Lecture 13: Nested Sampling for Bayesian Inference

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2025

Photo by Howard Jackman

Take-home Exam

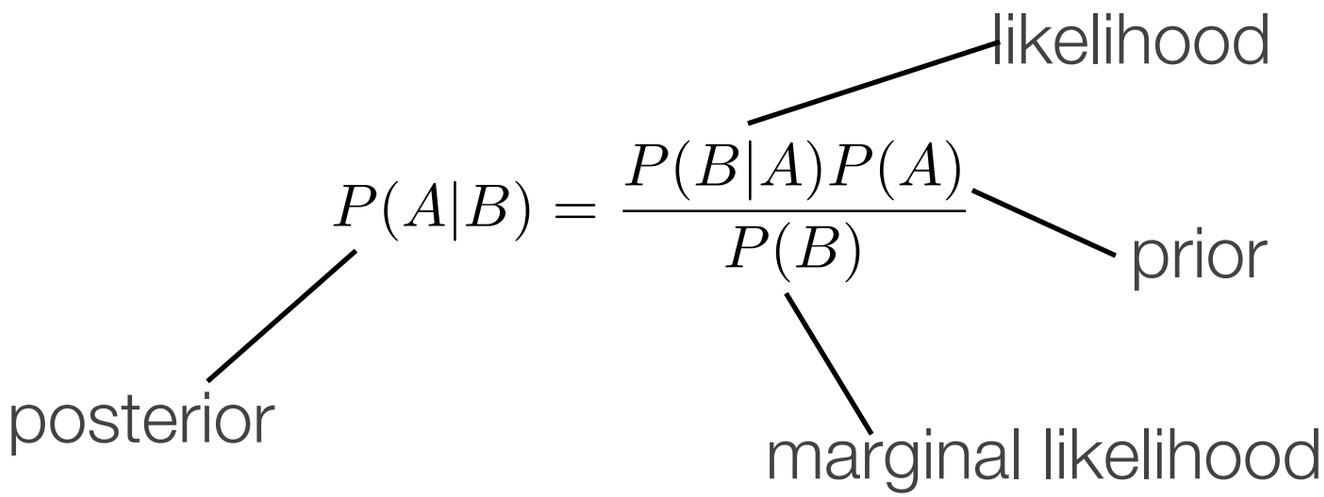
- 40% of the final course grade

Comments

- For the following nested sampling lecture, I have included more references at the end of the slides as well as on the course webpage
- As far as packages to use for nested sampling, I am fond of Nestle (see online links) and see that UltraNest might also be a nice option (along with Dynesty)

Bayes' Theorem

- One can solve the respective conditional probability equations for P(A and B) and P(B and A), setting them equal to give Bayes' theorem:



$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- The theorem applies to both frequentist and Bayesian methods. Differences stem from how the theorem is applied and, in particular, whether one extends probability to include some degree of belief.

Slight Notation Shift

- Previously, we have focused on the posterior distribution P(ΘlD,H) which is critical for parameter estimation, and then Markov Chain Monte Carlo can then calculate the marginal likelihood P(DlH)
- For model selection — versus parameter estimation — the marginal likelihood is important in its own right. The problem is that many MCMC methods are slow (simulated annealing).

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

D are data
Θ are parameters
H is hypothesis or model

New Task

- If model selection is important then comparing models can be done via the respective posterior distributions

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)} = \frac{Z_1 P(H_1)}{Z_0 P(H_0)}$$

- The “marginal likelihood” is now rebranded as the “Bayesian evidence” and noted as Z
- Reversing the traditional MCMC approach, the ‘evidence’ is now the primary target, and the posterior is a by-product
- Note: we won’t be doing model selection explicitly in this lecture, but it is the motivation for much of the following material

Nested Sampling

- In 2004, John Skilling came up with a new Monte Carlo sampling technique, known as nested sampling, to more efficiently evaluate the bayesian evidence (Z)

$$Z = \int \mathcal{L}(\Theta)\pi(\Theta)d\Theta$$

$$\mathcal{L} \text{ is the likelihood}$$
$$\pi \text{ is the prior}$$

- For higher dimensions of Θ the integral for the bayesian evidence becomes challenging

Nested Sampling

- If numerical integration in higher dimensions is troublesome, then we can transform the multi-dimensional integral to a one-dimensional integral, via

$$dX = \pi(\Theta)d\Theta$$

$$X(\lambda) = \int_{\mathcal{L}(\Theta)>\lambda} \pi(\Theta)d\Theta$$

- The new prior X is defined such that

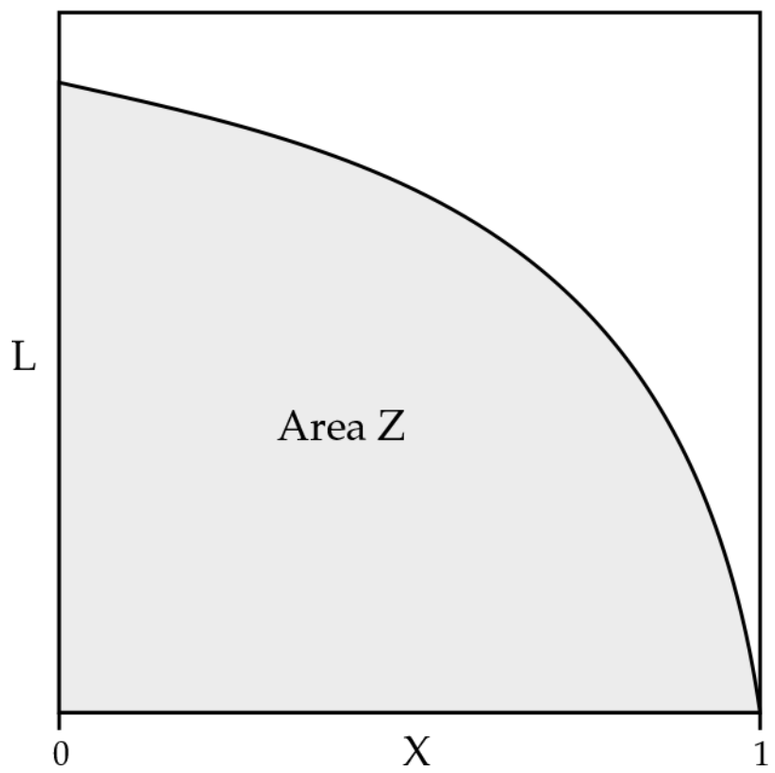
$$Z = \int_0^1 \mathcal{L}(X)dX$$

*For more justification, see paper by J .Skilling (DOI: 10.1214/06-BA127)

- Note that X is a probability function and can only be in the range from 0 to 1
- $\mathscr{L}(X)$ is also now a monotonically decreasing function
- A clever approx. to get X will be covered in later slides

New Likelihood in 1-D

- The bayesian evidence (Z) is now the 1-D integral of the re-parameterized likelihood ($\mathscr{L}(X)$) integrated over the re-parameterized prior (X)
  - The shape of $\mathscr{L}(X)$ could be any shape, but it **is** monotonically decreasing from 1→0, and by construction is bounded at 0 and 1.



$$dX = \pi(\Theta)d\Theta$$

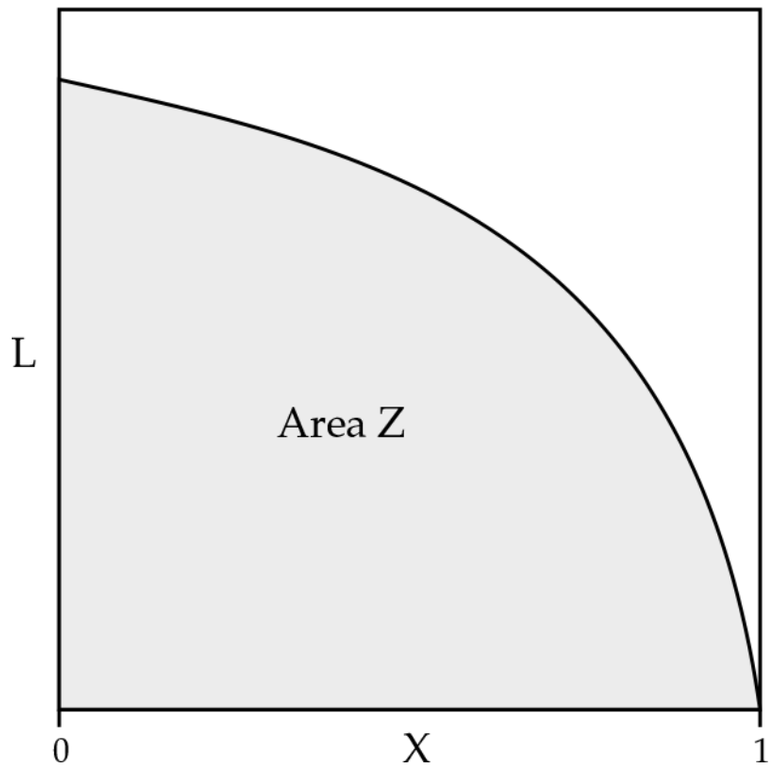
$$X(\lambda) = \int_{\mathcal{L}(\Theta)>\lambda} \pi(\Theta)d\Theta$$

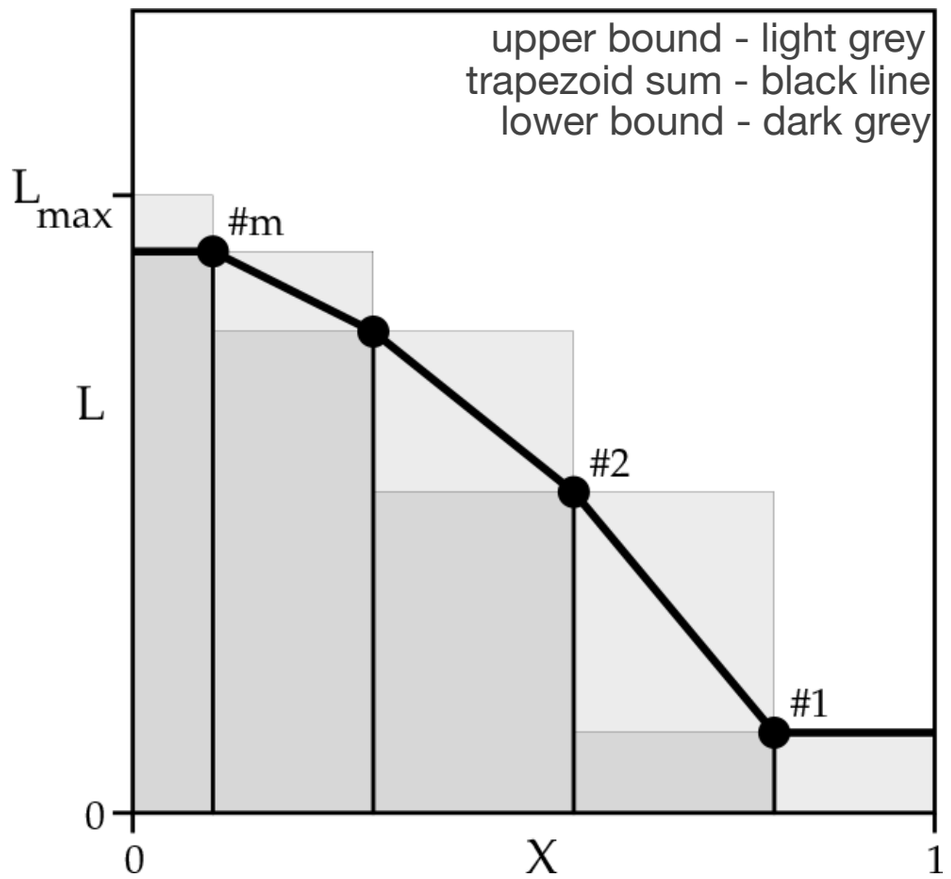
$$Z = \int_0^1 \mathcal{L}(X)dX$$

*J. Skilling

New Likelihood in 1-D

- The bayesian evidence is now the 1-D integral of the re-parameterized likelihood integrated over the re-parameterized prior
  - An analytic determination of the integral is not an option. If we could do it analytically, we wouldn't be using numerical integration.
  - Use points sampled in X to calculate the trapezoid sum
  - Diagram below (right) shows X and L for 4 sampled points

