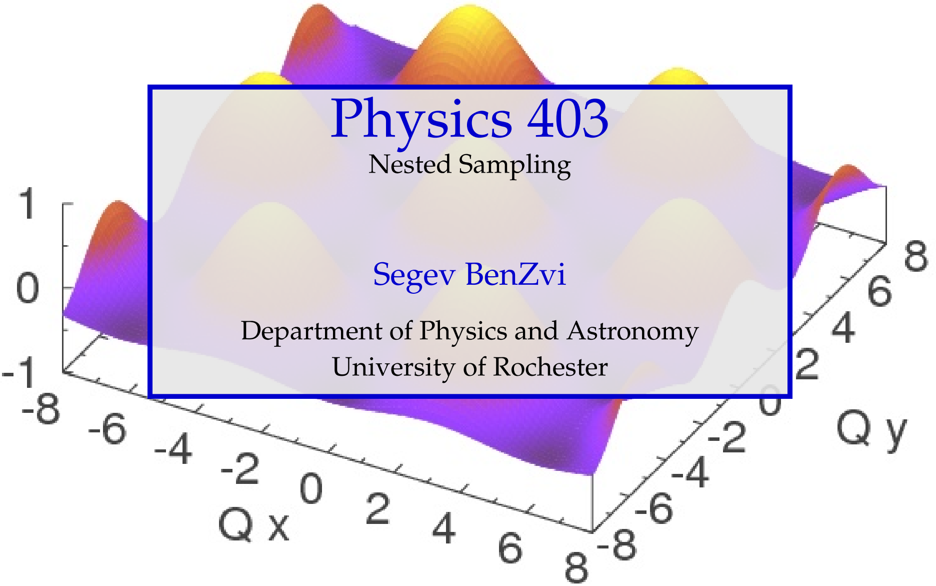
# Physics 403

Nested Sampling


Segev BenZvi

Department of Physics and Astronomy
University of Rochester

# Reading

- Sivia and Skilling, Ch. 9

# Evaluating Full Posterior Distributions

Recall the types of calculations we often have to do in a Bayesian analysis (from [1]):

$$
\begin{array}{ccccc}
p(D|\boldsymbol{x},I)\;p(\boldsymbol{x}|I) & = & p(D,\boldsymbol{x}|I) & = & p(D|I)\;p(\boldsymbol{x}|D,I) \\
\mathcal{L}(\boldsymbol{x}) \;\times\; \pi(\boldsymbol{x}) & = & \ldots & = & Z \;\times\; p(\boldsymbol{x}) \\
\text{likelihood} \times \text{prior} & = & \text{joint} & = & \text{evidence} \times \text{posterior} \\
\text{INPUT} & \Longrightarrow & \ldots & \Longrightarrow & \text{OUTPUT}
\end{array}
$$

To fully evaluate the posterior $p(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})/Z$ we have to evaluate integrals of the form

$$Z = \iint \ldots \int d\boldsymbol{x}\;\mathcal{L}(\boldsymbol{x})\;\pi(\boldsymbol{x})$$

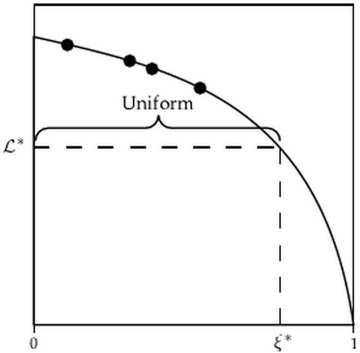
Often this can only be done numerically, so we need an efficient method of calculating high-dimensional integrals

# Nested Sampling

- Nested sampling is another kind of technique useful for high-dimensional integration and posterior sampling [2, 3]
- Advantages over MCMC: can handle pathologies in parameter spaces such as strong non-linear correlations and requires fewer samples (up to a factor 100 less) for evidence calculation
- The algorithm gives results that allow for model selection as well as best parameter estimates at once
- Several packages available in Python [4, 5]
- Basic concept: use a likelihood ordering scheme to evaluate integrals like

$$Z = \iint \ldots \int d\boldsymbol{x}\ \mathcal{L}(\boldsymbol{x})\ \pi(\boldsymbol{x})$$

