} \times T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (3)$$

higher sensitivity

higher SNR + longer signal  $\Rightarrow$  narrower posterior

overlapping signals  $\Rightarrow$  joint PE

duration  $\gtrsim 10\text{mins} \Rightarrow$  extra modeling

non-stationary noise

**Billions to quadrillions of CPU core hours for 1 month of ET data...** 2412.02651

## Heterodyning

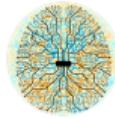
Exploits the smooth changes in GW waveforms near the maximum likelihood point.  
Assigns fiducial source parameter close to the true value and uses coarser frequency resolution.

## Reduced Order Quadrature

Approximates the waveform model as a linear combination of a small number of pre-trained bases evaluated at representative frequencies.

## Multibanding

Adjusts data resolution for gravitational wave signals across frequency bands, particularly during low-frequency early inspiral stage of CBCs.



## Established acceleration attempts

### Heterodyning

Exploits the smooth changes in GW waveforms near the maximum likelihood point.  
Assigns fiducial source parameter close to the true value and uses coarser frequency resolution.

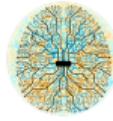
### Reduced Order Quadrature

Approximates the waveform model as a linear combination of a small number of pre-trained bases evaluated at representative frequencies.

### Multibanding

Adjusts data resolution for gravitational wave signals across frequency bands, particularly during low-frequency early inspiral stage of CBCs.

**Even with these, millions of CPU core hours for 1 month of ET data.** 2412.02651

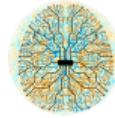


# Table of Contents

Background and context

Accelerating NS with ML

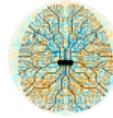
BLACKJAX NS



# ML for traditional methods

ROM, surrogates

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (4)$$

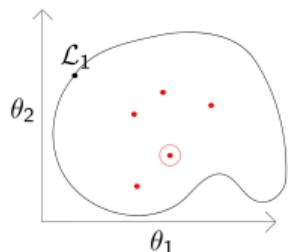


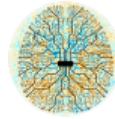
# ML for traditional methods

ROM, surrogates

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (4)$$

NESSAI





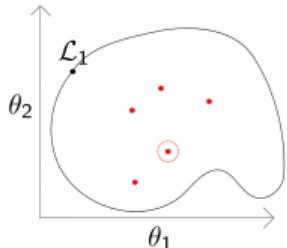
# ML for traditional methods

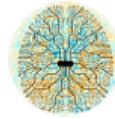
ROM, surrogates

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (4)$$

not as baked-in as you might think!

NESSAI





# ML for traditional methods

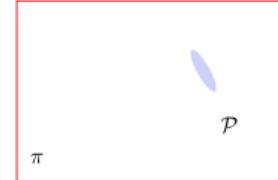
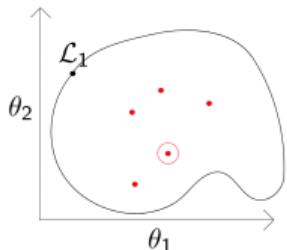
ROM, surrogates

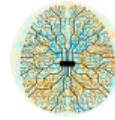
$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}}$$

(4)

focus of this section

NESSAI





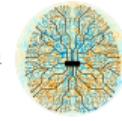
## Time of convergence of NS

$$T \propto T_{\mathcal{L}} \times f_{\text{sampler}} \times D_{\text{KL}} \times n_{\text{live}} \quad (5)$$

## Uncertainty in $\log \mathcal{Z}$

$$\sigma \propto \sqrt{D_{\text{KL}} / n_{\text{live}}} \quad (6)$$

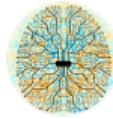
Precision-normalized runtime has quadratic dependence on KL divergence. 2212.01760



REACH

One way to do this (REACH):



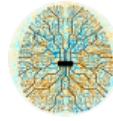


# Reducing $D_{\text{KL}}$

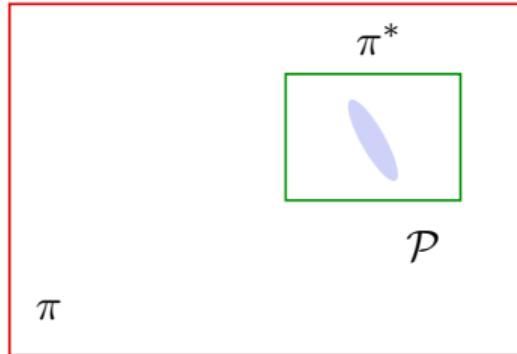
One way to do this (REACH):

- ▶ Perform low resolution (low live points) run first to roughly identify where **posterior** lies.

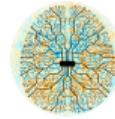




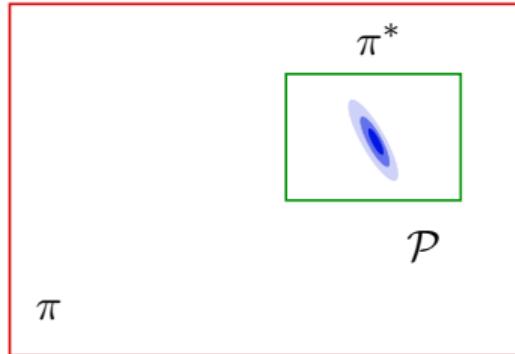
One way to do this (REACH):



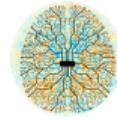
- ▶ Perform low resolution (low live points) run first to roughly identify where **posterior** lies.
- ▶ Then set off second, high resolution, run with **narrower** box **prior** (much quicker).



One way to do this (REACH):

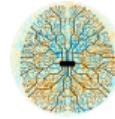


- ▶ Perform low resolution (low live points) run first to roughly identify where **posterior** lies.
- ▶ Then set off second, high resolution, run with **narrower** box **prior** (much quicker).
- ▶ **Evidence** has **changed** (since different prior), but easy to correct (multiply new evidence by  $\frac{V_{\pi^*}}{V_\pi}$ )

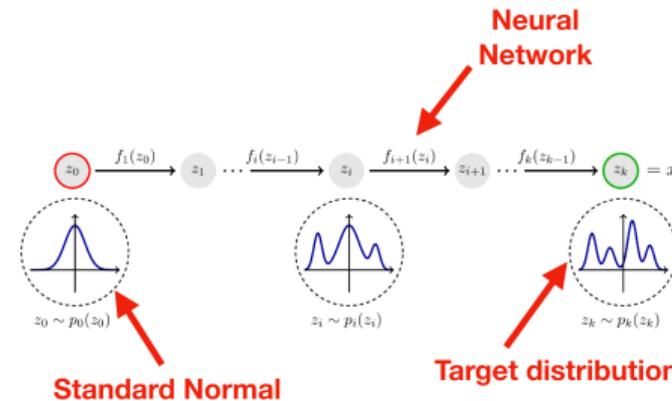


NFs

- ▶ Can iterate on this by using **normalizing flows** (NF) to learn the rough **posterior**.