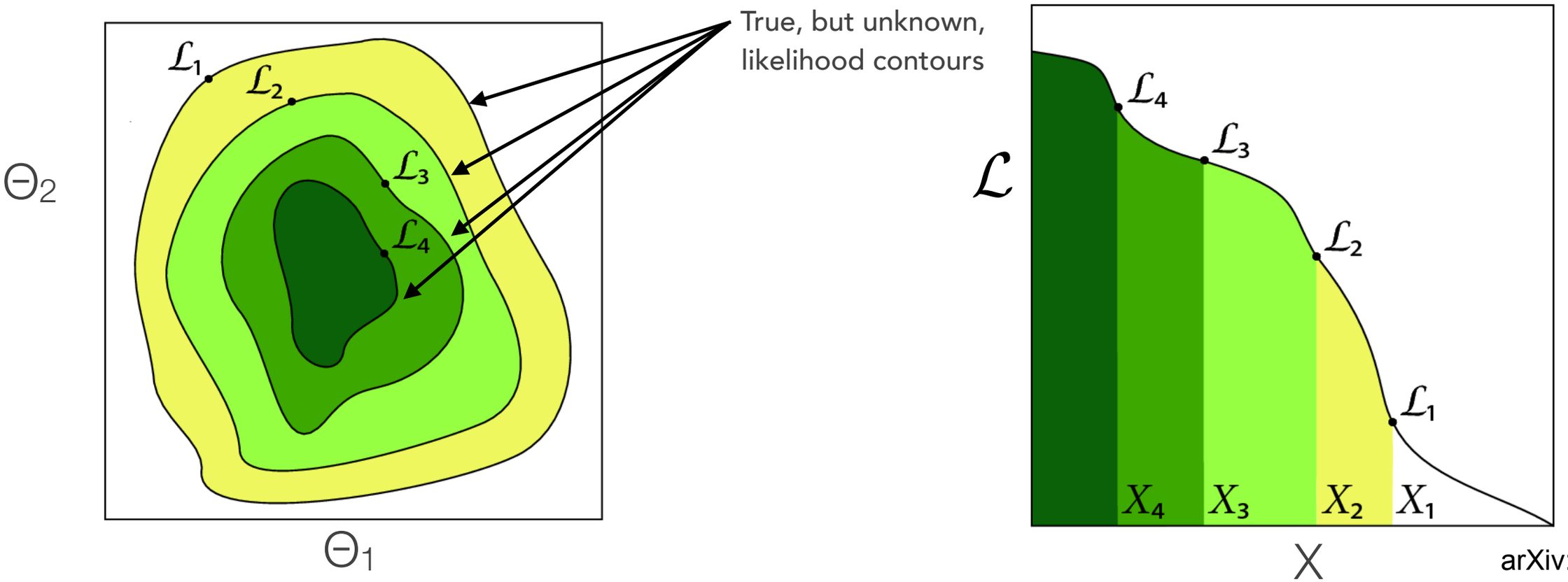




*J. Skilling

Simple Cartoon

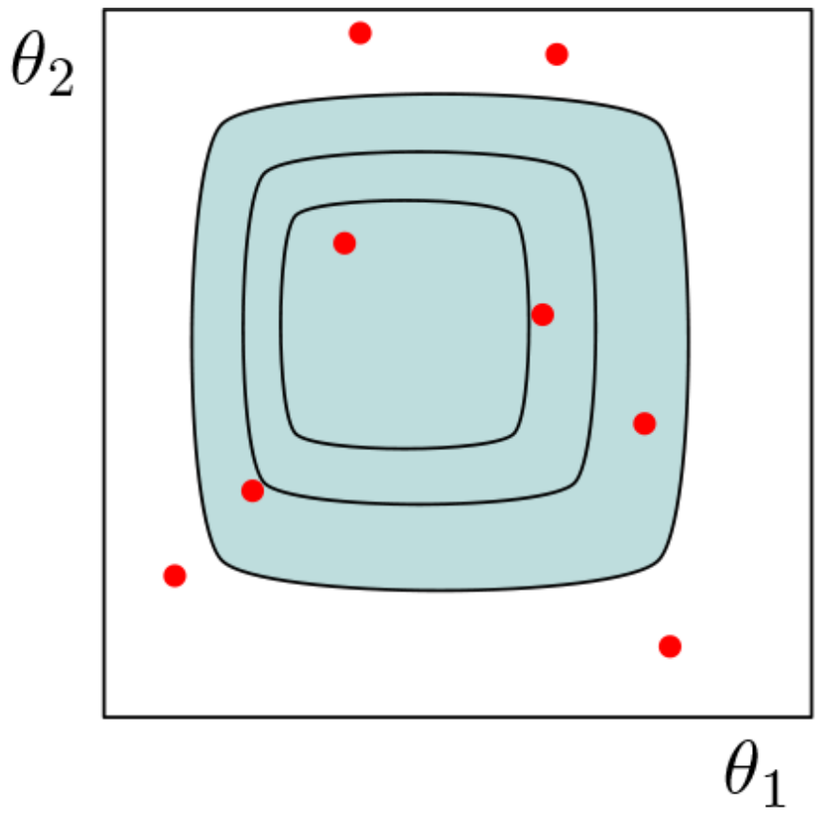
- For a simple 2-dimensional case, 4 ‘live’ points are sampled. The likelihood for each point ($\mathscr{L}_i$) is calculated.
  - Note that multiple points of (Θ_1, Θ_2) can have the same value of X. If $\mathscr{L}(\theta_1, \theta_2) = \mathscr{L}(\theta_1', \theta_2')$ for $(\theta_1, \theta_2) \neq (\theta_1', \theta_2')$, then both points will still have the same value of X.
  - This illustration nicely samples the space with only 4 points, which is uncommon and unrealistic



arXiv:1306.2144

Sampling

- Instead of relying on luck, it is better to sample the space *sparsely* where the new likelihood is *worse*, and sample **frequently** in the space where the likelihood is **better**



Shaded areas are the true underlying contours

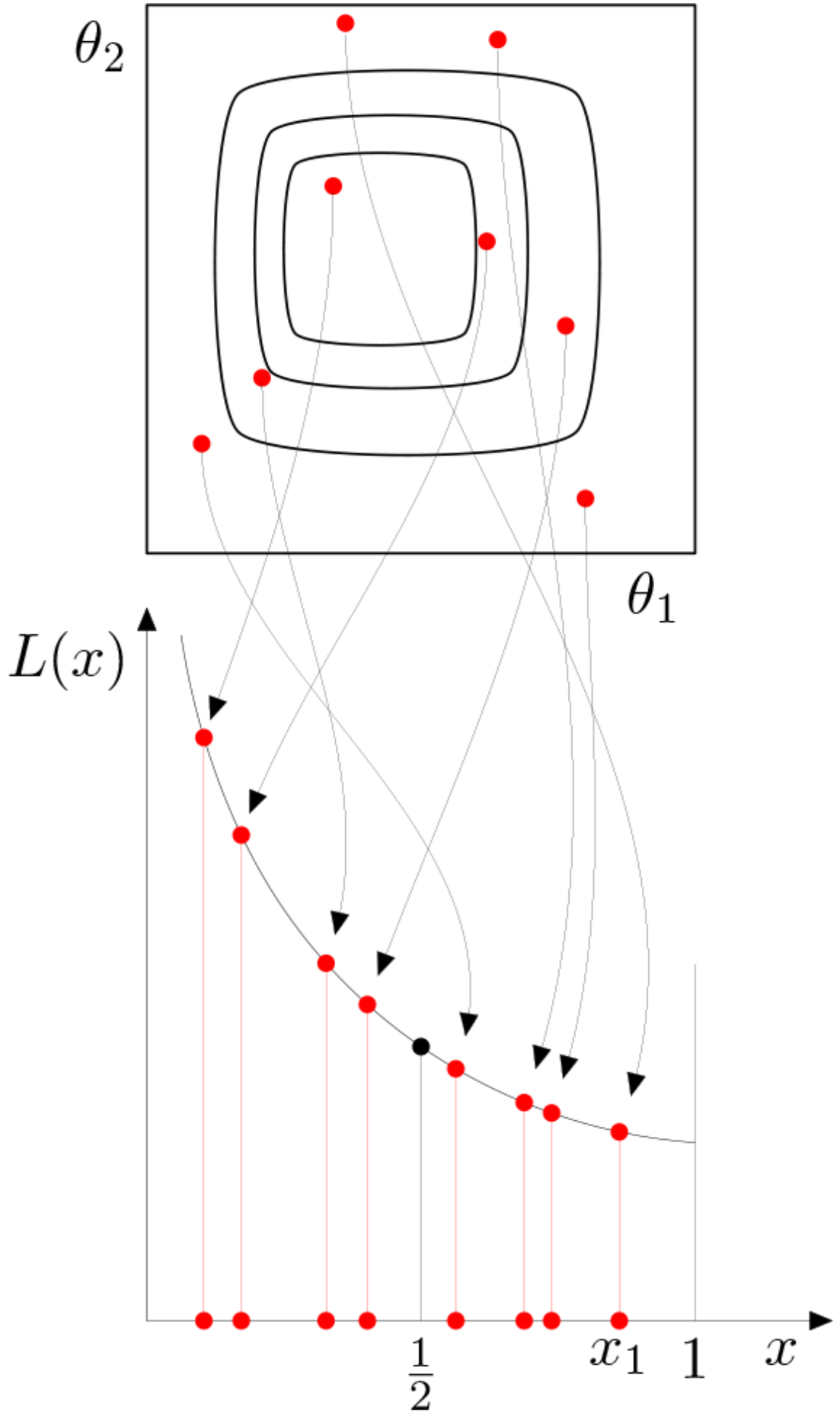
The sample points come from a flat prior in 2-D

Figure 51.3. $N = 8$ points drawn uniformly from the prior.

http://www.inference.phy.cam.ac.uk/bayesys/box/nested.pdf

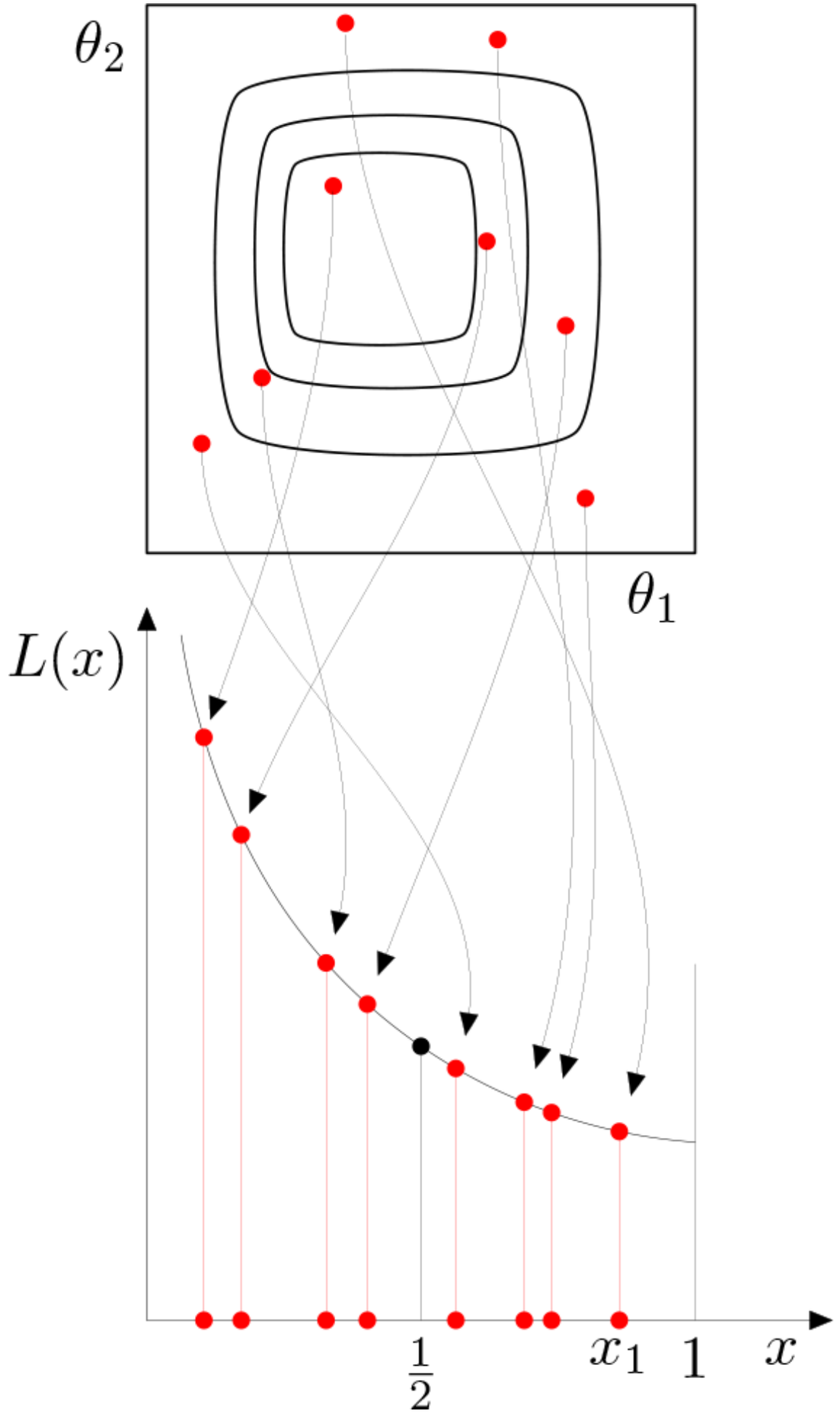
Sampling Start

- Each of the 8 sampled points has a likelihood value $\mathscr{L}(x_i)$ that can be ordered:
  $\mathscr{L}(x_1) < \mathscr{L}(x_2) < \mathscr{L}(x_3) < \ldots < \mathscr{L}(x_8)$
- To get the x-values, 8 values can be sampled from a uniform distribution in the range 0-1, and the largest x-value is defined as x_1
  - Second largest x-value is x_2, third largest is x_3, etc.
  - Sampling the x-values from a uniform distribution is not great, but is illustrative to convey that the x-values are not initially known.
- Can we use more than just the initial 8 points in some smart way?
  - **Absolutely!!**



Sampling Start cont.

- In order to better sample where the likelihood is high, the point with the lowest $\mathscr{L}_{lowest}$—i.e. x_1 in the diagram—is replaced by a new point.
  - A new point (θ_1, θ_2) is drawn from the sampling prior which produced the initial points. The x-value for the new (θ_1, θ_2) point is not yet known, but will be assigned when the new live point becomes the $\mathscr{L}_{lowest}$ at some iteration in the future.
  - Remove the point $\mathscr{L}_{lowest}$ (in the N-dimensional space (θ_1, θ_2) and 1D space as x_{lowest}), but store it's values to calculate the likelihood integral, e.g. bayesian evidence
- Next slide covers other approx. for values of x



Pseudo-Code

Generate N livepoints from the sampling prior
Loop where i increments as i=1,2,3,...
{

* Find the point $\mathscr{L}(\theta_1, \theta_2)$ with the lowest likelihood of the current live points, i.e. $\mathscr{L}_{lowest}$.

- Remove it from the population of livepoints, but store it for results. Estimate the value of

$$x_i = x_{lowest} = \left(\frac{N-1}{N}\right)^i.$$

* Add a new livepoint generated from sampling, where the new live point must satisfy that $\mathscr{L}_{new} > \mathscr{L}_{lowest}$.

