Nested Sampling

Skilling 2005

So far we have focused on the posterior prob. This method derives from another perspective.

inputs $P(D|M,I)\ P(M|I)$ Likelihood Prior $\mathcal{L}(x) \times \pi(x)$

$\downarrow$

$=$

$P(M,D|I)$ Joint Distribution

$=$

outputs $P(M|D,I)\ P(D|I)$ Posterior Evidence $P(x) \times Z$

Normalization implies that

$$\int P(M|I)\, dM = 1 \quad \text{and} \quad \int P(M|D,I)\, dM = 1$$

Together, this implies that

$$Z = \int P(D|M,I)\, P(M|I)\, dM = 1$$

$$Z = \int \mathcal{L}(x)\, \pi(M)\, dx$$

$$P(x) = \frac{\mathcal{L}(x)\, \pi(x)}{Z}$$

Nested Sampling : Sorted Likelihood

Consider a discrete 2D hypothesis space.
We assume a uniform prior and plot a map of the likelihood below

$\mathcal{L} =$

0	8	15	3
11	24	22	10
19	30	26	16
9	23	18	6

and sort the values into a vector

$\mathcal{L} = (30, 26, 24, 23, 22, 19, 18, 16, 15, 11, 10, 9, 8, 6, 3, 0)$

with

$\pi = \frac{1}{16} \times (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$

We now define the function

$\xi(\lambda) =$ proportion of the prior probability mass with likelihood greater than λ

Formally, this is

$$\xi(\lambda) = \int_{\mathcal{L}(x) > \lambda} \pi(x)\, dx$$

The element of prior mass is $d\xi = \pi(x)\, dx$

Nested Sampling: Density of States

Select $\lambda = 15$

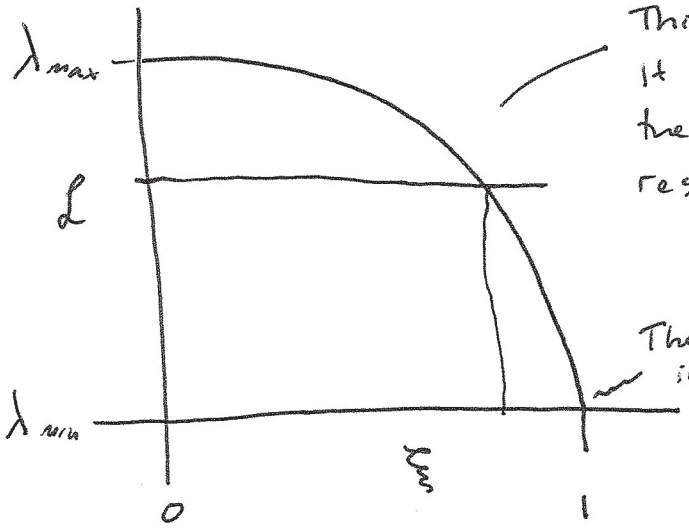
Then $\xi(\lambda=15) = 8/16 = 0.5$

We can look at the inverse of this function

$$\lambda = \xi^{-1}(\xi(\lambda))$$

and since this function returns a likelihood, Skilling calls $\xi^{-1} \equiv \mathcal{L}$

where $\mathcal{L}(\xi)$ is distinct from $\mathcal{L}(x)$



This is a "density of states" It describes how the states of the system are distributed with respect to their probability

The entire volume $\xi = 1$ includes all λ values $> \lambda_{min}$

Now $Z = \int \mathcal{L}(x)\,\pi(x)\,dx$

$\Rightarrow Z = \int_0^1 \mathcal{L}(\xi)\,d\xi$

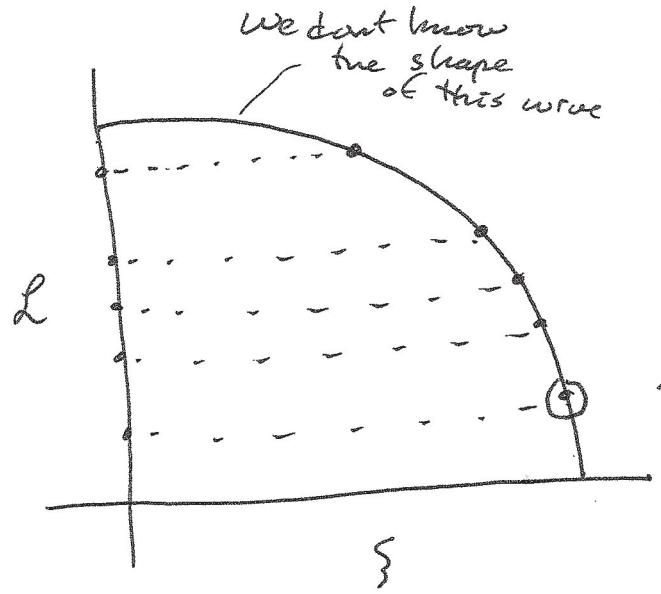
Let $\xi = \int \pi(x)\,dx$

$\Rightarrow d\xi = \pi(x)\,dx$

$\mathcal{L}(x) = \mathcal{L}(\xi)$

How do we perform this integral?

Nested Sampling: Stochastic Integration



model space

Choose N samples in X

Compute Likelihoods

Each sample represents $\frac{1}{N}$ of the total volume.

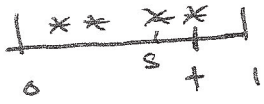
The worst sample defines $\mathcal{L}_k = \mathcal{L}^*$ which is the likelihood value that we now use as an implicit boundary.

