Advanced topics in Bayesian Statistics

Martin Völkl

Universität Heidelberg

2022-02-03

Statistical Methods in Particle Physics

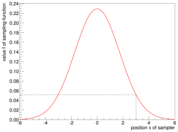
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Reminder: Bayesian analysis

- Bayes' theorem:

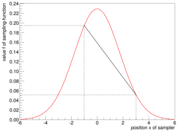
$$p(\vec{\lambda}|\vec{d}) = \frac{p(\vec{d}|\vec{\lambda})\ p(\vec{\lambda})}{\int \mathrm{d}\vec{\lambda}\ p(\vec{d}|\vec{\lambda})\ p(\vec{\lambda})}$$

Reminder: Markov-Chain Monte Carlo



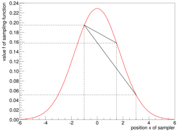
$x_{start} = 3$

Figure 1.5: A starting point is chosen.



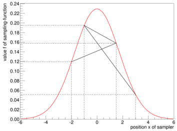
$$
\begin{aligned}
x_{old} &= x_{start} = 3 \\
p &= -4 \\
x_{new} &= x_{old} + p = -1 \\
\rho &= \min\left(1, \frac{f(x_{new})}{f(x_{old})}\right) = \min(1, 3.75) = 1 \\
u &= 0.9 \\
\rho &> u \Rightarrow \textbf{accept}
\end{aligned}
$$

Figure 1.6: For $f(x_{new}) > f(x_{old})$ the step is always accepted.



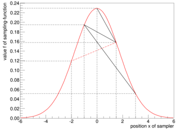
$$
\begin{aligned}
x_{old} &= -1 \\
p &= 2.5 \\
x_{new} &= x_{old} + p = 1.5 \\
\rho &= \min\left(1, \frac{f(x_{new})}{f(x_{old})}\right) = \min(1, 0.81) = 0.81 \\
u &= 0.4 \\
\rho &> u \Rightarrow \textbf{accept}
\end{aligned}
$$

Figure 1.7: For $f(x_{new}) < f(x_{old})$ it depends on u whether a step is accepted.



$$
\begin{aligned}
x_{old} &= 1.5 \\
p &= -3.5 \\
x_{new} &= x_{old} + p = -2 \\
\rho &= \min\left(1, \frac{f(x_{new})}{f(x_{old})}\right) = \min(1, 0.75) = 0.75 \\
u &= 0.8 \\
\rho &< u \Rightarrow \textbf{reject}
\end{aligned}
$$

Figure 1.8: For $\rho < u$ the step is rejected.



$$
\begin{aligned}
x_{old} &= 1.5 \\
p &= -1.5 \\
x_{new} &= x_{old} + p = 0 \\
\rho &= \min\left(1, \frac{f(x_{new})}{f(x_{old})}\right) = \min(1, 1.45) = 1 \\
u &= 0.6 \\
\rho &> u \Rightarrow \textbf{accept}
\end{aligned}
$$

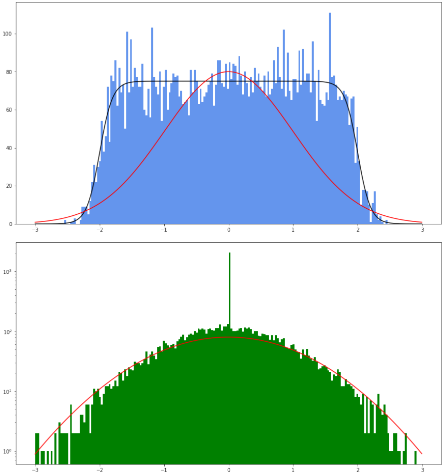
Figure 1.9: For $\rho > u$ the step is accepted again.

from Bachelor's thesis of Manuel Wittner

- Each new step depends only on previous point
- Distributed as true distribution in limit of infinite steps
- How do we know it has converged?

Has the chain converged?