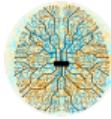


- ▶ Can iterate on this by using **normalizing flows** (NF) to learn the rough **posterior**.
- ▶ NFs perform density estimation, by learning a series of invertible mappings from the standard normal distribution to the target (posterior).





- ▶ Use **normalizing flows** (NF) to learn the rough **posterior**, and use this as our updated prior,  $\pi^*$ .
- ▶ In this case, can't do our trick of correcting the second **evidence** by volume ratio,  $\frac{V_{\pi^*}}{V_\pi}$ !
- ▶ Must rely on another technique to get around this!



- ▶ Use **normalizing flows** (NF) to learn the rough **posterior**, and use this as our updated prior,  $\pi^*$ .
- ▶ In this case, can't do our trick of correcting the second **evidence** by volume ratio,  $\frac{V_{\pi^*}}{V_\pi}$ !
- ▶ Must rely on another technique to get around this!

Posterior repartitioning (PR) can help us with this! (see e.g. 2212.01760)

Bayesian Analysis (0000)

00, Number 0, pp. 1

## Bayesian posterior repartitioning for nested sampling

Xi Chen<sup>\*†</sup>, Farhan Feroz<sup>‡</sup> and Michael Hobson<sup>†</sup>

Improving the efficiency and robustness of nested sampling using posterior repartitioning

Xi Chen · Michael Hobson · Saptarshi Das · Paul Gelderblom

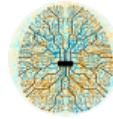


Article

SuperNest: accelerated nested sampling applied to astrophysics and cosmology<sup>†</sup>

Aleksandr Petrosyan<sup>1,2,3†</sup> & Will Handley<sup>1,2,4†</sup>





## Posterior repartitioning (PR)

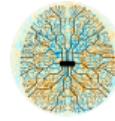
- Evidence and posterior only depend on product of  $\mathcal{L}$  and  $\pi$ :

$$\mathcal{Z} = \int \mathcal{L}(\theta) \pi(\theta) d\theta \quad (7)$$

$$\mathcal{P}(\theta) = \frac{\mathcal{L}(\theta) \pi(\theta)}{\mathcal{Z}} \quad (8)$$

We are free to redefine the likelihood and prior however we like - as long as the product is the same! arXiv:1908.04655

$$\tilde{\mathcal{Z}} = \int \tilde{\mathcal{L}}(\theta) \tilde{\pi}(\theta) d\theta = \int \mathcal{L}(\theta) \pi(\theta) d\theta = \mathcal{Z} \quad (9)$$

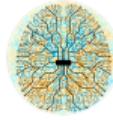


- ▶ Many sampling algorithms do not distinguish between  $\mathcal{L}$  and  $\pi$  at the algorithmic level.
- ▶ e.g. Metropolis-Hastings acceptance ratio only depends on the **joint distribution**,  $\mathcal{L}(\theta)\pi(\theta)$ .
- ▶ Nested sampling does distinguish between prior and likelihood at the algorithmic level, by '**sampling from the prior  $\pi$ , subject to the hard likelihood constraint,  $\mathcal{L}$** '.
- ▶  $\mathcal{Z}$  and  $\mathcal{P}$  will not change if we repartition  $\mathcal{L}$  and  $\pi$ , **but  $\mathcal{D}_{KL}$  will**.



# PR-NS w/ NFs

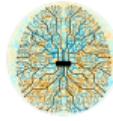
$$\pi(\theta) \longrightarrow \text{NF}(\theta)$$



# PR-NS w/ NFs

$$\pi(\theta) \longrightarrow \text{NF}(\theta)$$

$$\mathcal{L}(\theta) \longrightarrow \frac{\mathcal{L}(\theta)\pi(\theta)}{\text{NF}(\theta)}$$



PR-NS w/ NFs

$$\pi(\theta) \longrightarrow \text{NF}(\theta)$$

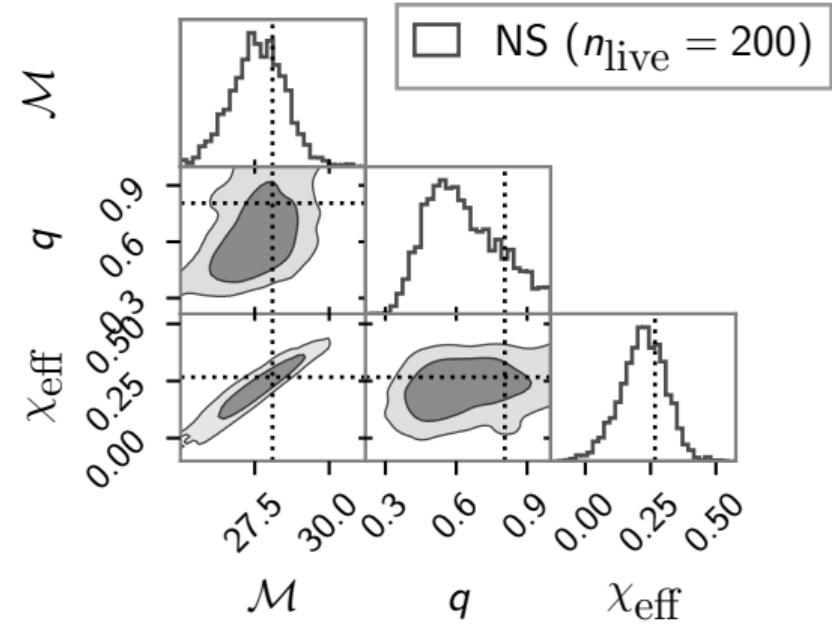
$$\mathcal{L}(\theta) \longrightarrow \frac{\mathcal{L}(\theta)\pi(\theta)}{\text{NF}(\theta)}$$

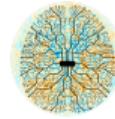
$$\mathcal{D}_{\text{KL}} \approx \log \frac{V_{\text{NF}}}{V_{\mathcal{P}}}$$



## Demo on simulated example

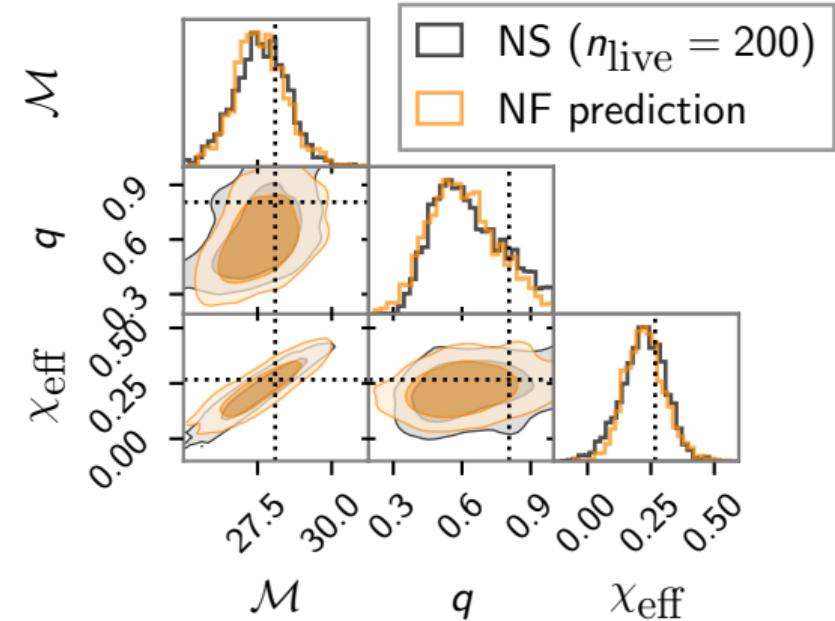
- ▶ Perform low resolution run on simulated data.