# Basics of Nested Sampling



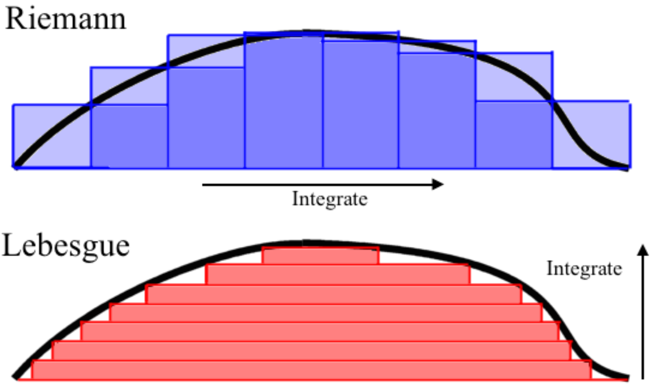

- Sample $N$ objects $\boldsymbol{x}$ with respect to the prior such that $\mathcal{L}(\boldsymbol{x}) > \mathcal{L}^*$
- Start with $\mathcal{L}^* = 0$, so that sampling begins over the entire prior
- We uniformly sample $\xi(\mathcal{L}^*)$, the proportion of the prior with likelihood greater than $\mathcal{L}^*$:

$$\xi(\mathcal{L}^*) = \iint\limits_{\mathcal{L}(\boldsymbol{x})>\mathcal{L}^*} \ldots \int \pi(\boldsymbol{x})\, d\boldsymbol{x}$$

- Slowly increase $\mathcal{L}^*$ so that we end up sampling in the high probability region

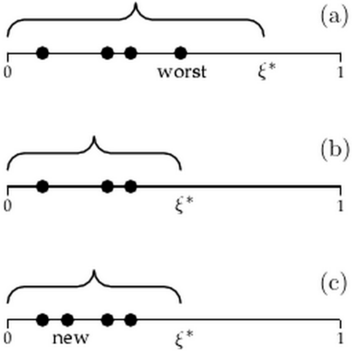
# Analogy: Riemann and Lebesgue Integration

The concept is similar to Lebesgue integration




Rather than partition the domain of $\mathcal{L}$ into subintervals, we partition the range of $\mathcal{L}$ and integrate "up the hill"

# Iteration Step




The algorithm in practice:

- Start with $N$ objects restricted to $\xi < \xi^*$
- Select the object with the largest $\xi$ (and hence smallest $\mathcal{L}$)
- Use the worst object's $(\xi, \mathcal{L})$ as the new $(\xi^*, \mathcal{L}^*)$ and then toss out the worst object
- There are now $N - 1$ objects in the new domain bounded by $\xi^*$, which is nested inside the old domain
- Generate a new object inside the smaller domain by uniformly sampling the prior
- Restart the loop, and proceed until $\mathcal{L}^* = \mathcal{L}_{\max}$

# Calculation of Marginal Evidence

- The shrinkage ratio $t = \xi/\xi^*$ at each iteration is distributed as

$$p(t) = Nt^{N-1}, \quad \text{with mean } \ln(t) = (-1 \pm 1)/N$$

- At each iteration $k$,

$$\mathcal{L}_k = \mathcal{L}^* \quad \text{and} \quad \xi_k = \xi^* \prod_{j=1}^{k} t_j$$

- Each shrinkage ratio is independently distributed according to $p(t)$ so

$$\ln \xi_k = (-k \pm \sqrt{k})/N$$

- If $\ln t = -1/N$ then $\xi_k = \exp(-k/n)$, and we can evaluate

$$Z = \int_0^1 \mathcal{L}(\xi)\, d\xi \approx \sum_k h_k\, \mathcal{L}_k,$$

where $h_k = \xi_{k-1} - \xi_k = \Delta\xi_k$

# Generating Quantities from the Posterior Distribution

- Each sequence in the parameter space $\{\boldsymbol{x}_k\}$ has an associated weight