- Simple example: Use normal distribution for proposal; try sampling from Fermi-distribution
- We have the freedom to select the width (and shape) of the proposal function – how to find a good value?
- Does it matter which starting point we use? And how can we find a good one?

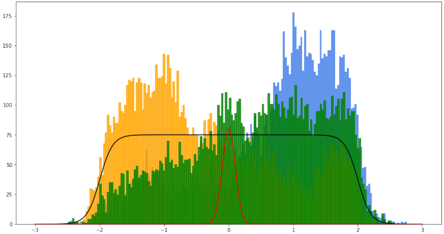


Proposal and actual steps

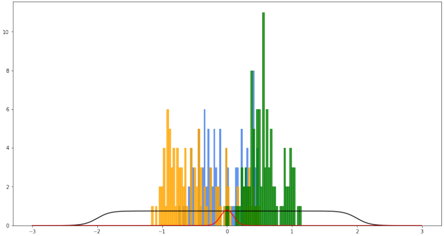
Comparing several outputs

- In example: Repeating MCMC several times clearly gives different results
- This is one way to assess convergence
- It is also costly; we would prefer to have 3x as much statistics in the chain we actually use instead
- For few iterations: chain does not traverse the entire distribution
- Random walk: standard deviation $\sim \sigma_{proposal}\sqrt{N}$
- To converge, the chain should traverse the distribution many times
- If sitze of distribution is σ_d, then this means:

$$\sigma_{proposal}\sqrt{N} \gg \sigma_d$$



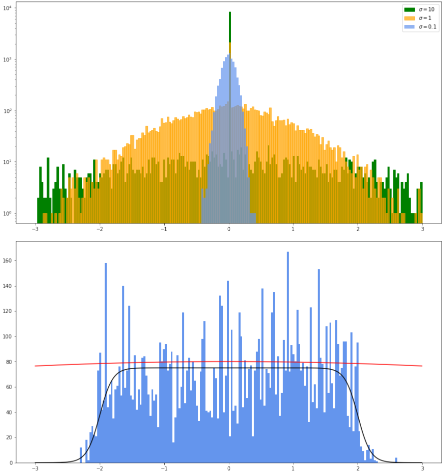
10000 steps



100 steps

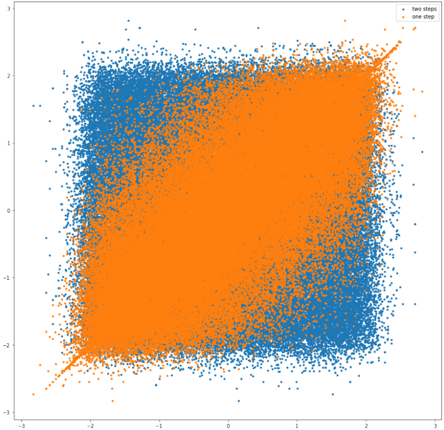
Large steps sizes

- Making the proposal steps very wide is also not good: The step often goes to regions of low probability – and is rejected
- This causes the chain to stay at one position for long amounts of time, convergence gets worse
- Somewhere in the middle there is a sweet spot
- In both cases, the problem is that sometimes we do not move a lot through the distribution – how to quantify?



Autocorrelation

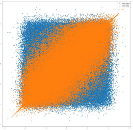
- Compare the position at step i to the position at step $i+1$, put in diagram
- Obviously not independent
- For additional step – less correlation
- On diagonal: Cases where step is rejected
- Quantify with Pearson coefficient
- If fully independent – coefficient is 0
- Now check what happens for different proposal step sizes

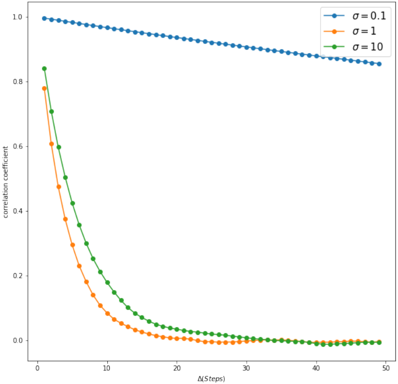


$\sigma = 1$

Autocorrelation (2)