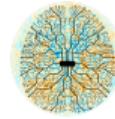




## Demo on simulated example

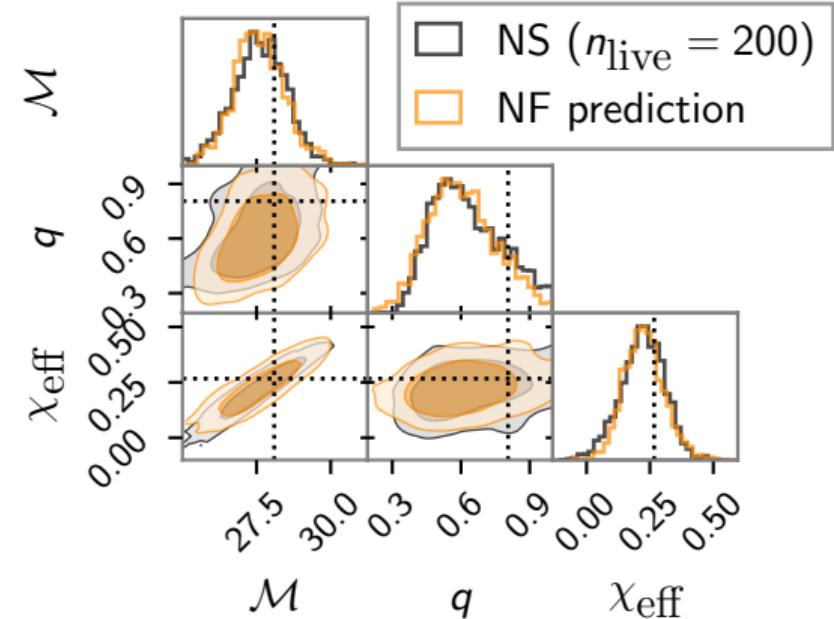
- ▶ Perform low resolution run on simulated data.
- ▶ Train NF on the weighted samples.

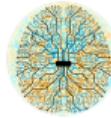




## Demo on simulated example

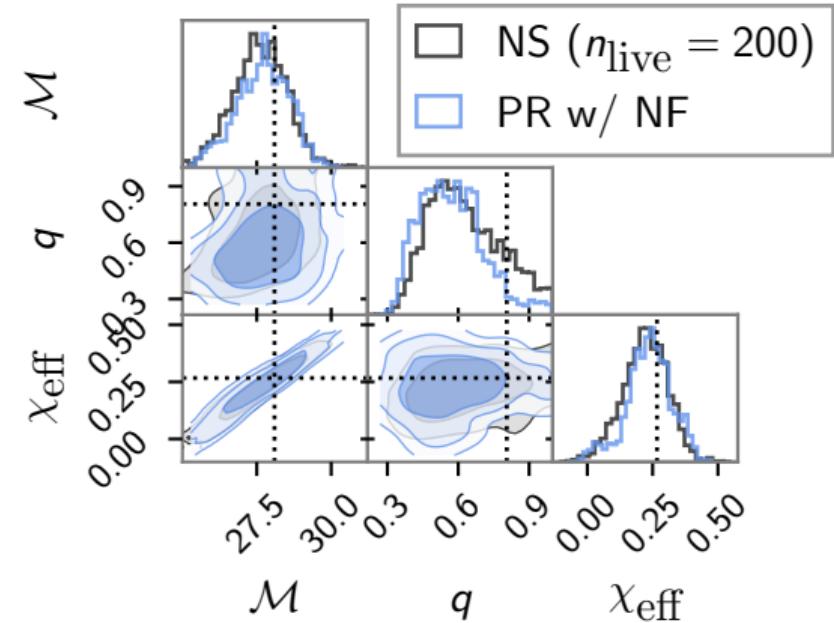
- ▶ Perform low resolution run on simulated data.
- ▶ Train NF on the weighted samples.
- ▶ Use this as ‘repartitioned prior’ for new high resolution run (using PR to also update likelihood accordingly to same evidences and posteriors out).

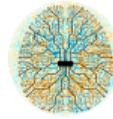




## Demo on simulated example

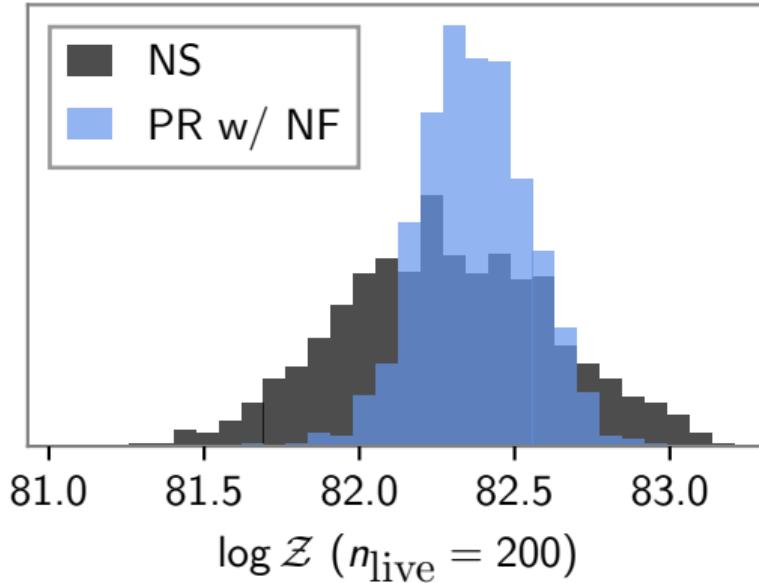
- ▶ Perform low resolution run on simulated data.
- ▶ Train NF on the weighted samples.
- ▶ Use this as ‘repartitioned prior’ for new high resolution run (using PR to also update likelihood accordingly to same evidences and posteriors out).



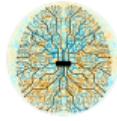


## Demo on simulated example

- ▶ Perform low resolution run on simulated data.
- ▶ Train NF on the weighted samples.
- ▶ Use this as ‘repartitioned prior’ for new high resolution run (using PR to also update likelihood accordingly to same evidences and posteriors out).



Same answer as doing a full resolution pass of NS, but **7x faster** (precision-normalized).

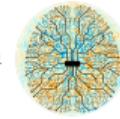


# Table of Contents

Background and context

Accelerating NS with ML

BLACKJAX NS



# CPUs & GPUs in Scientific Computing

## ► CPUs (Central Processing Units):

- ▶ Optimized for a wide range of tasks.
- ▶ Equipped with a limited number of powerful cores.
- ▶ Excel at sequential, complex calculations and intricate decision-making.
- ▶ Ideal for tasks featuring non-uniform workloads and conditional branching.