}

Other estimates of X_{lowest} can be
* $(N/(N+1))^i$
* exp(-i/N)

*G. F. Lewis

Sampling more

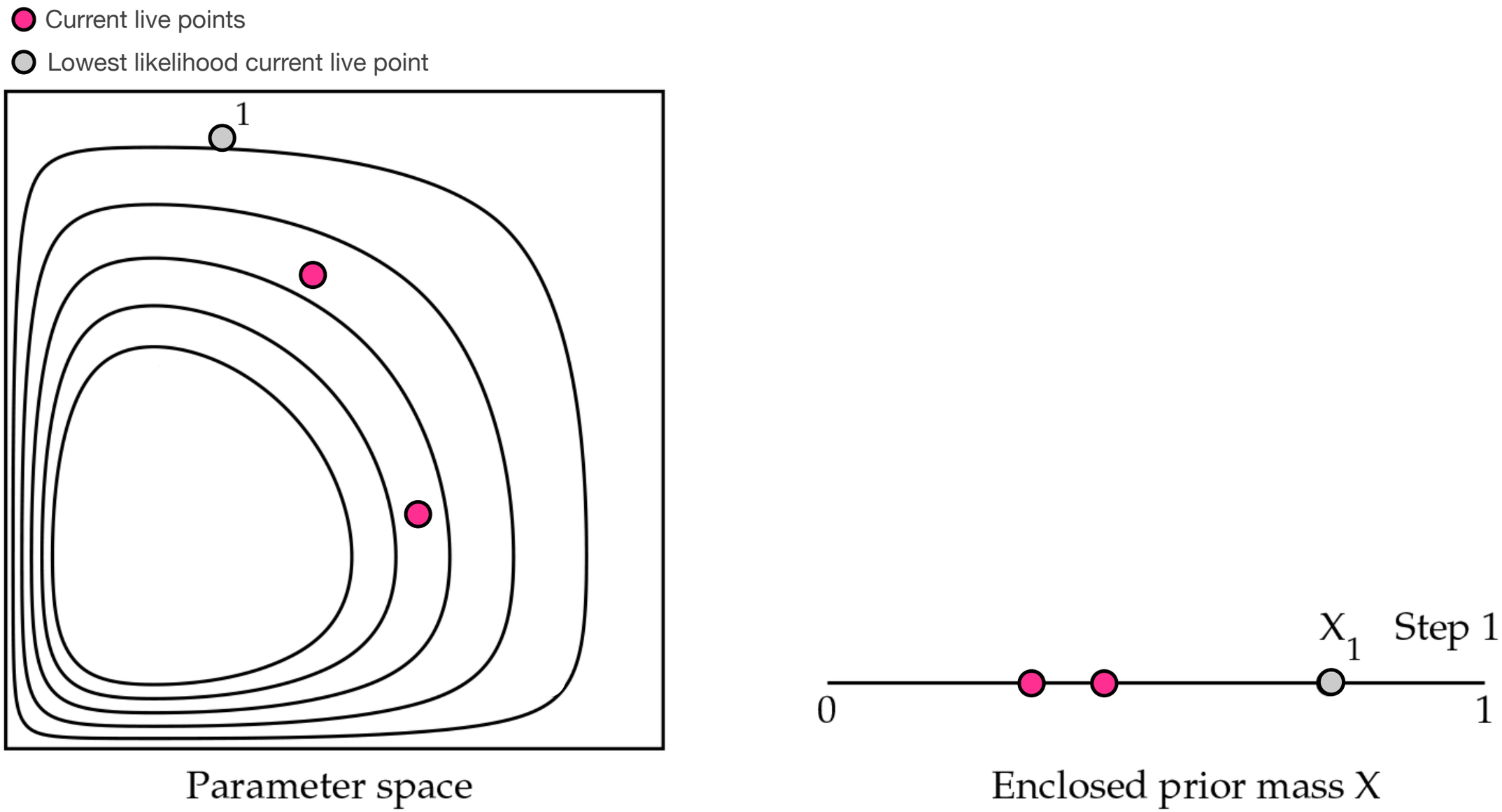


Figure 4: Nested sampling for five steps with a collection of three points. Likelihood contours shrink by factors $\exp(-1/3)$ in area and are roughly followed by successive sample points.

Modified from *J. Skilling 2006

Sampling more

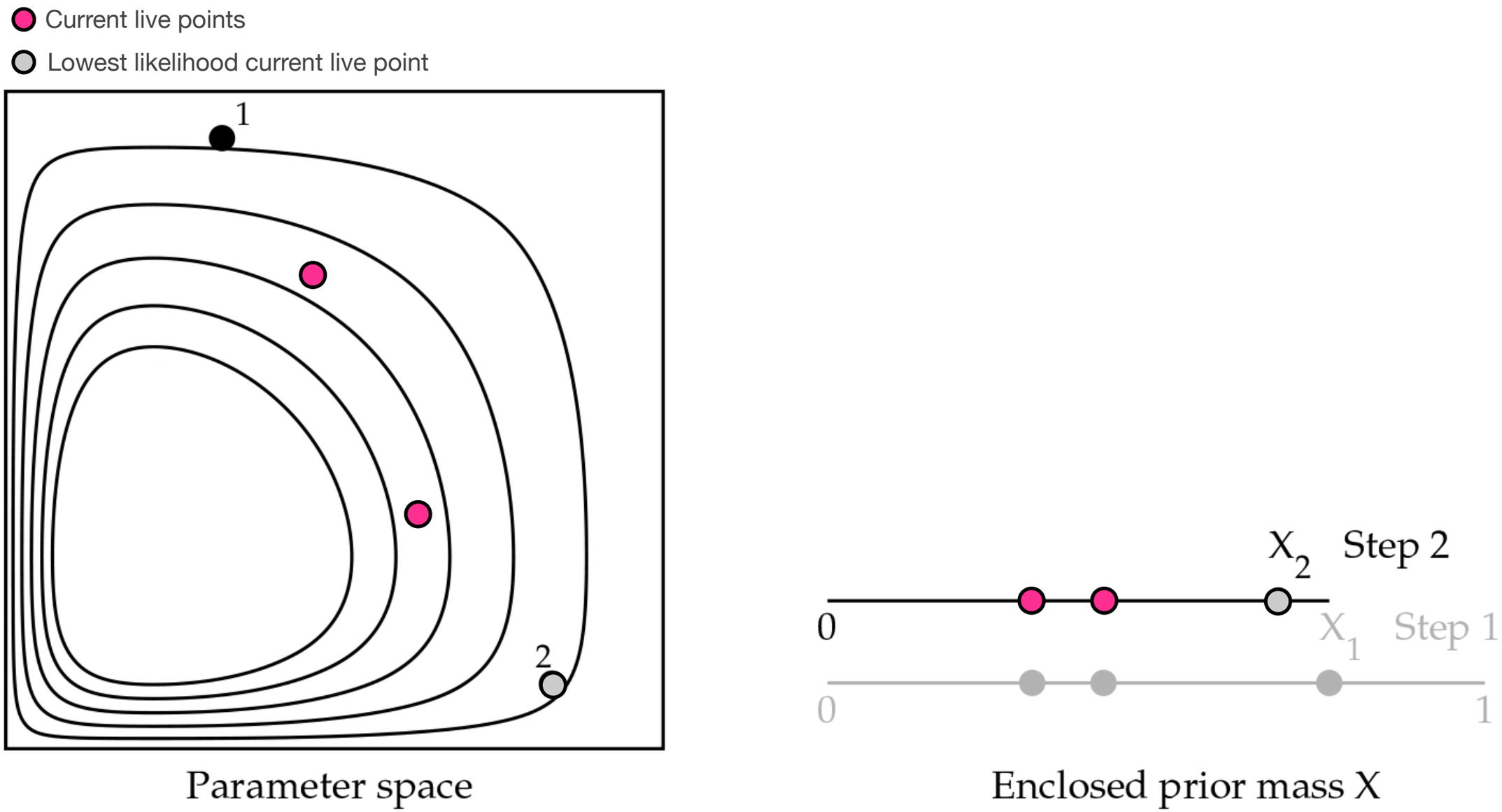


Figure 4: Nested sampling for five steps with a collection of three points. Likelihood contours shrink by factors $\exp(-1/3)$ in area and are roughly followed by successive sample points.

Modified from *J. Skilling 2006

Sampling more

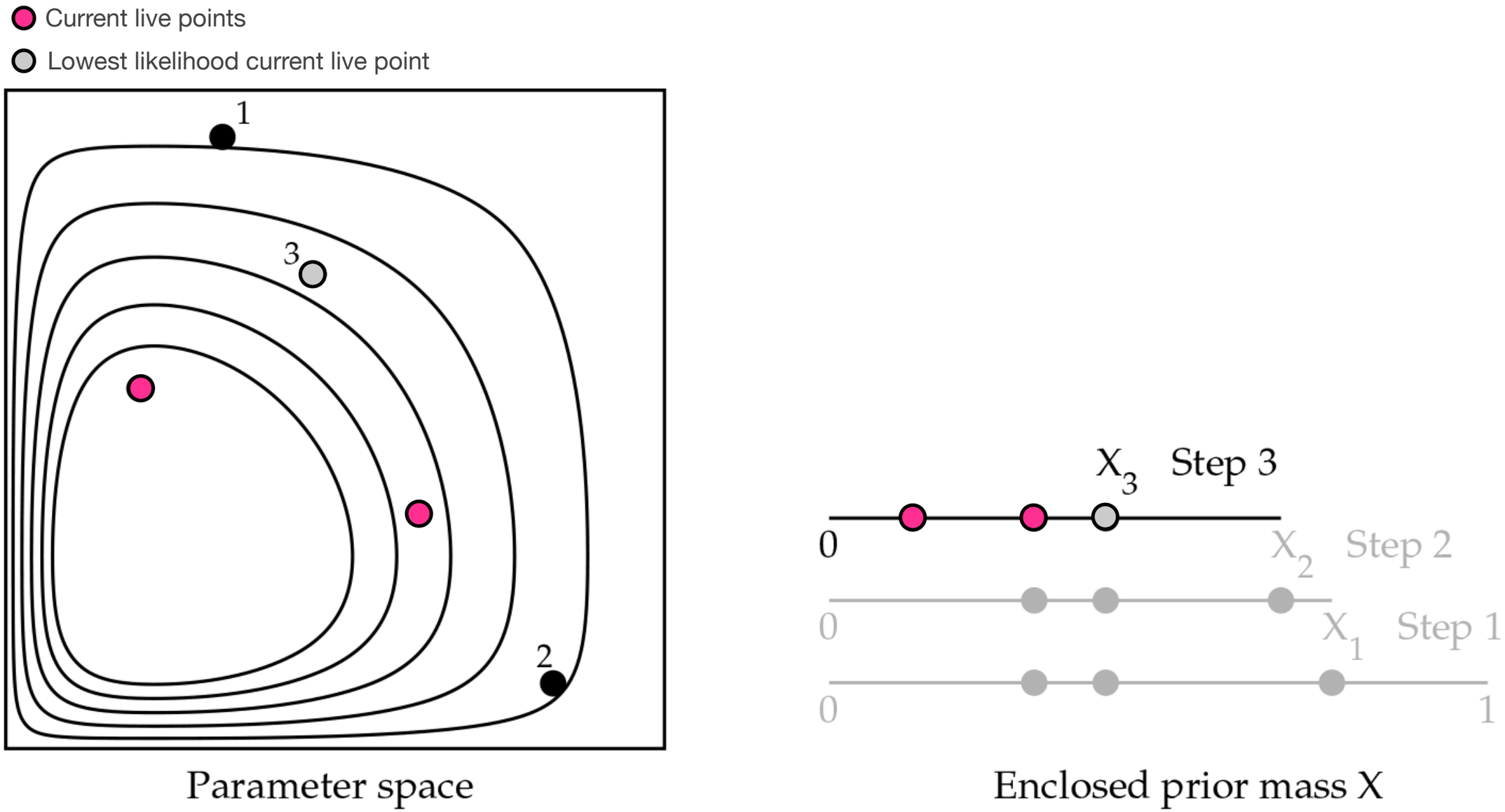


Figure 4: Nested sampling for five steps with a collection of three points. Likelihood contours shrink by factors $\exp(-1/3)$ in area and are roughly followed by successive sample points.