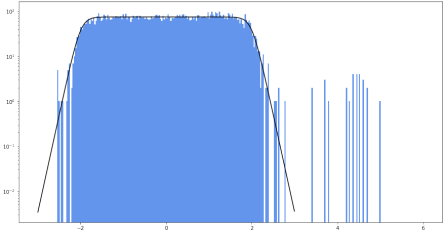
- Autocorrelation function shows how quickly sample points become independent
- The faster this happens, the better the proposal function
- Good proposal step sizes typically have around 50% acceptance ratio for $d \leq 2$ and about 25% for higher dimensionality
- Intermediate case is best of the ones tried here – almost independent after 10 steps
- To get an essentially uncorrelated sample, sometimes only every n^{th} step is used for the analysis
- This does throw away statistics though and is not generally recommended
- If N_{indep} is the amount of steps needed for the correlation to go down to near 0, then a neccessary condition for convergence is $N_{steps} \gg N_{indep}$





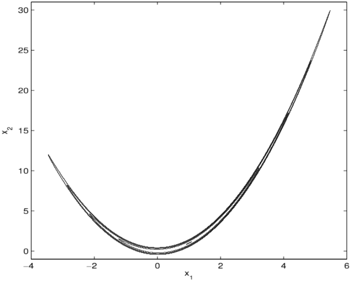
Burn-In

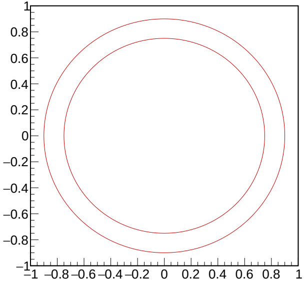
- Previously: Starting value at 0
- If it is at +5, then points always appear there even though pdf is very small
- Can cause problems when calculating distribution moments, credible intervals etc.
- Will always converge correctly but might take a very long time
- Testing and retesting starting value difficult in high dimensions
- However: Most of the points in the chain would be good starting values
- Idea: Run the MCMC for some time; then throw away these points and take current point as start
- Burn-In; e.g. use 10% of iterations to find starting value



Starting value +5

Difficult distributions

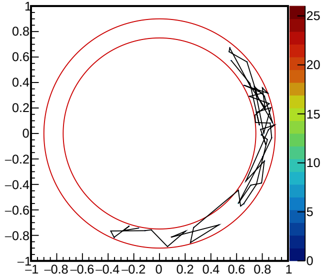


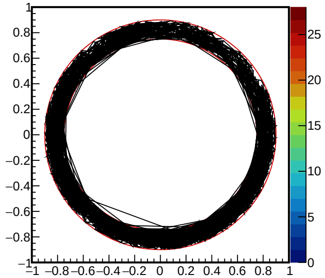


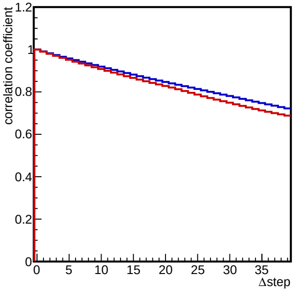
- Difficult-to-sample posterior distributions
- Reasons can be high dimensionality; but also distribution shape
- Particularly thin, bendy distributions difficult
- Simple Metropolis samplers an have difficulties traversing them
- Metropolis-Hastings algorithm actually very flexible
- Many techniques developed to deal with different problems
- Here, example of Gibbs-sampler
- Simple example: Annulus (Ring)

Metropolis sampler on the annulus

- If step size as big as circle, mostly rejected steps
- If step size as big as edge, takes a long time to traverse around
- Correlation time is very long