## ► GPUs (Graphics Processing Units):

- ▶ Architectured with thousands of smaller, efficient cores.
- ▶ Tailored for massive parallel arithmetic operations.
- ▶ Best suited for repetitive, data-parallel tasks.
- ▶ Increasingly used in scientific computing.



# Advantages of GPUs for GW inference

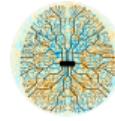
## ► Likelihood Evaluation:

- ▶ Frequency bins are independent, so waveforms can be evaluated across many bins in parallel.
- ▶ On Nvidia A100 GPU, we can evaluate the waveform model  $\approx O(10^9)$  times in a second for different frequencies or source parameters [2302.05333].

## ► Sampling:

- ▶ Multiple chains can be executed in parallel.
- ▶ For NS, if using a slice sampler, can run many slicing operations at once.

Parallel processing power of GPUs can speed up Bayesian inference by factor of thousands.



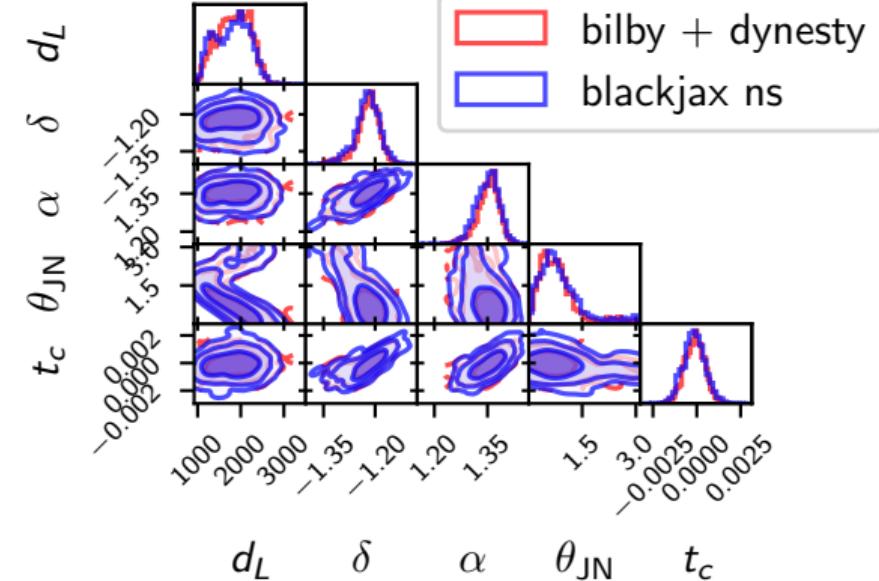
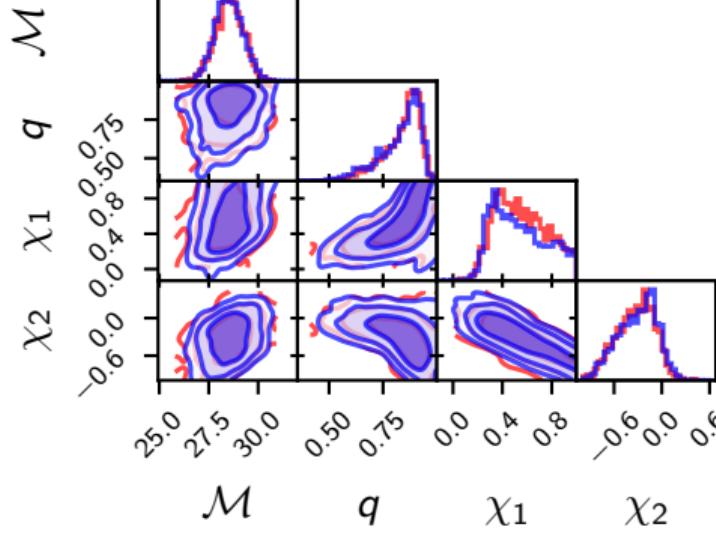
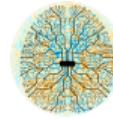
- ▶ RIPPLE: JAX implementation of models for GPU-accelerated waveform calls
- ▶ JIM uses RIPPLE + FLOWMC

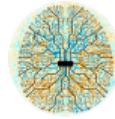


- ▶ RIPPLE: JAX implementation of models for GPU-accelerated waveform calls
- ▶ JIM uses RIPPLE + FLOWMC

### BLACKJAX NS

- ▶ Nested sampling implementation in BLACKJAX, developed by David Yallup.
- ▶ 11 parameter nested sampling run on GW150914 in minutes





# Future of scientific computing

## ► Shift in Hardware Paradigm:

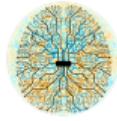
- ▶ Scientific computing is increasingly relying on architectures that favor parallel operations.
- ▶ Computing power will become heavily skewed towards GPUs.

## ► Implications for Research:

- ▶ GPUs are becoming the norm in many areas, including astrophysics.
- ▶ Adapting our algorithms to modern hardware is essential to keep pace with the cutting edge of computing power.

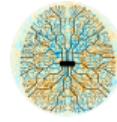
## ► For GWs:

- ▶ Need to get more waveforms models onto GPUs!



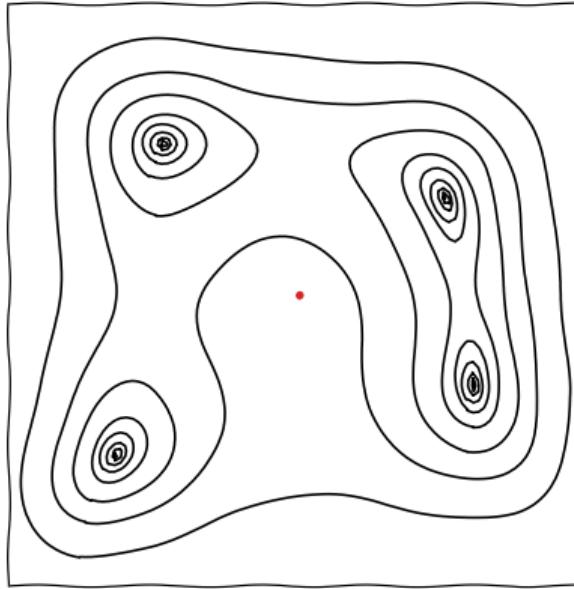
## Conclusions

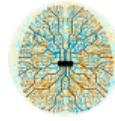
- ▶ 3G + LISA era presents significant challenges for data analysis
- ▶ Current standard inference methods, including traditional nested sampling, will not scale well
- ▶ ML-accelerated sampling and GPU-accelerated pipelines work towards addressing these challenges, and offer a complementary approach to SBI