Modified from *J. Skilling 2006

Sampling more

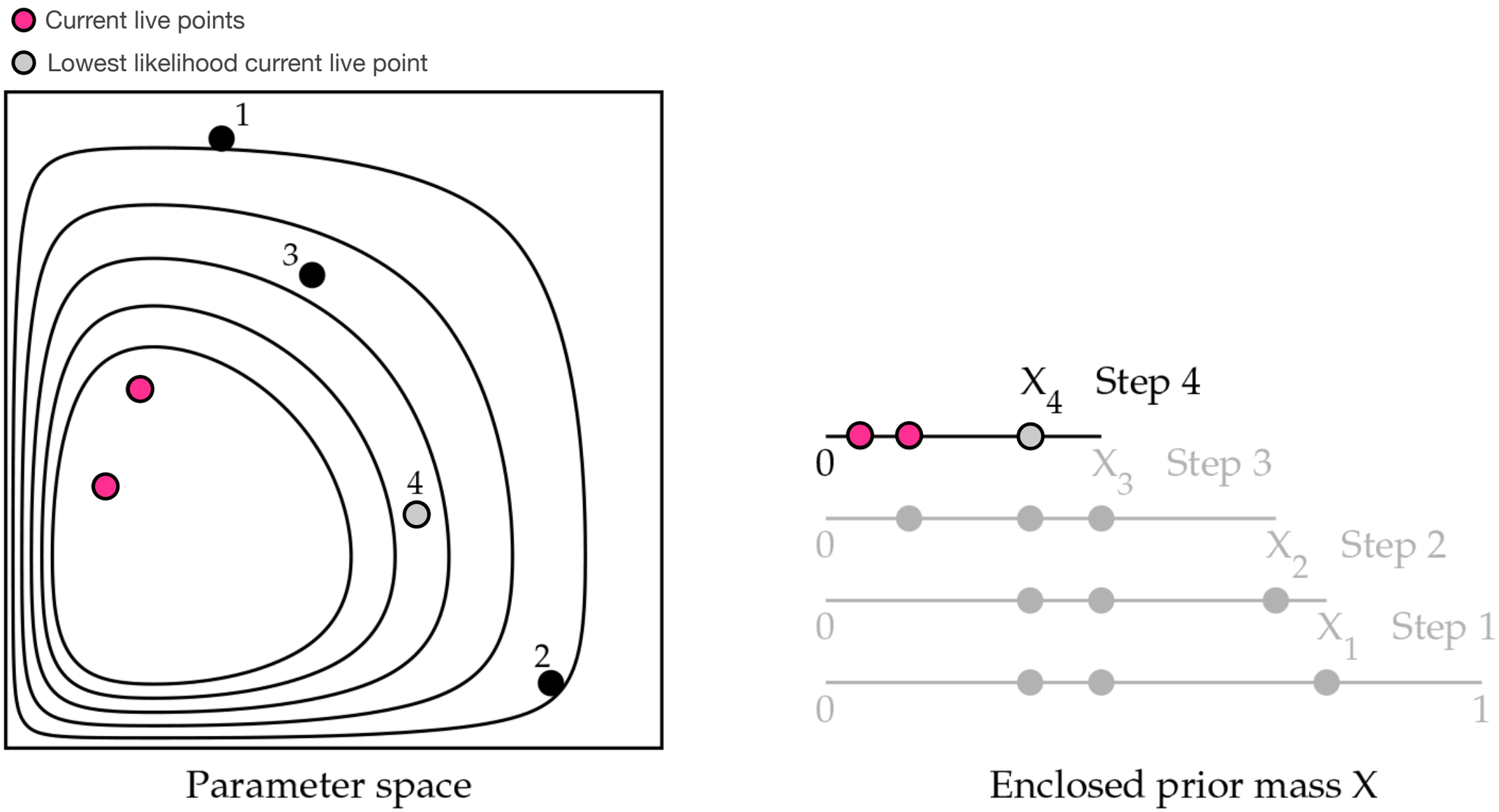


Figure 4: Nested sampling for five steps with a collection of three points. Likelihood contours shrink by factors $\exp(-1/3)$ in area and are roughly followed by successive sample points.

Modified from *J. Skilling 2006

Sampling more

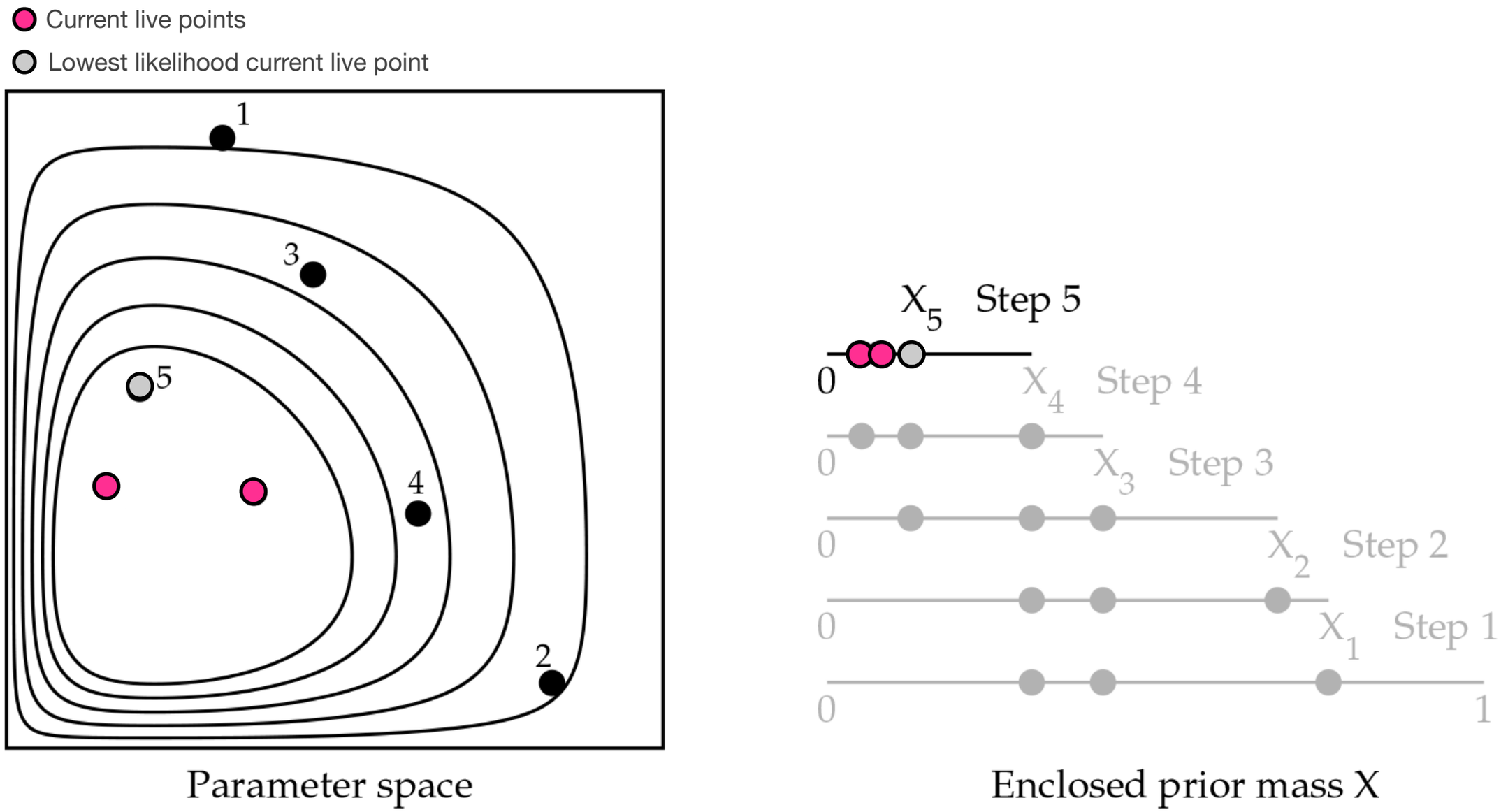


Figure 4: Nested sampling for five steps with a collection of three points. Likelihood contours shrink by factors $\exp(-1/3)$ in area and are roughly followed by successive sample points.

Modified from *J. Skilling 2006

Sampling more

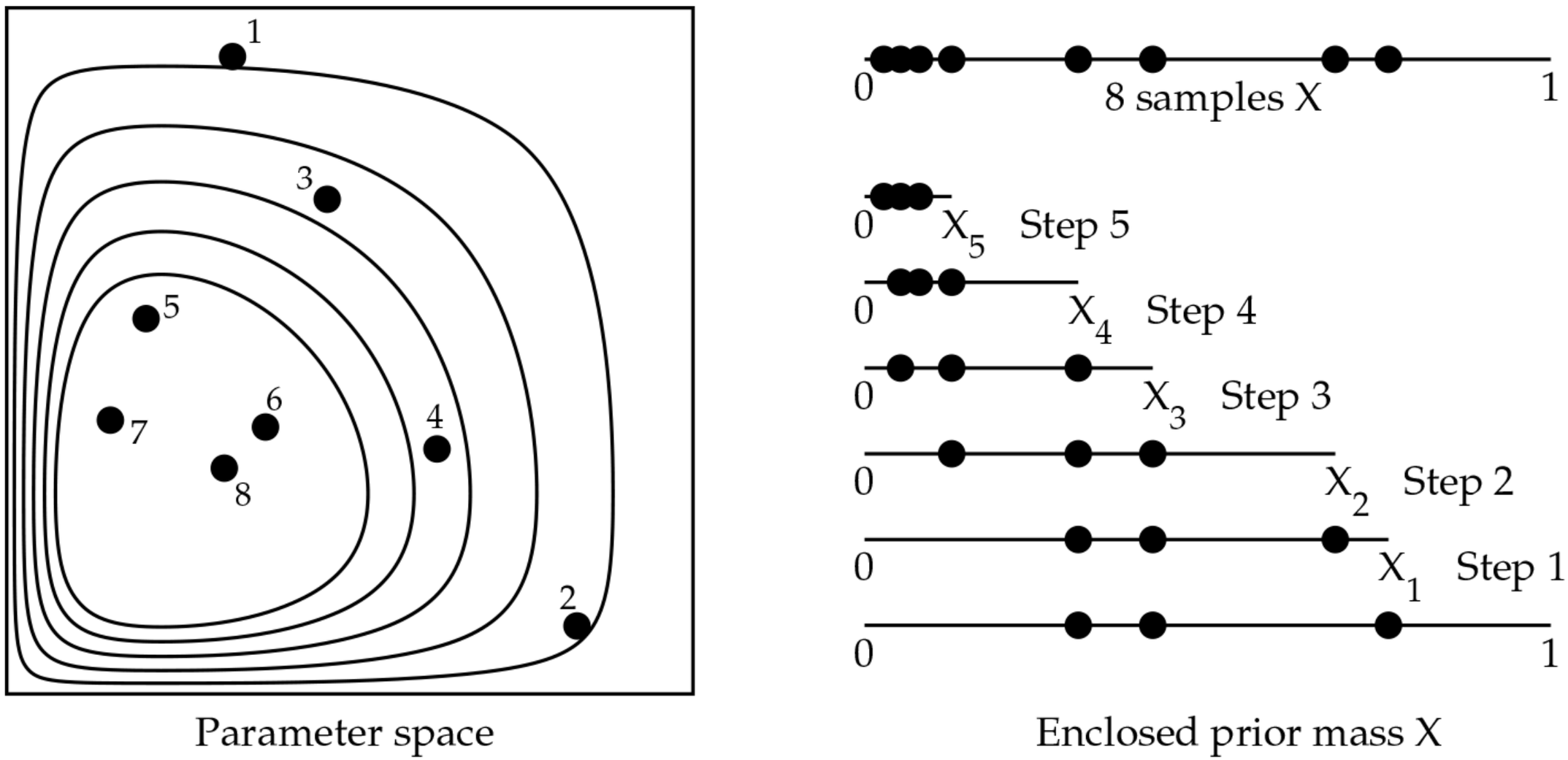
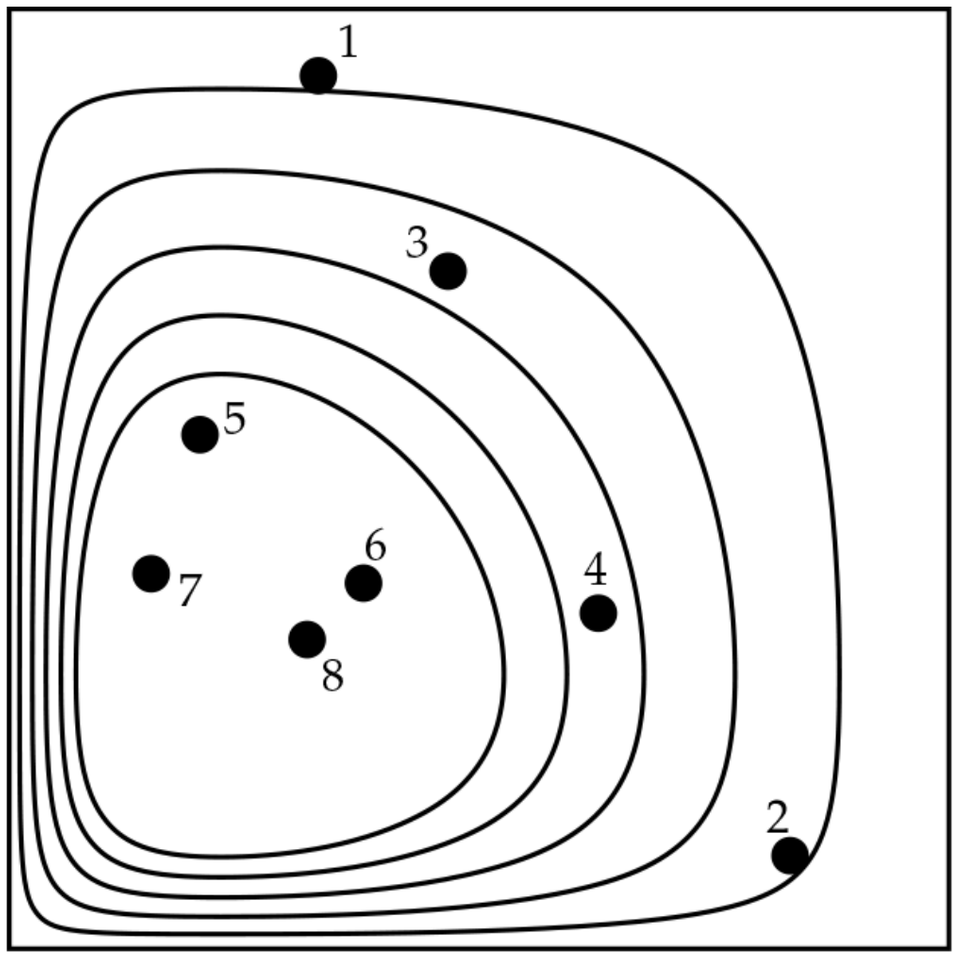