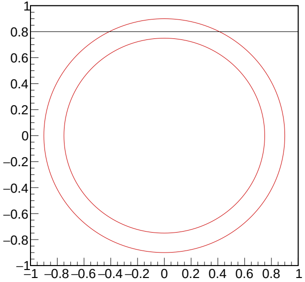




Conditional probability

- Want so sample $p(x_1, x_2, \ldots)$, (x_i the free parameters of the model)
- Consider leaving all coordinates except for one constant, e.g. $p(x_1|x_2, \ldots)$
- If it is possible to sample from this conditional probability efficiently, then the distribution is a good candidate for the Gibbs sampler

- For the annulus: For constant y, $R_2 > y > R_1$, one region of flat probability; sampling is easy
- For $y < R_1$ two regions of equal total probability. Flip a coin to select one, then sample flat probability
- Same for fixed x



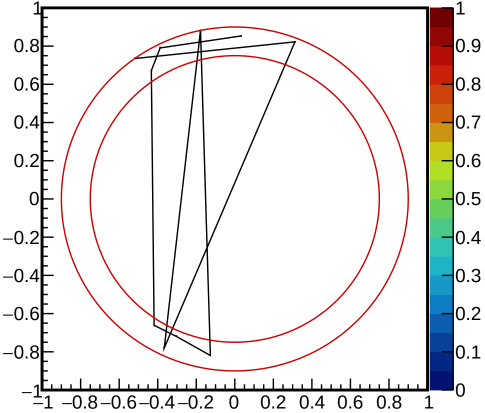
The Gibbs sampler algorithm

- Starting at point $\vec{x}^{(n)}$
- Keep all coordinates constant except for $x_1^{(n)}$; sample a new value for $x_1^{(n+1)}$ from this conditional probability $p(x_1|x_2^{(n)}, x_3^{(n)}, \ldots)$
- Update the value for x_1 and sample x_2 from the conditional distribution $p(x_2|x_1^{(n+1)}, x_3^{(n)}, x_4^{(n)}, \ldots)$
- Repeat this for all coordinates always using the updated values
- The new value is $\vec{x}^{(n+1)}$
- Steps are always accepted

- This is a special case of the Metropolis-Hastings algorithm

The Gibbs sampler in action

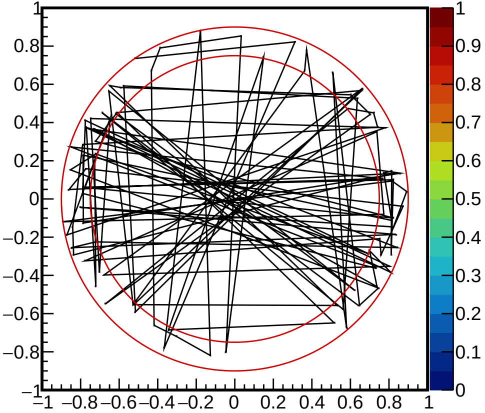
- Each step can now jump to other parts of the circle
- But now all states can be reached from all others
- Fills up distribution quickly



10 points

The Gibbs sampler in action

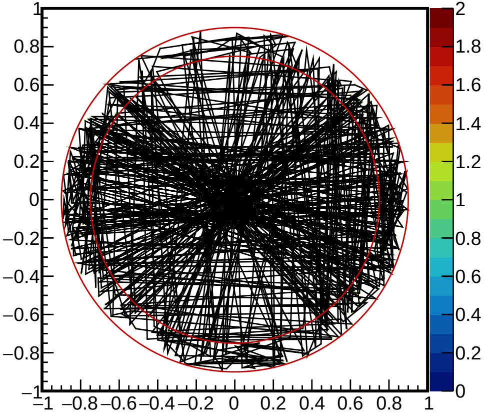
- Each step can now jump to other parts of the circle
- But now all states can be reached from all others
- Fills up distribution quickly