# MCMC vs. NS

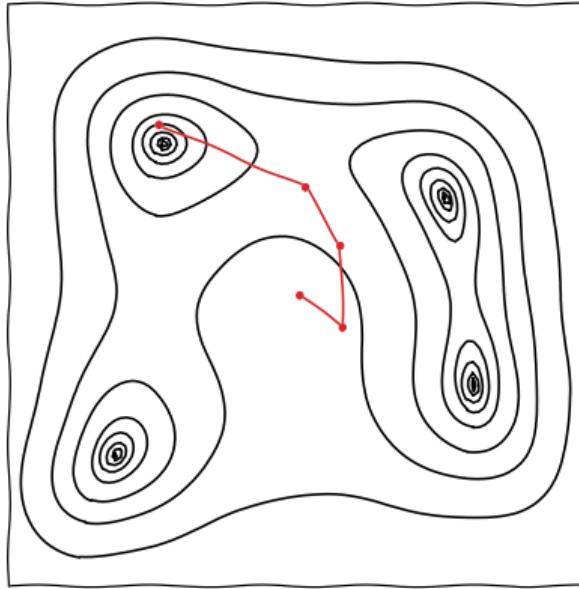
## MCMC

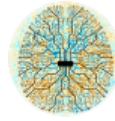




# MCMC vs. NS

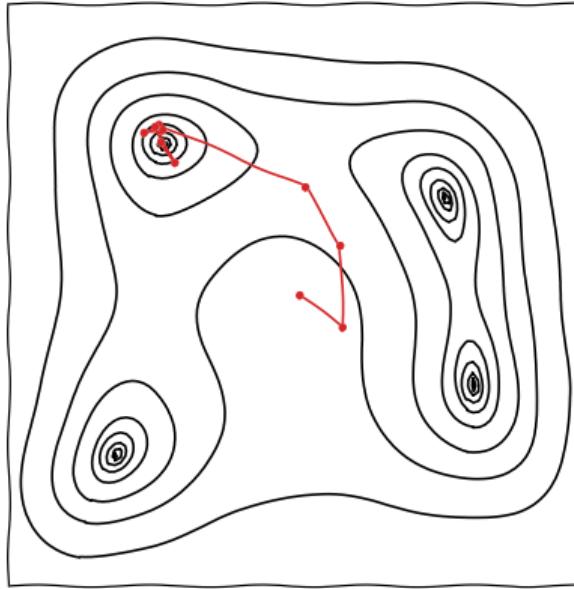
## MCMC

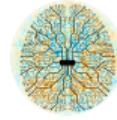




# MCMC vs. NS

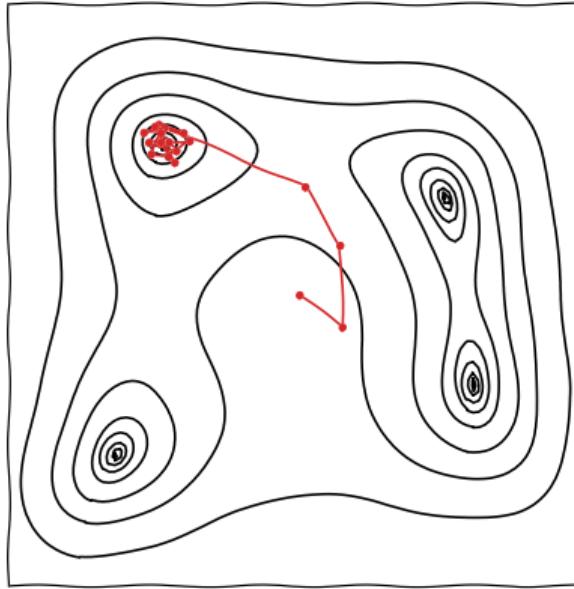
## MCMC

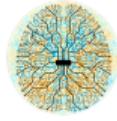




# MCMC vs. NS

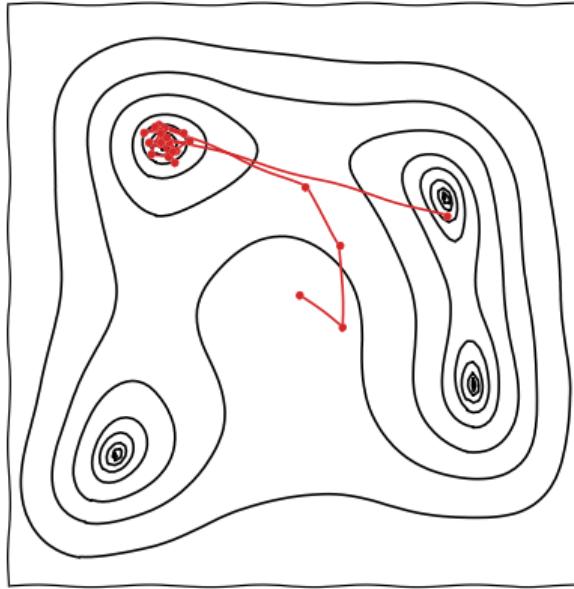
## MCMC

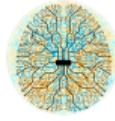




# MCMC vs. NS

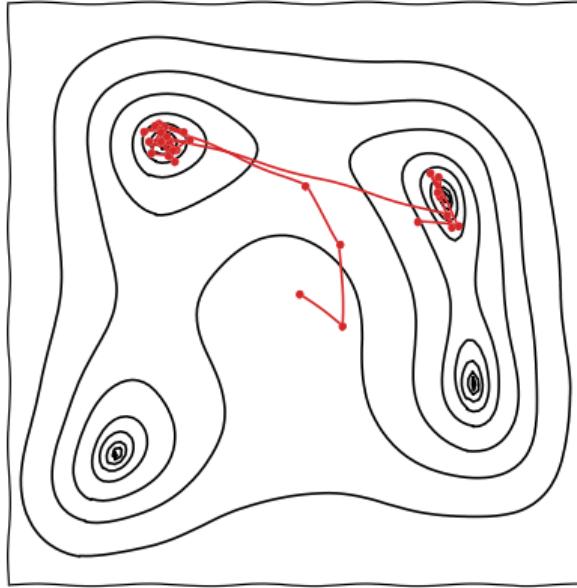