$$w_k = \frac{h_k\,\mathcal{L}_k}{Z}$$

where $h_k = \Delta\xi_k$ and $Z = \sum h_k\,\mathcal{L}_k$

- The weights define the posterior PDF. Any quantity $f(\boldsymbol{x})$ can be generated from the posterior in the usual way:

$$\langle f\rangle = \sum_k w_k f(\boldsymbol{x}_k)$$

$$\langle f\rangle = \sum_k w_k f^2(\boldsymbol{x}_k)$$

$$\mathrm{var}\,(f) = \langle f^2\rangle - \langle f\rangle$$

# Uncertainty in $Z$

- Given the estimate of $Z$, we can calculate the information or negative entropy

$$\mathcal{H} = \int p(\xi) \ln [p(\xi)] \, d\xi \approx \sum_k \frac{h_k \mathcal{L}_k}{Z} \ln \left[ \frac{\mathcal{L}_k}{Z} \right]$$
$$\approx (\#\text{ active components in data}) \times \ln (\text{signal}/\text{noise})$$

- If we count until $k = N\mathcal{H}$ then the accumulated values of $\ln \xi$ are subject to an uncertainty $\sqrt{N\mathcal{H}}/N$
- This uncertainty also applies to $\ln Z$, so that

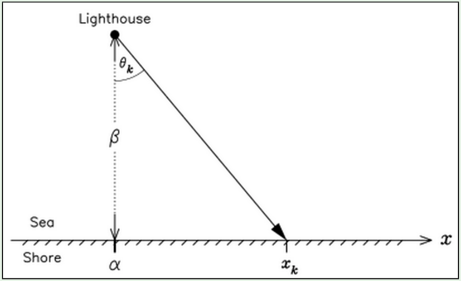
$$\ln Z \approx \ln \left( \sum_k h_k \mathcal{L}_k \right) \pm \sqrt{\frac{\mathcal{H}}{N}}$$

- Convergence criterion: no rigorous approach. Use your judgment. Typical: choose upper limit on the number of iterations

# Lighthouse Problem

## Example

A lighthouse is somewhere off the coast at position $\alpha$ along the shore and $\beta$ out to sea. It emits a series of short collimated flashes at random intervals (and hence, random azimuths)




$N$ flashes are detected at positions $\{x_k\}$ along the coast. Given the $\{x_k\}$, where is the lighthouse?

# Parameterization of the Lighthouse Problem

- Since the lighthouse emissions are random, the azimuth angle of the $k^{\text{th}}$ emission is uniform over $\theta = \pm 90^\circ$:

$$p(\theta_k|\alpha, \beta, I) = 1/\pi$$

- The azimuth angle is related to the position along the coast $x_k$ by

$$\beta \tan\theta_k = x_k - \alpha$$

- Change variables to find the likelihood of the $x_k$:

$$\begin{aligned}
p(x_k|\alpha, \beta, I) &= p(\theta_k|\alpha, \beta, I) \left| \frac{\partial\theta_k}{\partial x_k} \right| \\
\beta \sec^2\theta \frac{\partial\theta}{\partial x} &= 1 \\
\beta[1 + \tan^2\theta] \frac{\partial\theta}{\partial x} &= \beta \left[ 1 + \left( \frac{x - \alpha}{\beta} \right)^2 \right] \frac{\partial\theta}{\partial x} = 1
\end{aligned}$$

# Parameterization of the Lighthouse Problem

- Using the Jacobian we find the likelihood of the $x_k$:

$$p(x_k|\alpha,\beta,I) = \frac{\beta}{\pi\left[\beta^2 + (x_k-\alpha)^2\right]}$$

$$p(\boldsymbol{x}|\alpha,\beta,I) = \prod_{k=1}^{N} p(x_k|\alpha,\beta,I)$$

- What we really want is the posterior distribution of $\alpha$:

$$p(\alpha,\beta|\boldsymbol{x},I) = \frac{1}{Z} p(\boldsymbol{x}|\alpha,\beta,I)\, p(\alpha,\beta|I),$$

where we expect that $p(\alpha,\beta|I) = p(\alpha|I)p(\beta|I)$ is uniform:

$$p(\alpha,\beta|I) = \begin{cases} \frac{1}{\alpha_{\max}-\alpha_{\min}}\frac{1}{\beta_{\max}-\beta_{\min}}, & \alpha\in[\alpha_{\min},\alpha_{\max}], \beta\in[\beta_{\min},\beta_{\max}] \\ 0 & \text{otherwise} \end{cases}$$