100 points

The Gibbs sampler in action

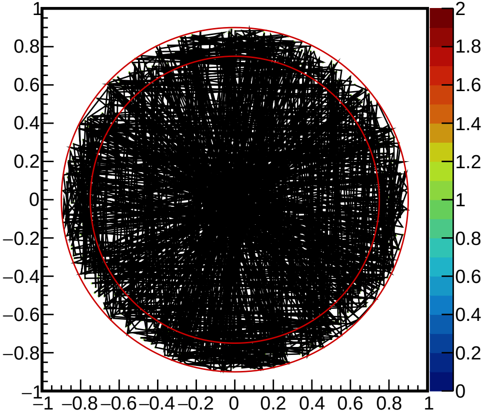
- Each step can now jump to other parts of the circle
- But now all states can be reached from all others
- Fills up distribution quickly



500 points

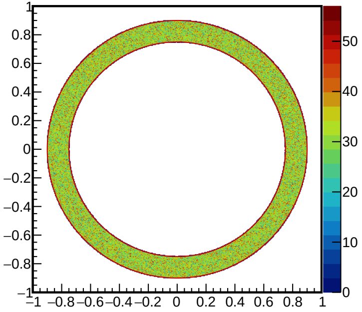
The Gibbs sampler in action

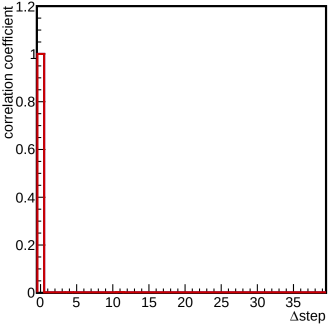
- Each step can now jump to other parts of the circle
- But now all states can be reached from all others
- Fills up distribution quickly



1k points

Correlation coefficient





- Sampler converges nicely
- Correlation immediately jumps to 0
- But not uncorrelated
- Symmetry of the jumps means that $\pm x$ and $\pm y$ are each equally likely
- $\rightarrow$ Pearson coefficient is not a good measure of the correlations

Reminder: Bayesian Hypothesis testing

- Bayes' theorem for a set of parameters $\vec{\lambda}$ and data $\vec{d}$:

$$p(\vec{\lambda}|\vec{d}) = \frac{p(\vec{d}|\vec{\lambda})\ p(\vec{\lambda})}{\int \mathrm{d}\vec{\lambda}\ p(\vec{d}|\vec{\lambda})\ p(\vec{\lambda})}$$

- Bayes' theorem for Hypotheses H_0 and H_1

$$p(H_1|\vec{d}) = \frac{p(\vec{d}|H_1)\ p(H_1)}{p(\vec{d}|H_1)\ p(H_1) + p(\vec{d}|H_0)\ p(H_0)}$$

- If we are interested in only some parameters, the others can be "integrated out" by marginalization:

$$p(\lambda_1|\vec{d}) = \int \mathrm{d}\lambda_2\ p(\lambda_1, \lambda_2|\vec{d})$$

- Using MCMC we can walk through the model/parameter space and find the marginals
- But what happens if the different hypotheses have different sets of parameters? (e.g. GW signal from BH merger vs. only background)
- Not easily possible with the type of MCMC discussed so far

Simple Hypothesis testing

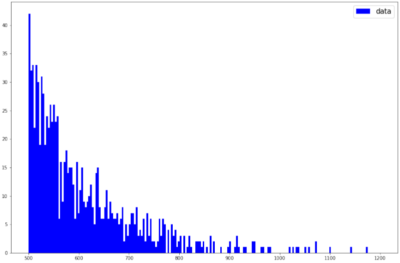
- Exercise 7.3 "Significance of a Peak"
- In exercise: Compare two simple hypotheses: Signal vs. no Signal using likelihood ratio; Test statistic was a log-likelihood ratio of 2.6
- Bayes:

$$p(H_1|\vec{d}) = \frac{p(\vec{d}|H_1)\ p(H_1)}{p(\vec{d}|H_1)\ p(H_1) + p(\vec{d}|H_0)\ p(H_0)}$$

- The normalization drops out in the ratio:

$$\frac{p(H_1|\vec{d})}{p(H_0|\vec{d})} = \frac{p(\vec{d}|H_1)\ p(H_1)}{p(\vec{d}|H_0)\ p(H_0)}$$