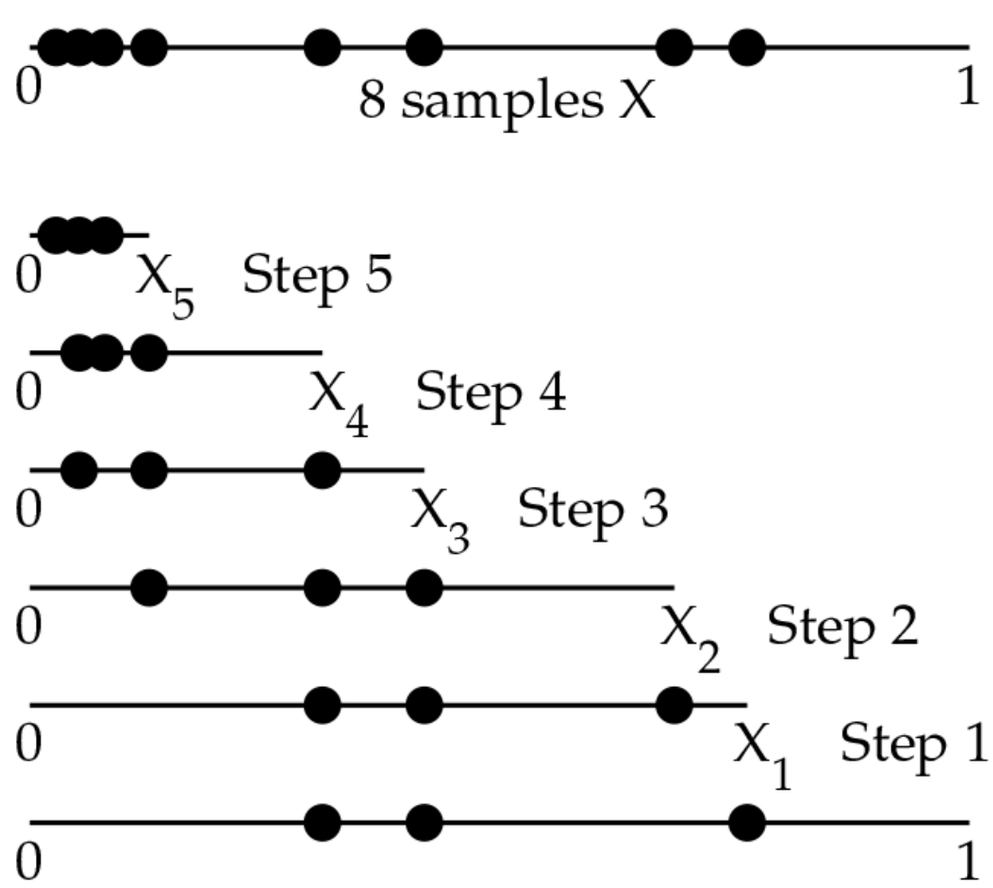
Figure 4: Nested sampling for five steps with a collection of three points. Likelihood contours shrink by factors $\exp(-1/3)$ in area and are roughly followed by successive sample points.

*J. Skilling 2006

The Integration



Parameter space



Enclosed prior mass X

- After 5 steps, we reach some stopping criteria, and have 8 total sample points to estimate the N-dimensional Bayesian evidence (Z) via the 1D integral of $\mathscr{L}(X)$, i.e. $Z = \int_0^1 \mathscr{L}(X)dX.$

*J. Skilling 2006

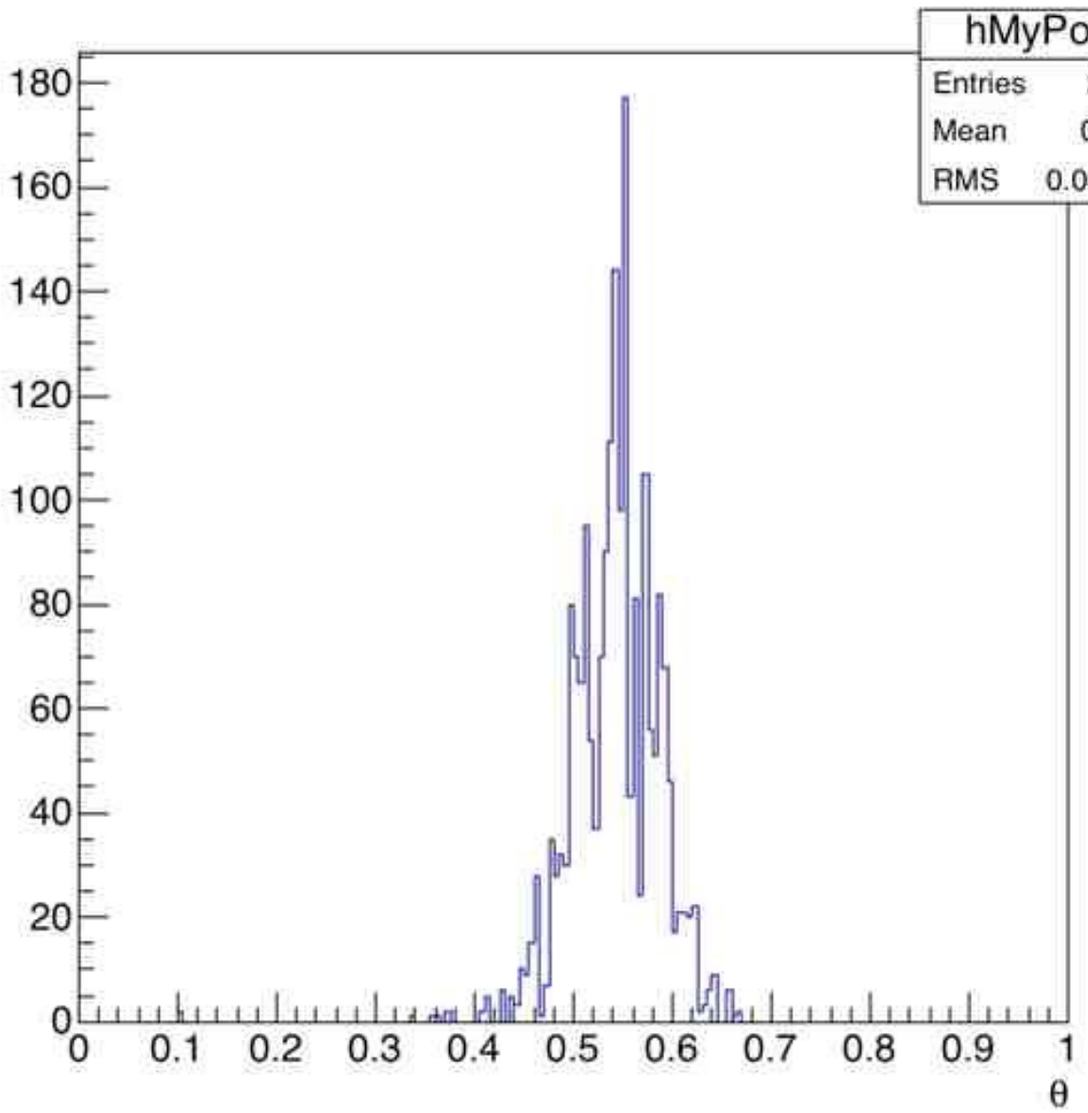
Course Evaluation & Break

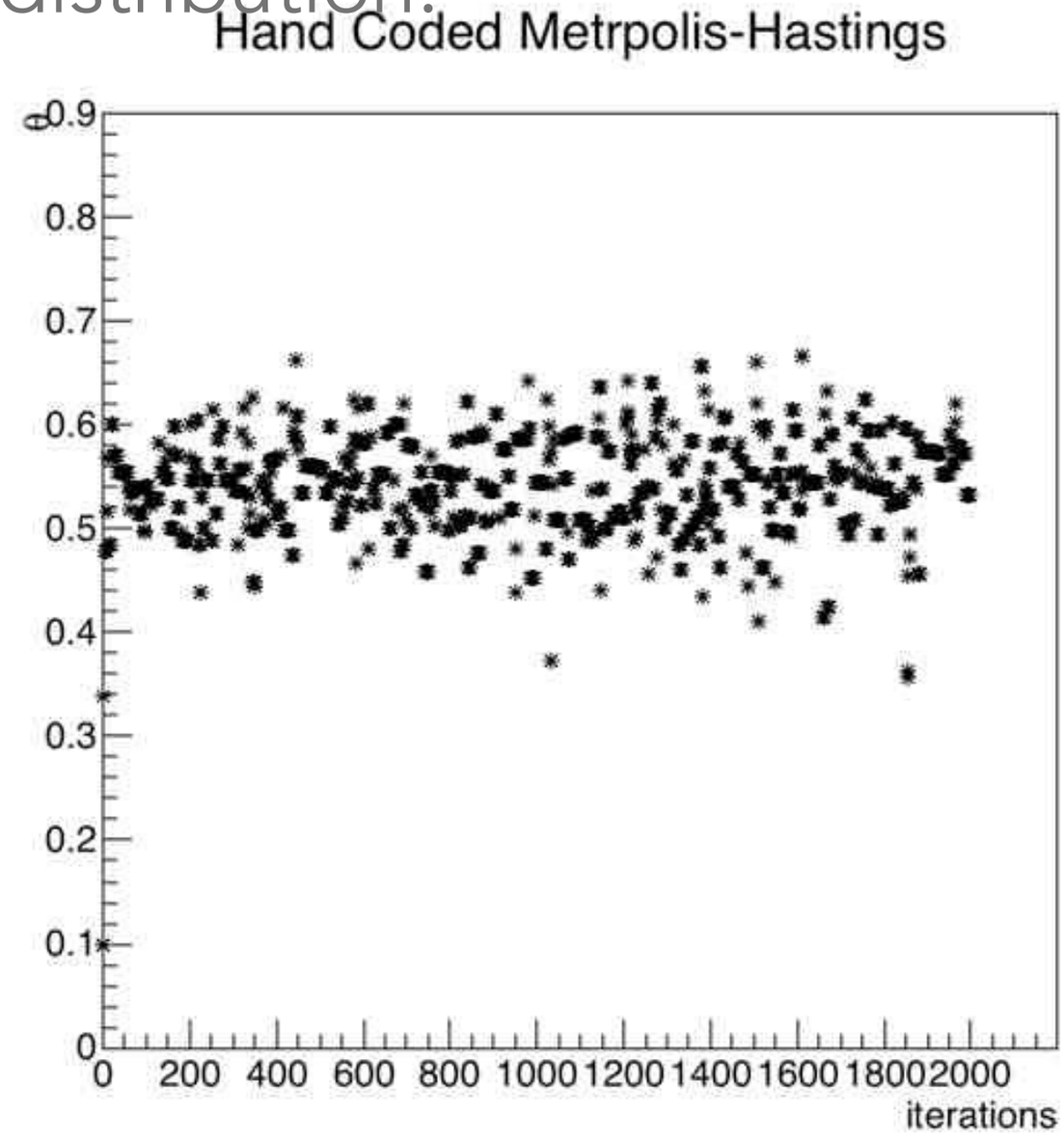
- Please do the course evaluation, it helps identify things to keep (or change) in the course

Exercise #3 (cont.)

*Reminder from the lecture about Markov Chain Monte Carlo

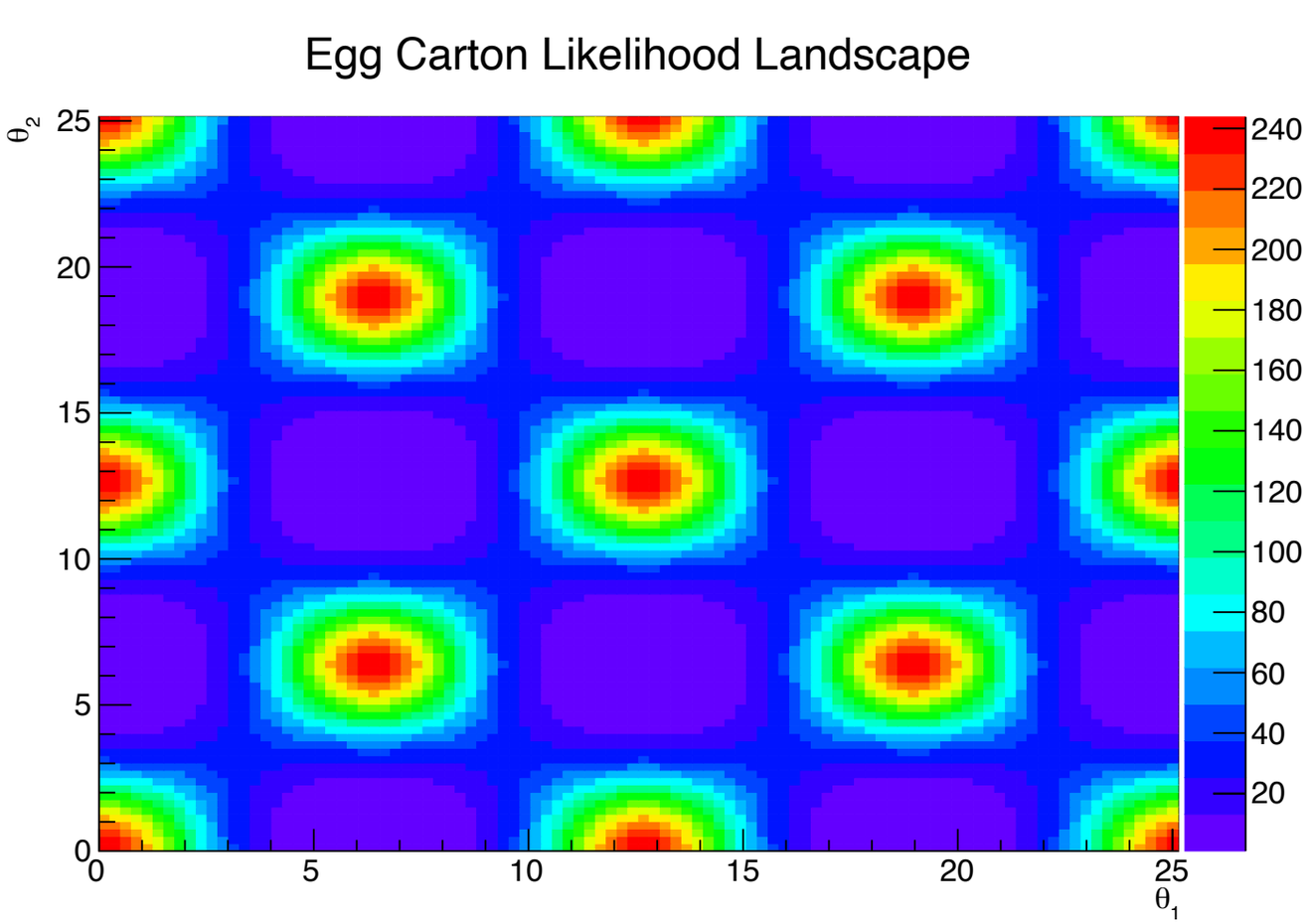
- For 2000 iterations plot Markov Chain Monte Carlo samples as a function of iteration, as well as a histogram of the samples, i.e. the posterior distribution.