# Calculating the Likelihood

The likelihood we use for nested sampling is

$$\mathcal{L}(\alpha, \beta) = \prod_{k=1}^{N} \frac{\beta}{\pi\left[\beta^2 + (x_k - \alpha)^2\right]}$$

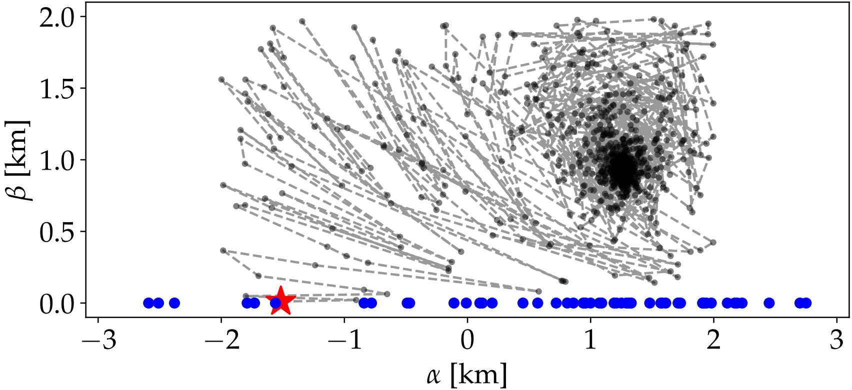
$$\ln \mathcal{L} = \ln \beta - \ln \pi - \sum_{k=1}^{N} \left(\beta^2 + (x_k - \alpha)^2\right)$$

The algorithm we apply is:

1. Generate $N$ values of $\alpha$ and $\beta$ from the uniform priors
2. Calculate $\mathcal{L}$ (or $\ln \mathcal{L}$) using the $N$ points and the $\{x_k\}$
3. Pick the value with the lowest $\mathcal{L}$ and set it to $\mathcal{L}^*$
4. Use $\mathcal{L}^*$ to estimate new limits $\alpha^*$ and $\beta^*$ and generate new values of $\alpha$ and $\beta$ subject to these limits. Proceed until termination

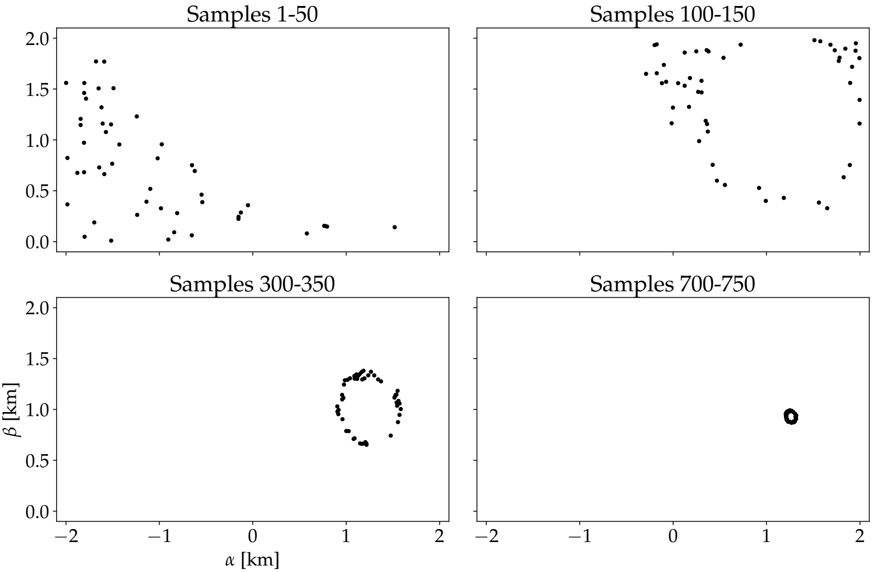
# Lighthouse Problem

Chooose $\alpha \in [-2, 2]$ and $\beta \in [0, 2]$. Update $\alpha$ and $\beta$ with uniform steps (easy to implement; could have used a Gaussian)