- The factor $p(\vec{d}|H_1)/p(\vec{d}|H_0)$ shows how the ratio of probabilities of the probabilities changes with the addition of the data; it is called the Bayes' factor
- The Bayes' factor is $\exp(2.6) \approx 13.5$; at this point the Bayesian analysis is done



Hypotheses with parameters

- Bayes' theorem for Hypotheses H_0 and H_1 with parameters $\vec{\theta_1}$, $\vec{\theta_0}$

$$p(H_1, \vec{\theta_1}|\vec{d}) = \frac{p(\vec{d}|H_1, \vec{\theta_1})\ p(H_1, \vec{\theta_1})}{\int \mathrm{d}\vec{\theta_1} p(\vec{d}|H_1, \vec{\theta_1})\ p(H_1, \vec{\theta_1}) + \int \mathrm{d}\vec{\theta_0} p(\vec{d}|H_0, \vec{\theta_0})\ p(H_0, \vec{\theta_0})}$$

- We can marginalize out the other parameters:

$$p(H_1|\vec{d}) = \int \mathrm{d}\vec{\theta_1}\ p(\vec{d}|H_1, \vec{\theta_1})\ p(H_1, \vec{\theta_1})$$

- If we take the ratio of the two posteriors again and make use of $p(H_1, \vec{\theta_1}) = p(\vec{\theta_1}|H_1)\ p(H_1)$, then we get:

$$\frac{p(H_1|\vec{d})}{p(H_0|\vec{d})} = \frac{\int \mathrm{d}\vec{\theta_1}\ p(\vec{d}|H_1, \vec{\theta_1})\ p(\vec{\theta_1}|H_1)}{\int \mathrm{d}\vec{\theta_0}\ p(\vec{d}|H_0, \vec{\theta_0})\ p(\vec{\theta_0}|H_0)}\ \frac{p(H_1)}{p(H_0)}$$

- Bayes factor contains normalization constants for the posterior for the single models!
- This integral very much depends on the absolute normalization, thus MCMC cannot be used!
- The factor $\int \mathrm{d}\vec{\theta_1}\ p(\vec{d}|H_1, \vec{\theta_1})\ p(\vec{\theta_1}|H_1)$ only depends on one model, it is sometimes called the evidence

Hypothesis with parameters

- Make signal hypothesis slightly more complex: Variable signal fraction f_s
- Prior on signal fraction; flat interval for simplicity
- For H_1, calculate

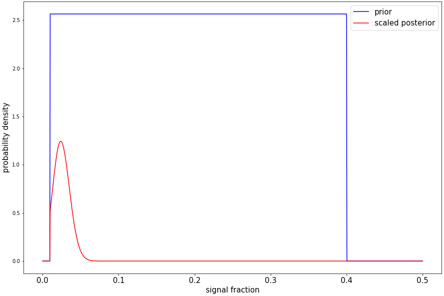
$$\int \mathrm{d}f_s \; p(\vec{d}|f_s, H_1) \; p(f_s|H_1)$$

- For H_0, we just need the likelihood: $p(\vec{d}|H_0)$, which was done in the exercise
- Result:

$$\frac{p(H_1|\vec{d})}{p(H_0|\vec{d})} \approx 1.02$$

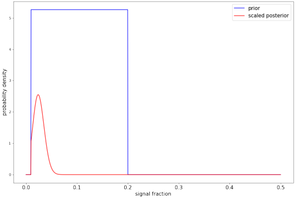
- For a fixed signal this was 13.5!