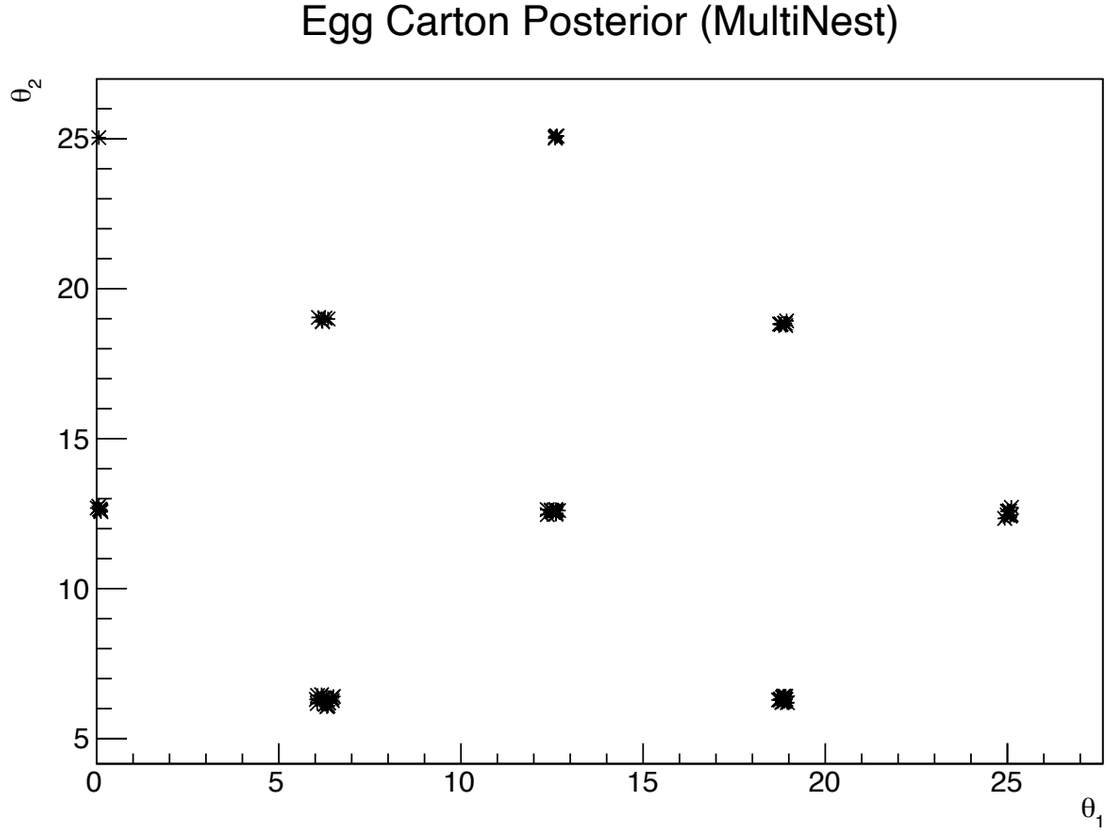




Nested Sampling in Action

- The ‘Egg Carton’ likelihood landscape is a benchmark test-statistic landscape for difficulty and 'stress testing’ of bayesian sampling technique's





Nested Sampling Benefits

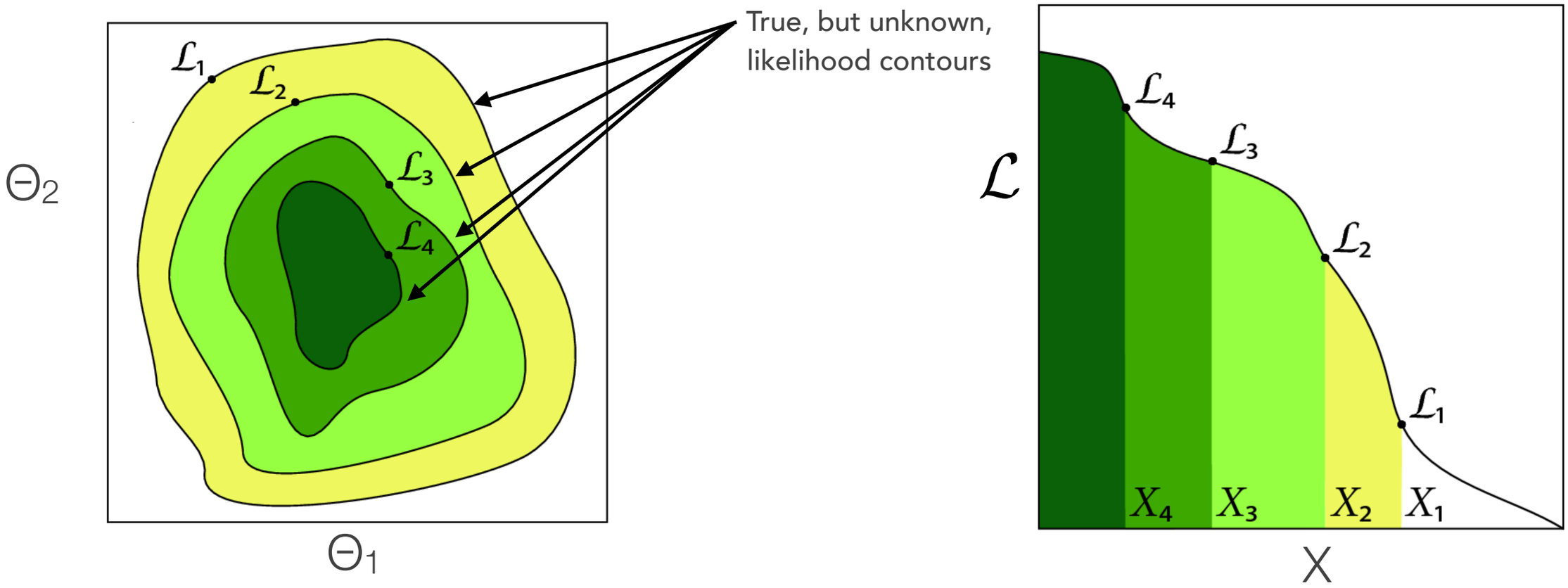
- Samples sparsely in low likelihood regions and samples densely where the likelihood is high
- Can handle irregular likelihood/test-statistic landscapes
- Many applications require nothing more than setting the range over which to generate ‘live points’
  - Does not require lots of tuning
  - Often, the initial sampling prior is uniform, i.e. flat
- The true value of the best-fit parameters or hypothesis estimator is not essential to be known, it just needs to be within the region where the points are sampled
- Efficient when compared to other MCMC methods

Cons

- Similar to every other fitting technique, there is no guarantee that any best-fit values are global best-fit values
- No rigorous termination criterion
  - There is always the possibility that there exist some unsampled regions in X which have ‘large’ likelihood values $\mathscr{L}(X)$ which will contribute to the bayesian evidence value Z
- Trapezoidal summing will induce some uncertainty and *possibly* small bias

Big Issue

- How do we actually sample new nested points X' that are better than the current X_{lowest}, where X_{lowest} has the lowest likelihood?
- In N-dimensions and without knowing the true likelihood contours, this is problematic.

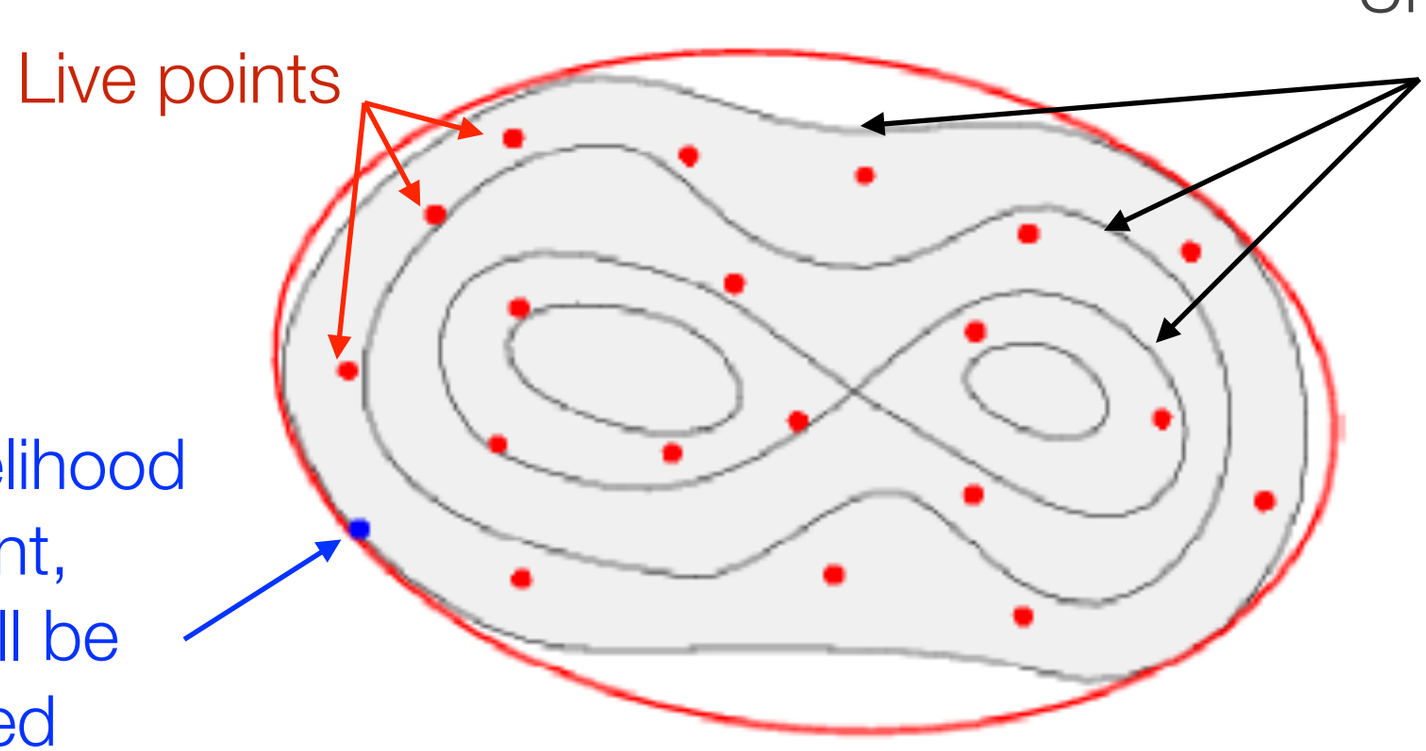


MultiNest Application

- Crude nested sampling was somewhat inefficient when it came to multi-modal likelihood landscapes
  - But, much better than conventional maximum likelihood fitters when it comes to not getting stuck in local minima
- Instead of using a multi-dimensional uniform prior for each replacement point, use an N-dimensional ellipsoid for resampling
  - The hyper-ellipsoid is defined by the current iteration live points
  - The hyper-ellipsoid for re-sampling has a small enlargement margin as a safeguard
  - If the test-statistic landscape becomes disjoint, create additional hyper-ellipsoids for sampling

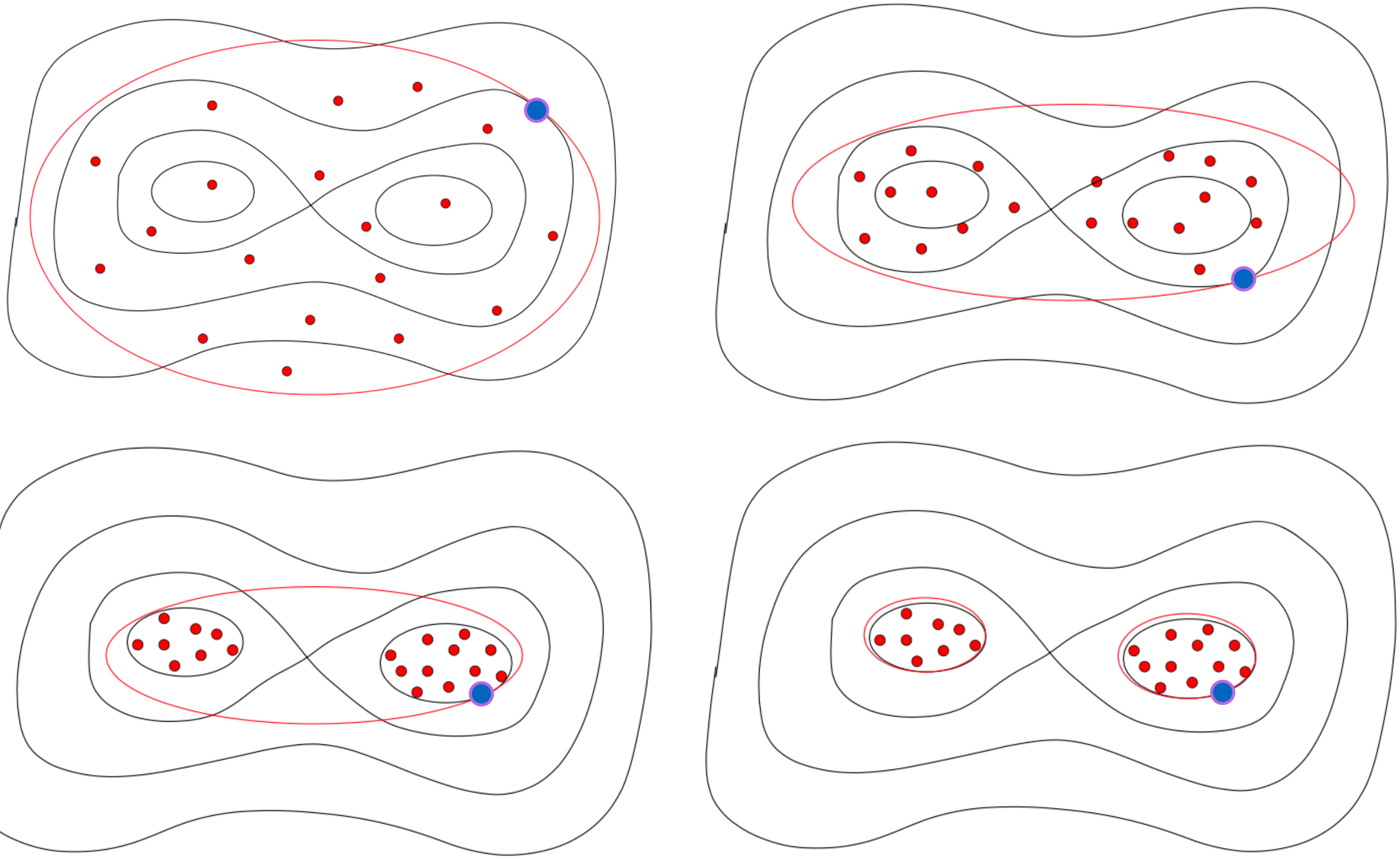
MultiNest Ellipsoid Sampling

- Start with a sample of live points using a uniform prior in an n-dimensional hyper-cube
- After a few iterations resampling within an ellipsoid we have:



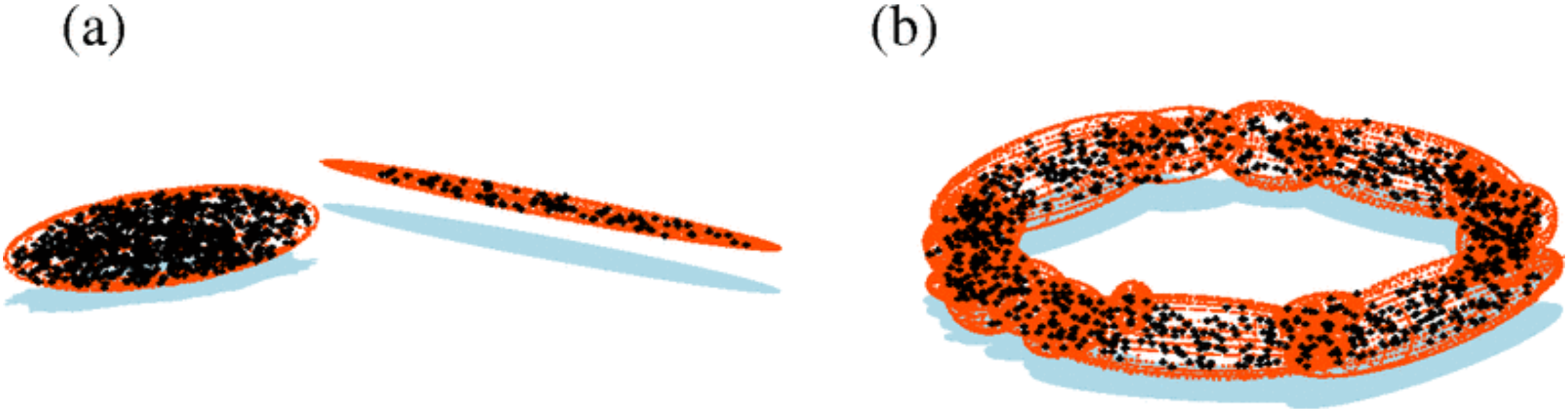
*F. Feroz

MultiNest Evolution



MultiNest Pictures

- Dis-joint regions, as in Fig. (a), as well as multi-dimensional multi-modal regions, as in Figs. (a) and (b), can be found efficiently without continual resampling of the whole space



Nested Sampling

- Can be an excellent method to map out a likelihood/ probability landscape that is complicated
- MultiNest is very nice, but the base package requires Fortran, even though there are nice wrapper packages in other software languages

Packages

- In Python there are a handful of nestling sampling packages
  - nestle (http://kbarbary.github.io/nestle/)
  - UltraNest (https://github.com/JohannesBuchner/UltraNest & https://arxiv.org/abs/2101.09604)
  - Dynesty (https://dynesty.readthedocs.io/en/stable/index.html)

Exercise Egg Carton

- The task is to produce a posterior-like distribution using a (hopefully) nested sampling algorithm for the classic 2-dimensional egg carton likelihood

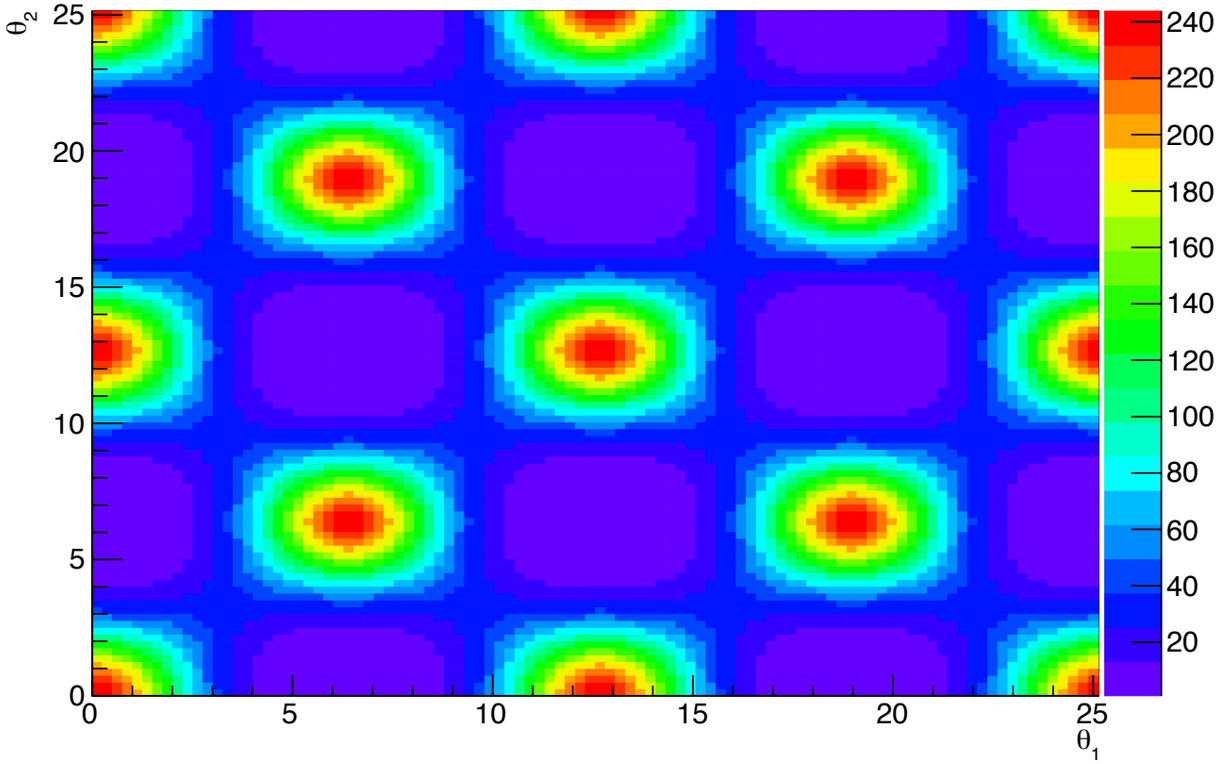
$$\mathcal{L}(\theta_1, \theta_2) \propto \cos(\theta_1)\cos(\theta_2)$$

- First, make sure you have a nested sampling algorithm package installed
- Second, make a plot of the raster scan of the the 2-D likelihood for reference
- Third, make a plot of the posterior-like distribution from the sampling algorithm

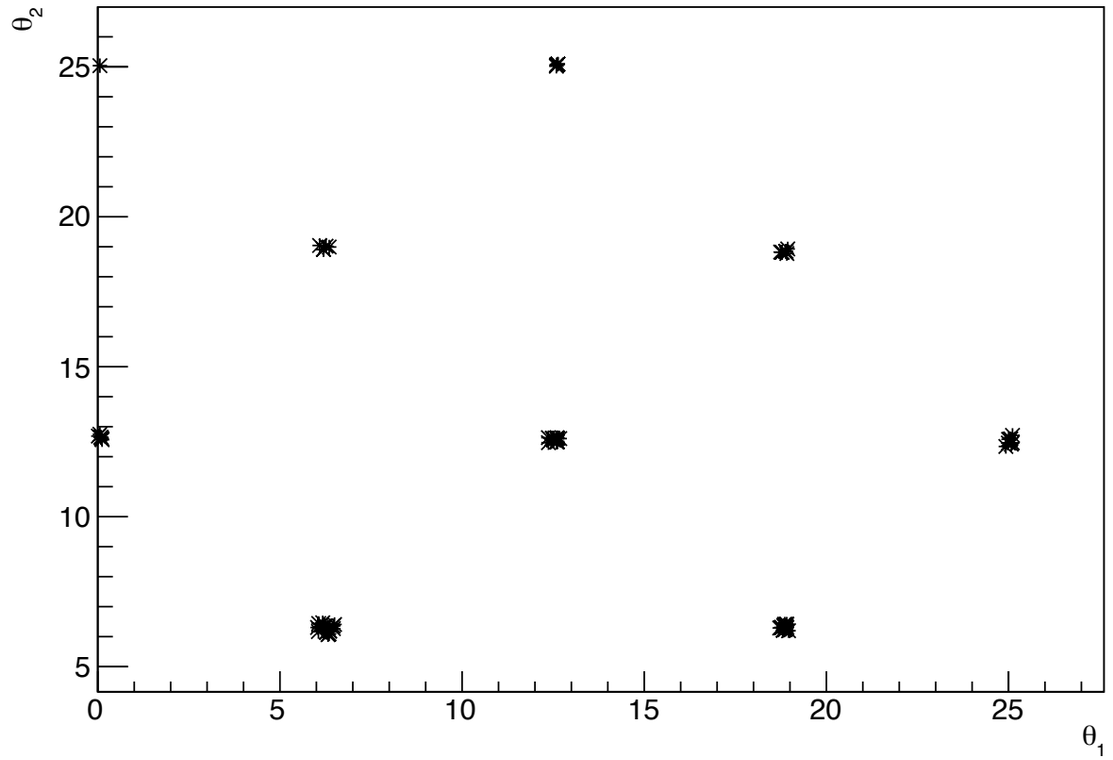
Exercise Egg Carton cont.

- The raster scan across θ_1 and θ_2 and the posterior distribution

Egg Carton Likelihood Landscape



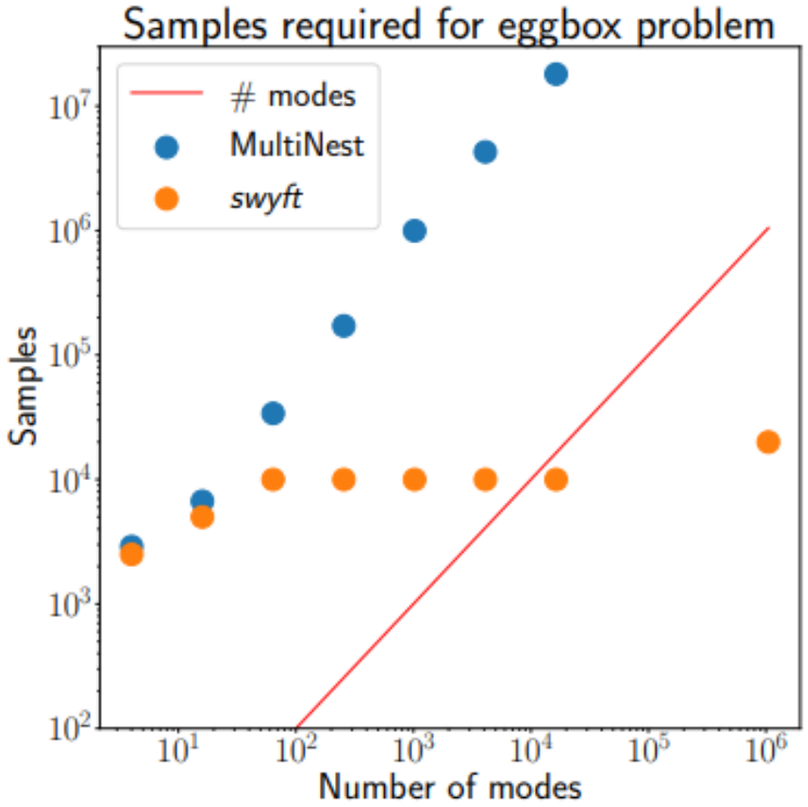
Egg Carton Posterior (MultiNest)

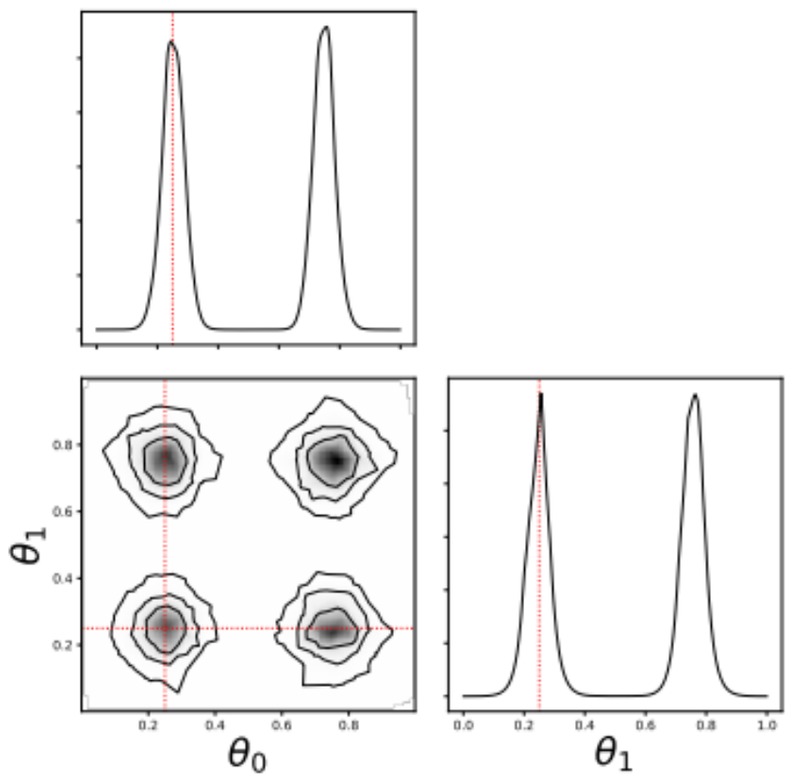


Algorithmic Comparisons to Multinest

This method can give inference super powers

- Consider a high-dimensional eggbox posterior, with two modes in each direction. Assuming 20 parameters, this give $2^{20} \sim 10^6$ modes.
- We can effectively marginalize over likelihoods with 1 Mio modes, using only 10 thousand samples.