## MCMC

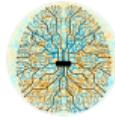




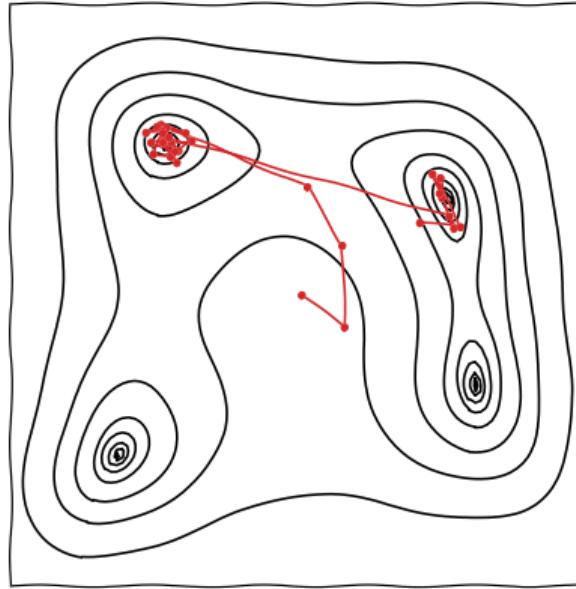
# MCMC vs. NS

## MCMC

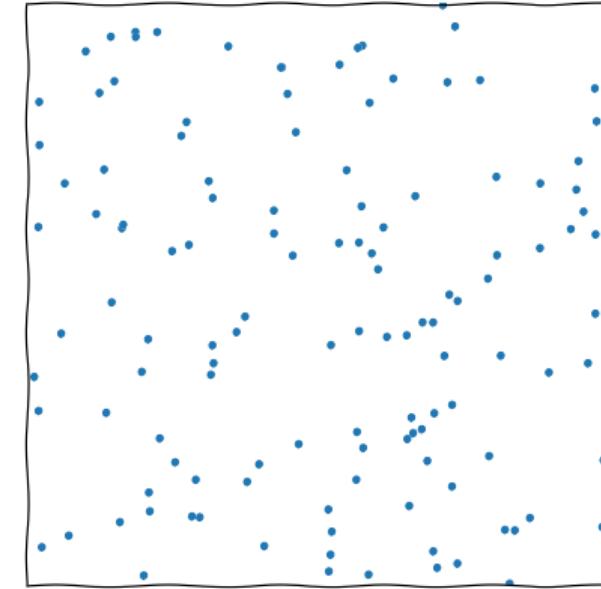


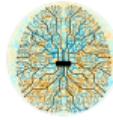


## MCMC

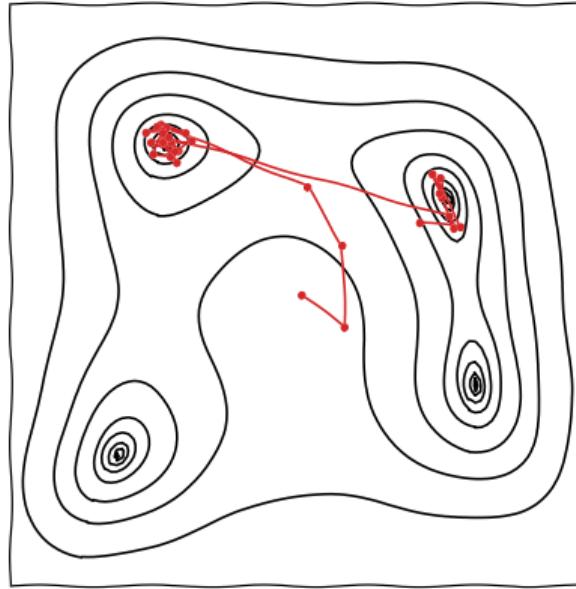


## Nested sampling

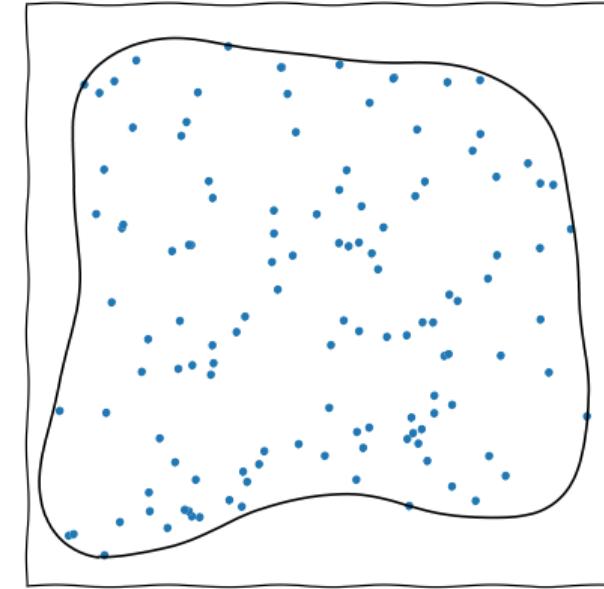


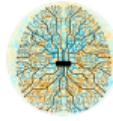


## MCMC

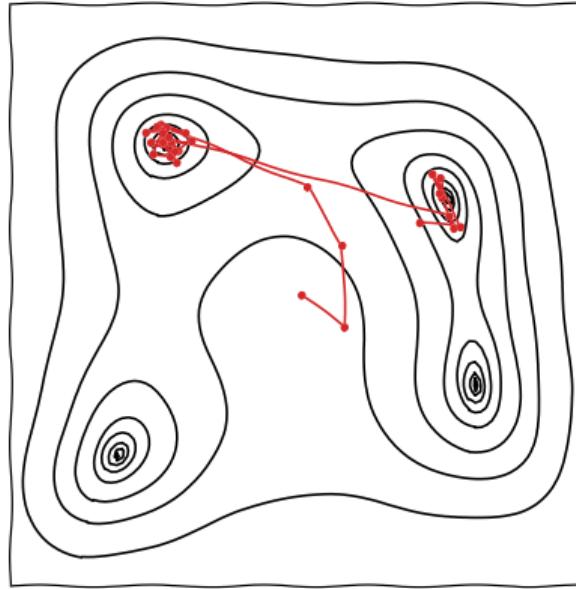


## Nested sampling

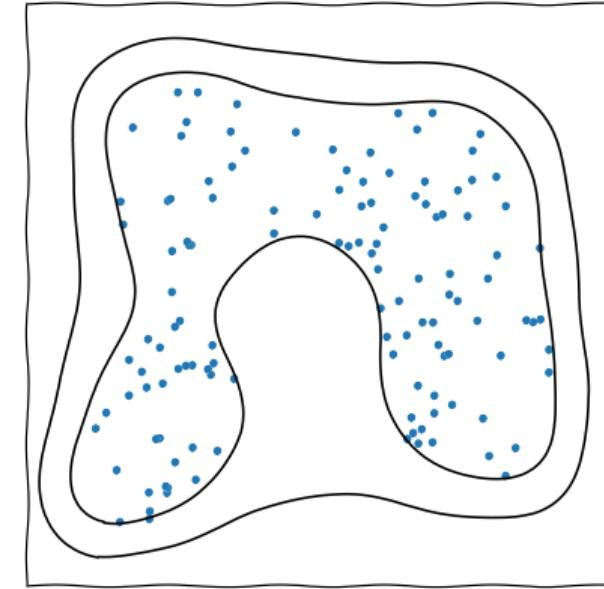


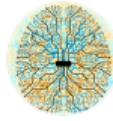