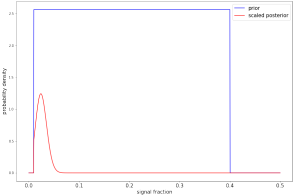
```
def model(m, fsig):
    """normalized pdf for the mass distribution, fsig = signal fraction (0 <= fs
    return (1-fsig)*bkg.pdf(m) + fsig*sig.pdf(m)
plt.plot(m, model(m, 0.1), label='$H_1$')

lowedge = 0.01
upedge = 0.4
def prior(mgg):
    return np.greater(mgg, lowedge)*np.greater(upedge,mgg)/(upedge-lowedge)

def LLikelihood(evts, fsig):
    return np.sum(np.log(model(evts, fsig)))
```

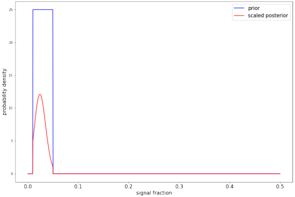


Comparing different priors

- Why is the probability so much lower if the parameter is not fixed?
- Large region of the prior in H_1 is actually excluded by the data
- Remaining prior mass is very low
- Compare Bayes factor B for different priors (from 0.01 to a)
  - $a = 0.20$, $B = 2.1$
  - $a = 0.4$, $B = 1.02$
  - $a = 0.05$, $B = 9.7$
- The clearer the model is, the more evidence it gives for the hypothesis
- With more free parameters this effect is even stronger; the peak region will be a small part of the parameter space
- This encodes something like Occam's razor: More complex models are disfavoured; more parameters and more available parameter space decreases the Bayes factor

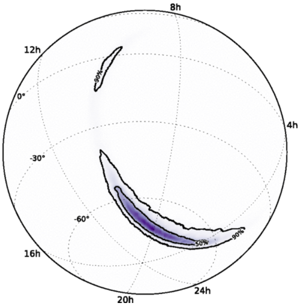






Nested Sampling

- Likelihood-distribution has lines of equal value
- Most of the distribution is in a fairly small region of the available space
- Nested Sampling calculates evidence and also gives posterior distribution of parameters
- Set of points which is constrained within regions of larger and larger likelihood



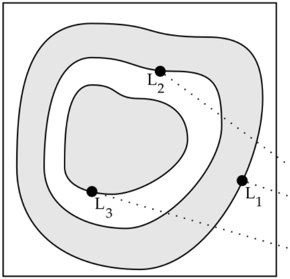
Algorithm:

1. Start with a number of N points sampled from the prior, $i = 0$
2. Find the point with the lowest likelihood, and set L_i to this value
3. Remove this point and replace it by another point sampled from the prior, but only allowing points with a likelihood larger than L_i
4. Repeat steps 2+3 a number of times filling some variables for each step

- Finding the new point with the constraints is not trivial, but can be done for example with MCMC

Nested Sampling

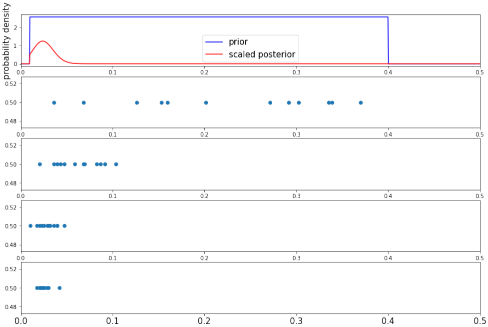
- For each step i and likelihood L_i (with N points):
  - $X_i = \exp(-i/N)$
  - $w_i = X_{i-1} - X_i$
  - $Z \to Z + L_i \cdot w_i$
- Then Z converges towards the evidence



Parameter space

Nested Sampling - Example

- For 1d finding the region is simple
- 12 points shown after 0, 20, 40, and 60 steps



Black hole merger

- For high dimensions, not so easy to sample from confined prior
- Also not so easy to see when algorithm converged
- Need to take into account possibility of "islands"
- MCMC methods work for this
- Bayes factor of e^{289} for signal vs. background-only hypothesis