*C. Weniger, 2022 AMAS Guest Lecture &
arXiv:2011.13951

Exercise Gaussian Shell/Cylinder

- Another example is the 2- or 3-dimensional gaussian shell
  - The probability is highest, i.e. centered, on the surface of a sphere or cylinder, and has a gaussian width
  - Looking at 3D gaussian surfaces is tough, so we will do a projection into 2D for visualization

$$\mathcal{L}(\vec{\theta}) = \mathrm{circ}(\vec{\theta}; \vec{c}, r, \sigma)$$

$$\mathrm{circ}(\vec{\theta}; \vec{c}, r, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(|\vec{\theta} - \vec{c}| - r)^2}{2\sigma^2} \right]$$

c is the center of the sphere/cylinder

r is the radius

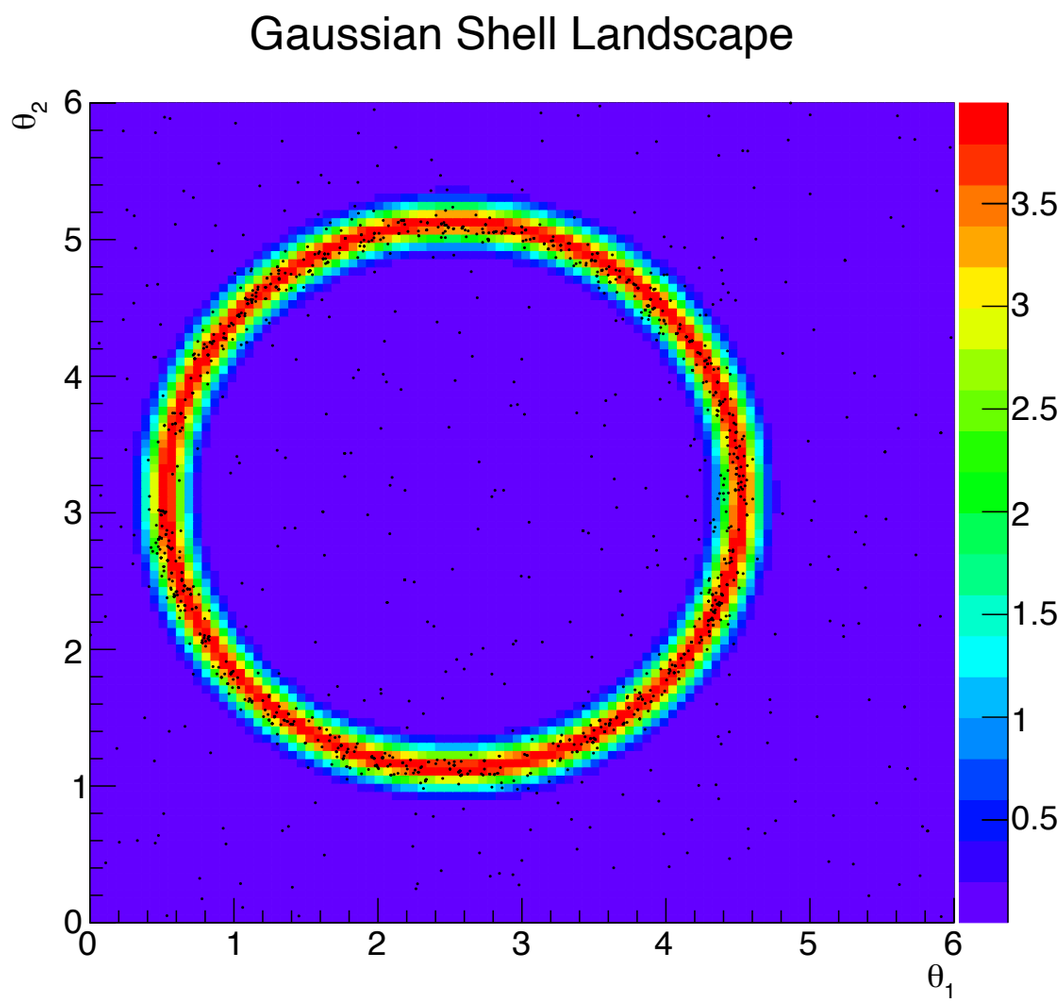
σ is the gaussian width

θ is a/the sample point as a vector, e.g. (x,y,z,…) in cartesian coordinates

$|\vec{\theta} - \vec{c}|$ is the norm

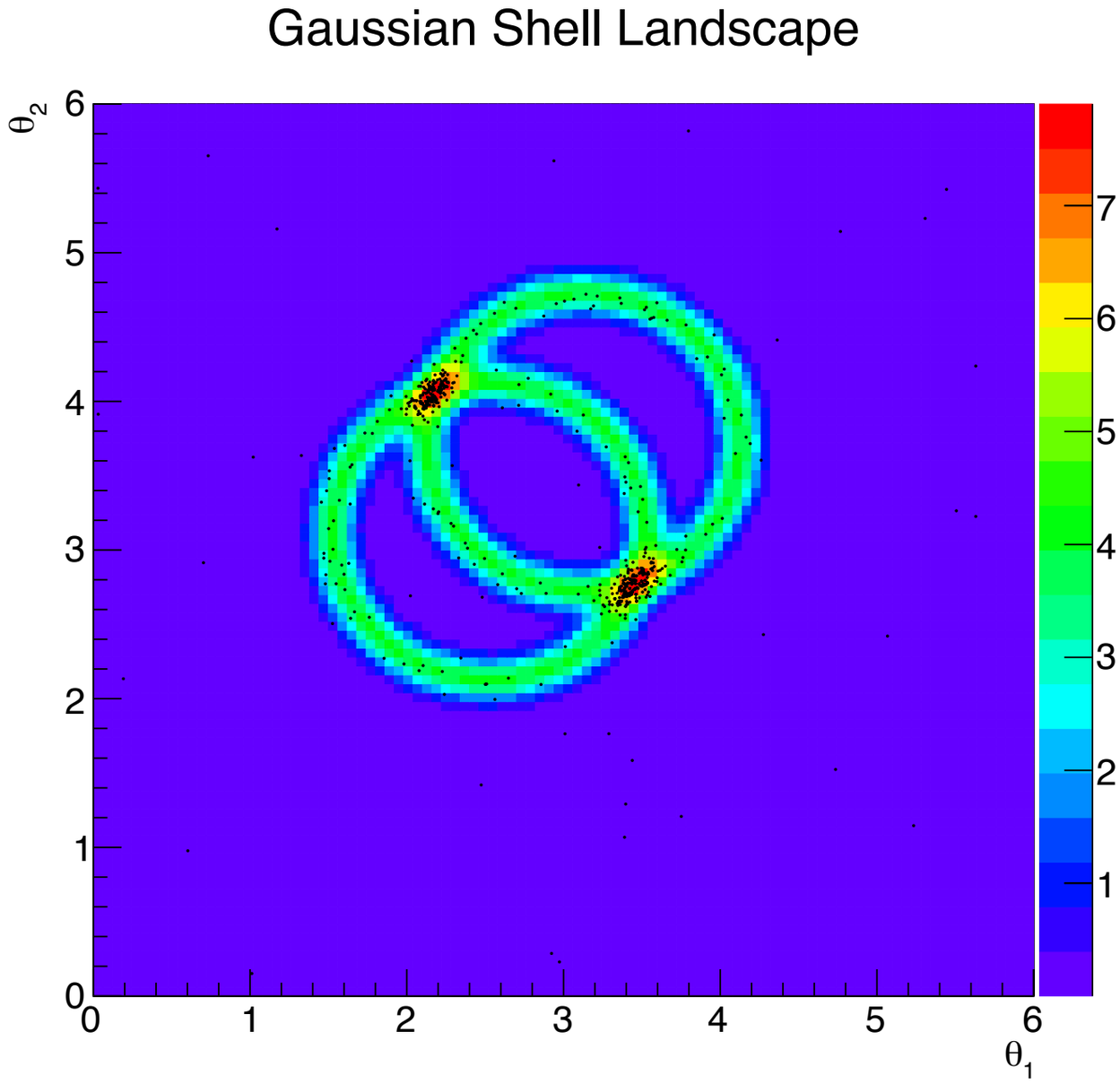
Exercise Gaussian Shell/Cylinder cont.

- Similar to the Egg Carton exercise generate the following plots:
  - For a single cylinder/sphere of r=2, σ=0.1, centered at c=(2.5, 3.1)
  - Plot the underlying probability/likelihood space
  - Plot the posterior sampling
- Note that there might be issues with the computer/machine precision when calculating exp() or ln() for negative, extremely large, or extremely small values related to the likelihood



Exercise Gaussian Shell/Cylinder cont.

- Repeat the previous task with two overlapping spheres/ cylinders
  - For r=1, σ=0.1, with one centered at c1=(2.5, 3.1) and the other at c2=(3.1, 3.7)



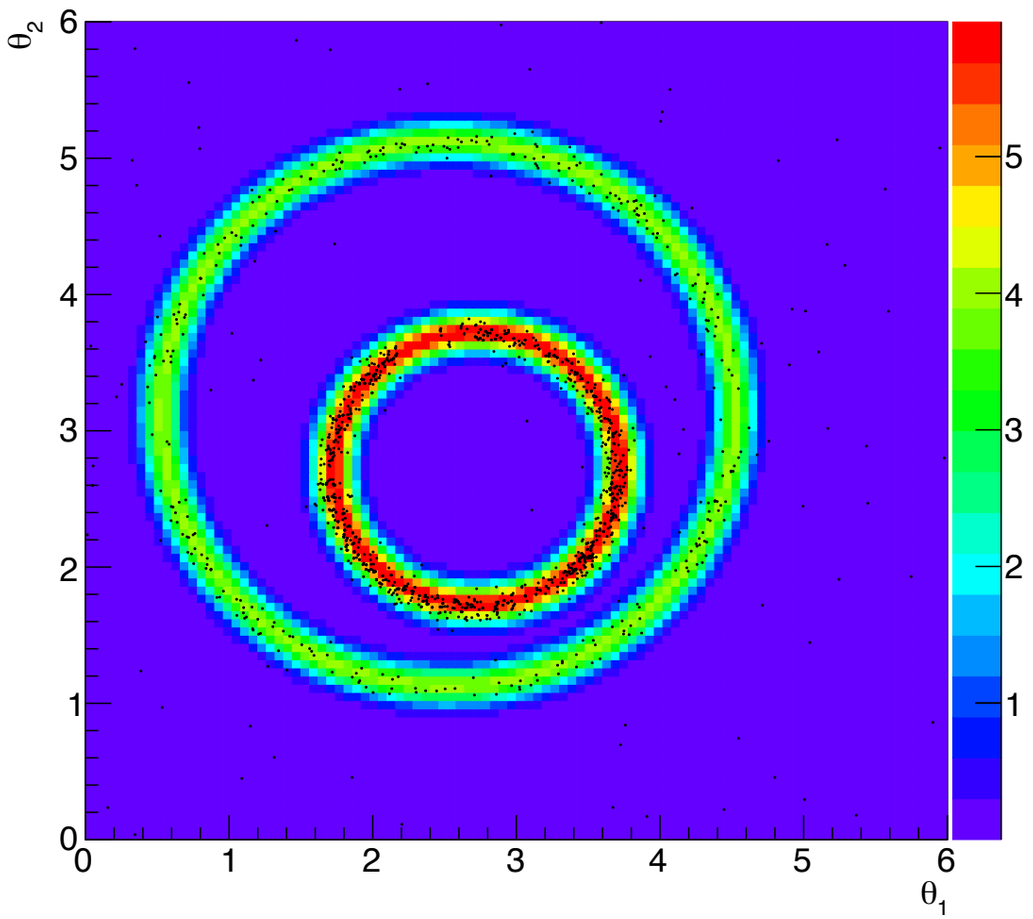
Exercise Nested Cylinder

- Using the following likelihood for the two cylinders plot the underlying likelihood and posterior distribution:

$$\mathcal{L}(\vec{\theta}) = \mathrm{circ}(\vec{\theta}; \vec{c}_1, r_1, \sigma_1) + 1.5\ \mathrm{circ}(\vec{\theta}; \vec{c}_2, r_2, \sigma_2)$$

  - $\sigma_{1,2}$=0.1, c_1=(2.5, 3.1) and c_2=(2.7, 2.7) and r_1=2 and r_2=1

Gaussian Shell Landscape



Extra

- Try higher dimensionality landscapes, e.g. 16-dimensions, and see if the sampler starts to slow down dramatically for the gaussian shell hyper-sphere likelihood

References

- Excellent and readable paper by developer John Skilling
  - http://projecteuclid.org/euclid.ba/1340370944
- MultiNest
  - Slides by F. Feroz (http://www.ics.forth.gr/ada5/pdf_files/Feroz_talk.pdf)
  - Papers (http://arxiv.org/abs/0809.3437, http://arxiv.org/abs/1306.2144)
- "Nested Sampling Methods" by Johannes Buchner
  - https://arxiv.org/pdf/2101.09675.pdf