## MCMC

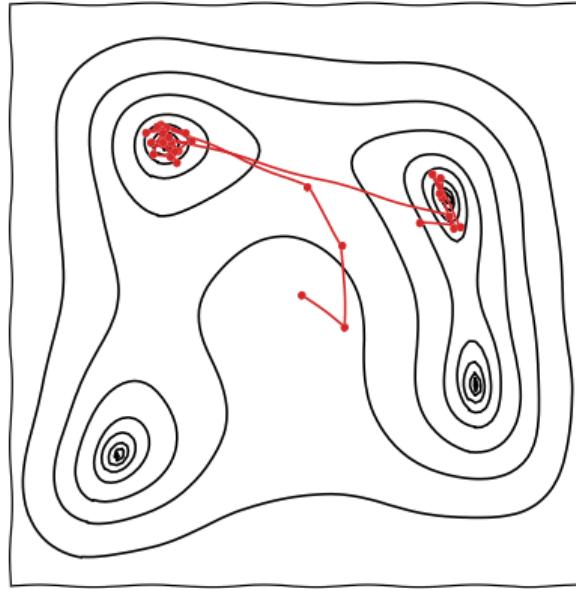


## Nested sampling

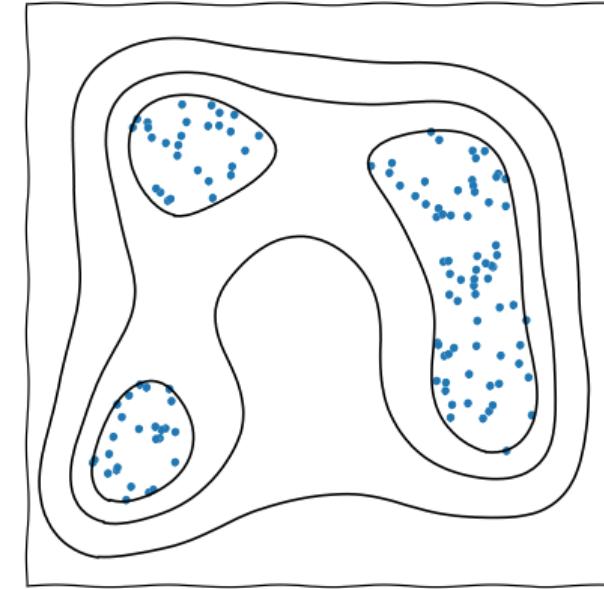


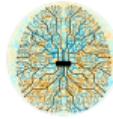


## MCMC

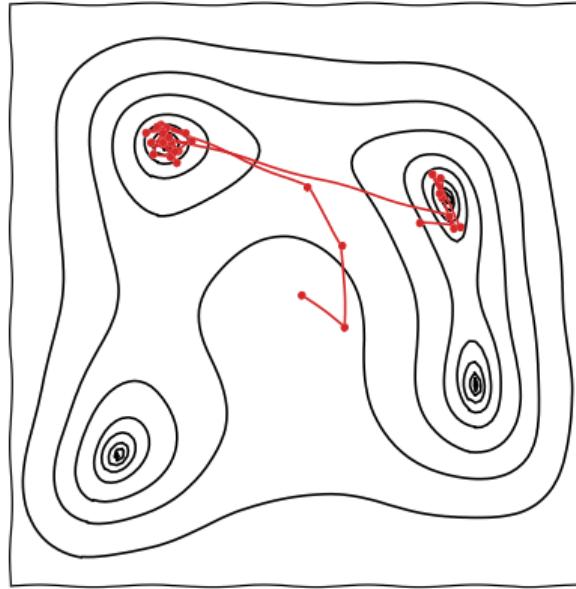


## Nested sampling

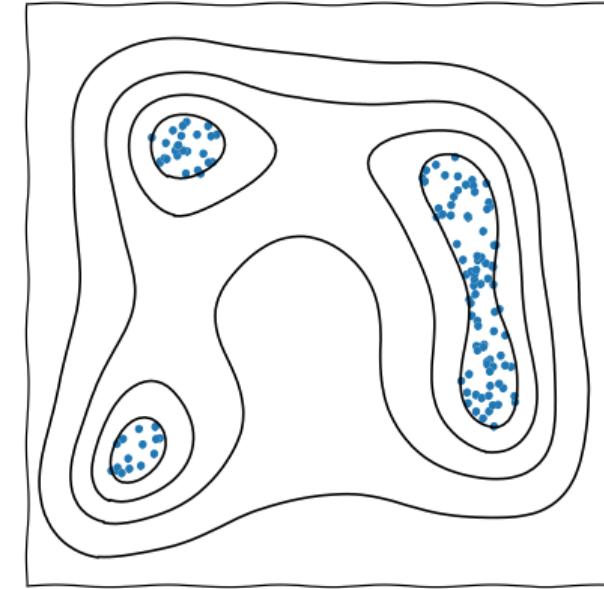


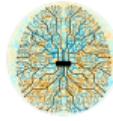


## MCMC

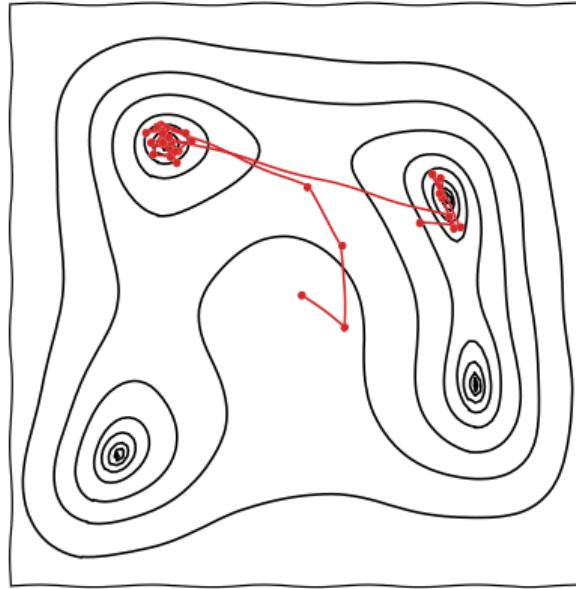


## Nested sampling

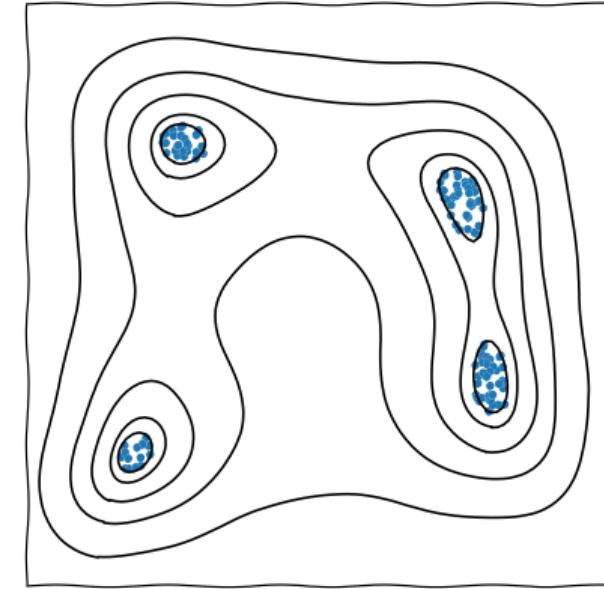


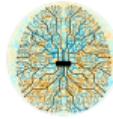


## MCMC

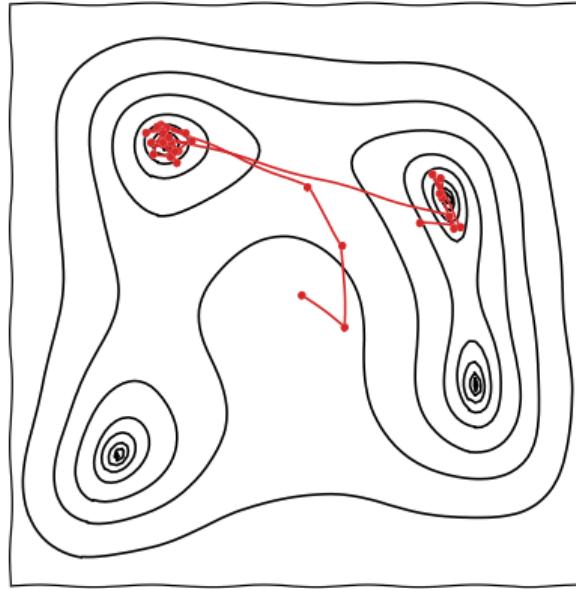