$(\alpha, \beta)$ moves from starting point (red star) to the region of highest probability

# Sampling of the Posterior vs. Time



# Best Estimate of $\alpha$, $\beta$

- Using the liklihood weights from each sample

$$w_k = \frac{h_k\,\mathcal{L}_k}{Z}$$

  we can get the mean $\alpha$ and $\beta$:

$$\langle\alpha\rangle = \sum_k w_k \alpha_k = 1.25 \pm 0.18\text{ km}$$
$$\langle\beta\rangle = \sum_k w_k \beta_k = 1.01 \pm 0.20\text{ km}$$

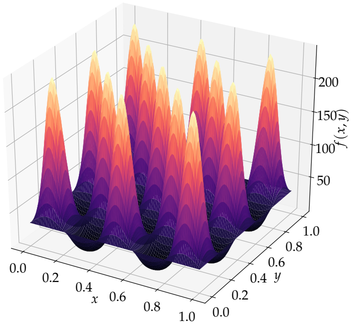
- The estimate of the evidence $\ln Z$ is

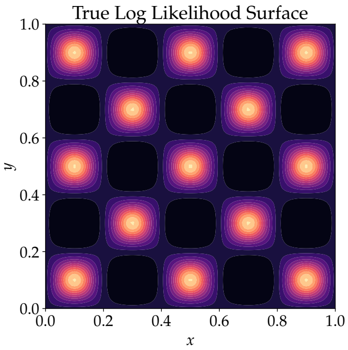
$$\ln\left(Z/\text{km}^{64}\right) = -160.53 \pm 0.17$$

- Note that $Z$ has dimensions of km$^{64}$ because of the 64 $\{x_k\}$

# Highly Multimodal Distributions

Handles very multimodal distributions like the eggbox function

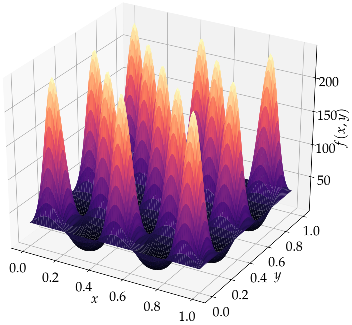


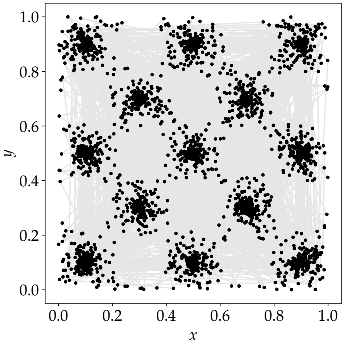




Note: the acceptance rate for points $\mathcal{L} > \mathcal{L}^*$ can be poor unless some effort is made to split up the sampling region

# Highly Multimodal Distributions

Handles very multimodal distributions like the eggbox function




Note: the acceptance rate for points $\mathcal{L} > \mathcal{L}^*$ can be poor unless some effort is made to split up the sampling region

# References I


[1] D.S. Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press, 1998.

[2] J. Skilling. "Nested Sampling". In: *Proc. Bayesian Inference and Maximum Entropy Methods*. Vol. 735. Garching, Germany: AIP, July 2004, p. 395.

[3] J. Skilling. "Nested sampling for general Bayesian computation". In: *Bayesian Anal.* 1.4 (Dec. 2006), pp. 833–859.

[4] F. Feroz et al. *MultiNest: Efficient and Robust Bayesian Inference*. 2015. URL: http://ccpforge.cse.rl.ac.uk/gf/project/multinest/.

[5] K. Barbary. *Nestle Nested Sampling Package*. 2015. URL: https://github.com/kbarbary/nestle.