	EOBNR	IMRPhenom	Overall
Detector-frame total mass $M/M_\odot$	$70.3^{+5.3}_{-4.8}$	$70.9^{+4.0}_{-3.9}$	$70.6^{+4.6\pm0.5}_{-4.5\pm1.3}$
Detector-frame chirp mass $\mathcal{M}/M_\odot$	$30.2^{+2.5}_{-1.9}$	$30.6^{+1.8}_{-1.8}$	$30.4^{+2.1\pm0.2}_{-1.9\pm0.5}$
Detector-frame primary mass $m_1/M_\odot$	$39.4^{+5.5}_{-4.9}$	$38.5^{+5.6}_{-3.6}$	$38.9^{+5.6\pm0.6}_{-4.3\pm0.4}$
Detector-frame secondary mass $m_2/M_\odot$	$30.9^{+4.8}_{-4.4}$	$32.2^{+3.6}_{-4.8}$	$31.6^{+4.2\pm0.1}_{-4.7\pm0.9}$
Detector-frame final mass $M_f/M_\odot$	$67.1^{+4.6}_{-4.4}$	$67.6^{+3.6}_{-3.5}$	$67.4^{+4.1\pm0.4}_{-4.0\pm1.2}$
Source-frame total mass $M^{\text{source}}/M_\odot$	$65.0^{+5.0}_{-4.4}$	$65.0^{+4.0}_{-3.6}$	$65.0^{+4.5\pm0.8}_{-4.0\pm0.7}$
Source-frame chirp mass $\mathcal{M}^{\text{source}}/M_\odot$	$27.9^{+2.3}_{-1.8}$	$28.1^{+1.7}_{-1.6}$	$28.0^{+2.0\pm0.3}_{-1.7\pm0.3}$
Source-frame primary mass $m_1^{\text{source}}/M_\odot$	$36.3^{+5.3}_{-4.5}$	$35.3^{+5.2}_{-3.4}$	$35.8^{+5.3\pm0.9}_{-3.9\pm0.1}$
Source-frame secondary mass $m_2^{\text{source}}/M_\odot$	$28.6^{+4.4}_{-4.2}$	$29.6^{+3.3}_{-4.3}$	$29.1^{+3.8\pm0.1}_{-4.3\pm0.7}$
Source-frame final mass $M_f^{\text{source}}/M_\odot$	$62.0^{+4.4}_{-4.0}$	$62.0^{+3.7}_{-3.3}$	$62.0^{+4.1\pm0.7}_{-3.7\pm0.6}$
Mass ratio q	$0.79^{+0.18}_{-0.19}$	$0.84^{+0.14}_{-0.20}$	$0.82^{+0.17\pm0.01}_{-0.20\pm0.03}$
Effective inspiral spin parameter χ_{eff}	$-0.09^{+0.19}_{-0.17}$	$-0.05^{+0.13}_{-0.15}$	$-0.07^{+0.16\pm0.01}_{-0.17\pm0.05}$
Dimensionless primary spin magnitude a_1	$0.32^{+0.45}_{-0.28}$	$0.32^{+0.53}_{-0.29}$	$0.32^{+0.49\pm0.06}_{-0.29\pm0.01}$
Dimensionless secondary spin magnitude a_2	$0.57^{+0.40}_{-0.51}$	$0.34^{+0.54}_{-0.31}$	$0.44^{+0.50\pm0.08}_{-0.40\pm0.02}$
Final spin a_f	$0.67^{+0.06}_{-0.08}$	$0.66^{+0.04}_{-0.06}$	$0.67^{+0.05\pm0.01}_{-0.07\pm0.02}$
Luminosity distance D_L/Mpc	390^{+170}_{-180}	440^{+150}_{-180}	$410^{+160\pm20}_{-180\pm40}$
Source redshift z	$0.083^{+0.033}_{-0.036}$	$0.093^{+0.029}_{-0.036}$	$0.088^{+0.032\pm0.005}_{-0.037\pm0.008}$
Upper bound on primary spin magnitude a_1	0.65	0.74	0.69 ± 0.08
Upper bound on secondary spin magnitude a_2	0.93	0.78	0.89 ± 0.13
Lower bound on mass ratio q	0.64	0.68	0.66 ± 0.03
Log Bayes factor $\ln \mathcal{B}_{s/n}$	288.7 ± 0.2	290.3 ± 0.1	...