## Nested sampling

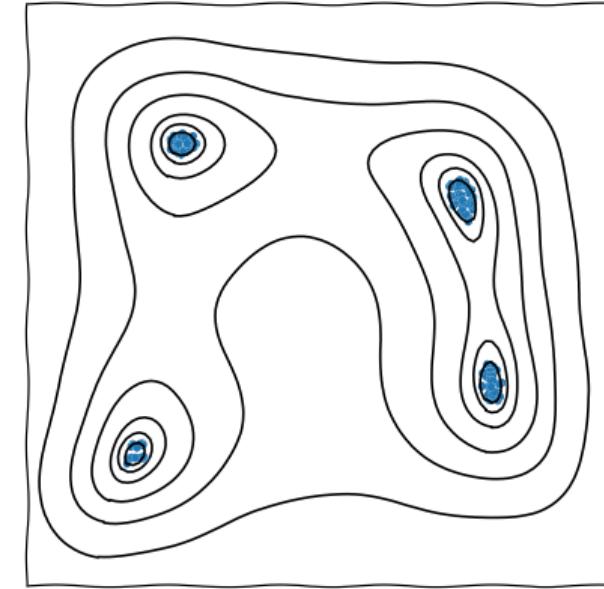


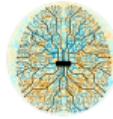


## MCMC

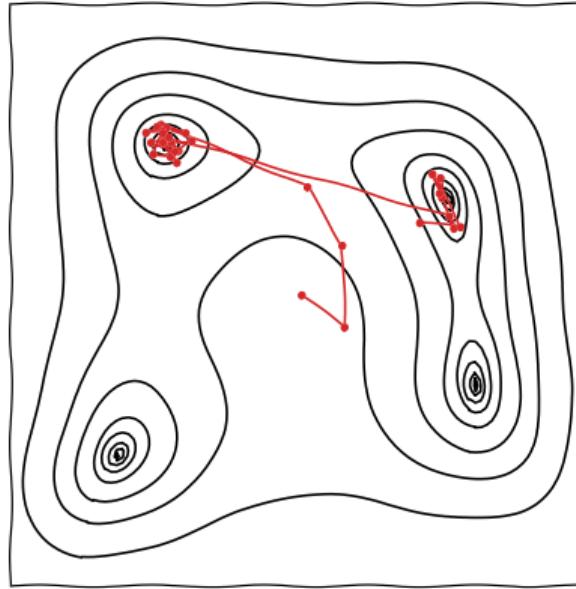


## Nested sampling

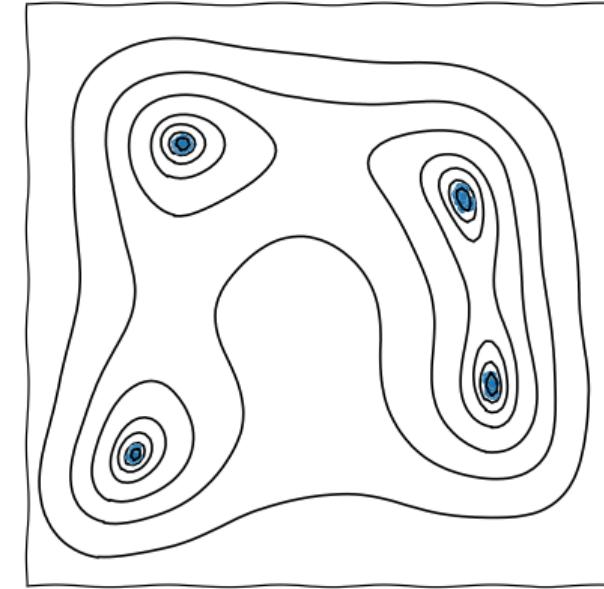


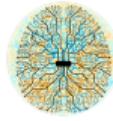


## MCMC

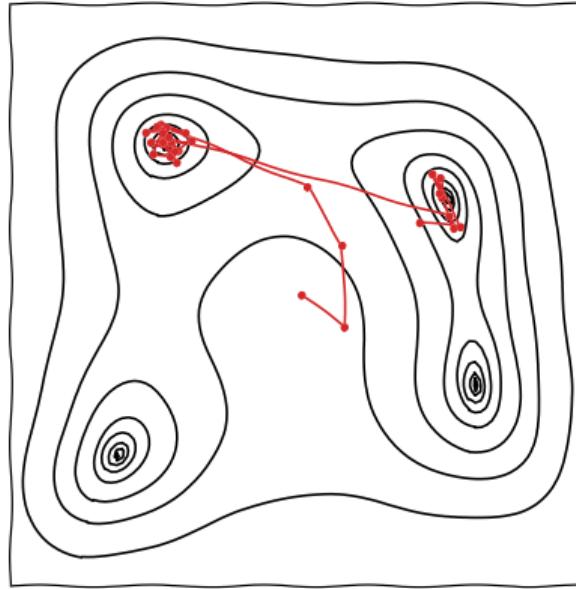


## Nested sampling

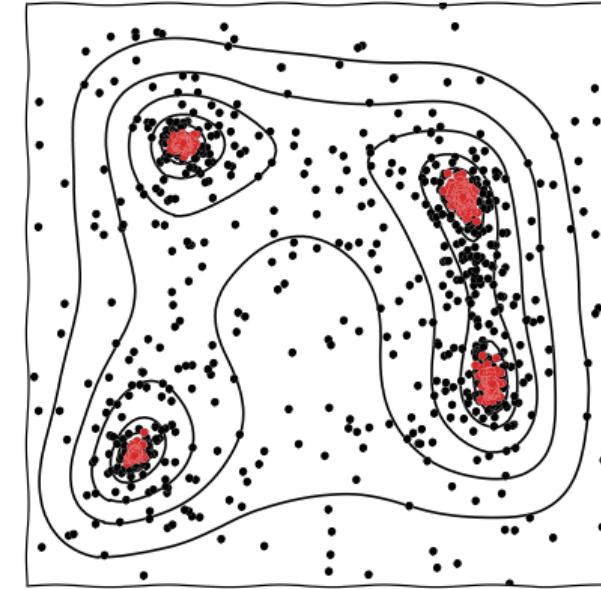


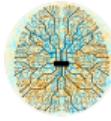


## MCMC



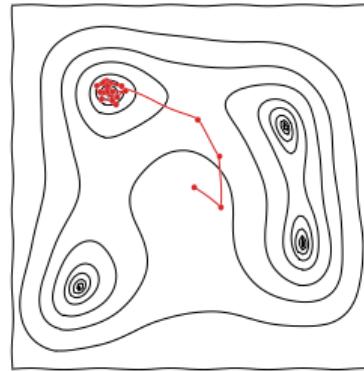
## Nested sampling





MCMC vs. NS

MCMC



Nested sampling