Order Statistics

When the worst sample is rejected, the interval implicitly defined by $\mathcal{L}^*$ shrinks.

By how much?

Consider the unit interval



with n samples.

One sample is the rightmost sample.

What is the probability that s will be to the left of t?

$$P(s<t \mid I) = t$$

For $n-1$ samples

$$P(\text{all } n-1 < t \mid I) \propto t^{n-1}$$

$$\Rightarrow \text{Prob}\left(\tfrac{s}{s^*} = t\right) \propto t^{n-1}$$

$$P(t) = c\,t^{n-1}$$

$$\int_0^1 P(t)\,dt = 1$$

$$\int_0^1 c\,t^{n-1} = 1 \Rightarrow P(t) = n\,t^{n-1}$$

$$\Rightarrow c\,\frac{1}{n} = 1$$

$$\Rightarrow c = n$$

Order Statistics

Consider the unit interval covered with n samples...

Let t be the greatest sample

What is the probability that another sample $s < t$?

$$P(s<t \mid I) = t$$

For $n-1$ samples, what is the probability that $n-1$ samples are less than t?

$$P(s_1<t, s_2<t, \ldots, s_n<t \mid I) \propto \prod_{i=1}^{n-1} P(s_i<t \mid I)$$

Same as the Probability that t is the largest of n samples

$$\propto \prod_{i=1}^{n-1} t$$

$$P(t \mid I) \propto t^{n-1}$$

Normalize it

$$P(t \mid I) = c t^{n-1}$$

$$\int P(t \mid I)\, dt = \int_0^1 c t^{n-1} dt = 1$$

$$\Rightarrow c \frac{1}{n} = 1 \quad \Rightarrow \quad P(t) = n t^{n-1}$$

$$\Rightarrow c = n$$

Order Statistics

Find mean value of $\log t$ ...

$t = \xi/\xi^*$ shrinkage ratio

$P(t|I) = n t^{n-1}$

Let $s = \log t \Rightarrow \frac{ds}{dt} = \frac{1}{t}$

$P(s) = P(t)\left|\frac{dt}{ds}\right|$

$= t\,p(t)$

$= n t^n$

$= n e^{ns}$ (since $s = \log t \Rightarrow t = e^s$)

$\langle \log t \rangle = \langle s \rangle$

$= \int_{-\infty}^{0} s\, n\, e^{ns}\, ds$

$= \left.\frac{e^{ns}(ns-1)}{n}\right|_{-\infty}^{0}$

$= -\frac{1}{n} - 0$

Similarly for variance s.t. $\langle \log t \rangle = -\frac{1}{n} \pm \frac{1}{n}$

Nested Sampling

At the k^{th} iteration

$$\mathcal{L}_k = \mathcal{L}^*$$

$$\xi_k = \xi^* = \prod_{j=1}^{k} t_j$$

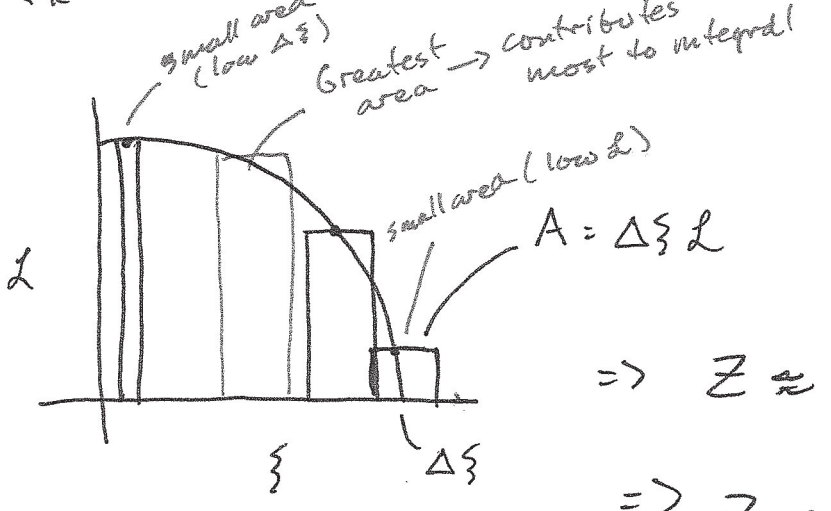
$$\Rightarrow \quad \log \xi_k = (-k \pm \sqrt{k})/n = -\frac{k}{n} \pm \frac{\sqrt{k}}{n}$$

Ignoring our uncertainty about $\log \xi_k$

$$\Rightarrow \quad \log \xi_k = -k/n$$

$$\Rightarrow \quad \xi_k = e^{-k/n}$$

Better to sample $\log \xi_k$!



$$\Rightarrow \quad Z \approx \sum_k \mathcal{L}_k \Delta\xi_k$$

$$\Rightarrow \quad Z \approx \sum_k A_k$$

where $\Delta\xi_k \approx \xi_{k-1} - \xi_k$

The entropy can also be found:

$$\mathcal{H} = \int P(\xi) \log[P(\xi)] \, d\xi \approx \sum_k \frac{A_k}{Z} \log\left[\frac{\mathcal{L}_k}{Z}\right]$$

$$\Rightarrow \boxed{\log Z \approx \log\left(\sum A_k\right) \pm \sqrt{\mathcal{H}}}$$

Nested Sampling : Key Points

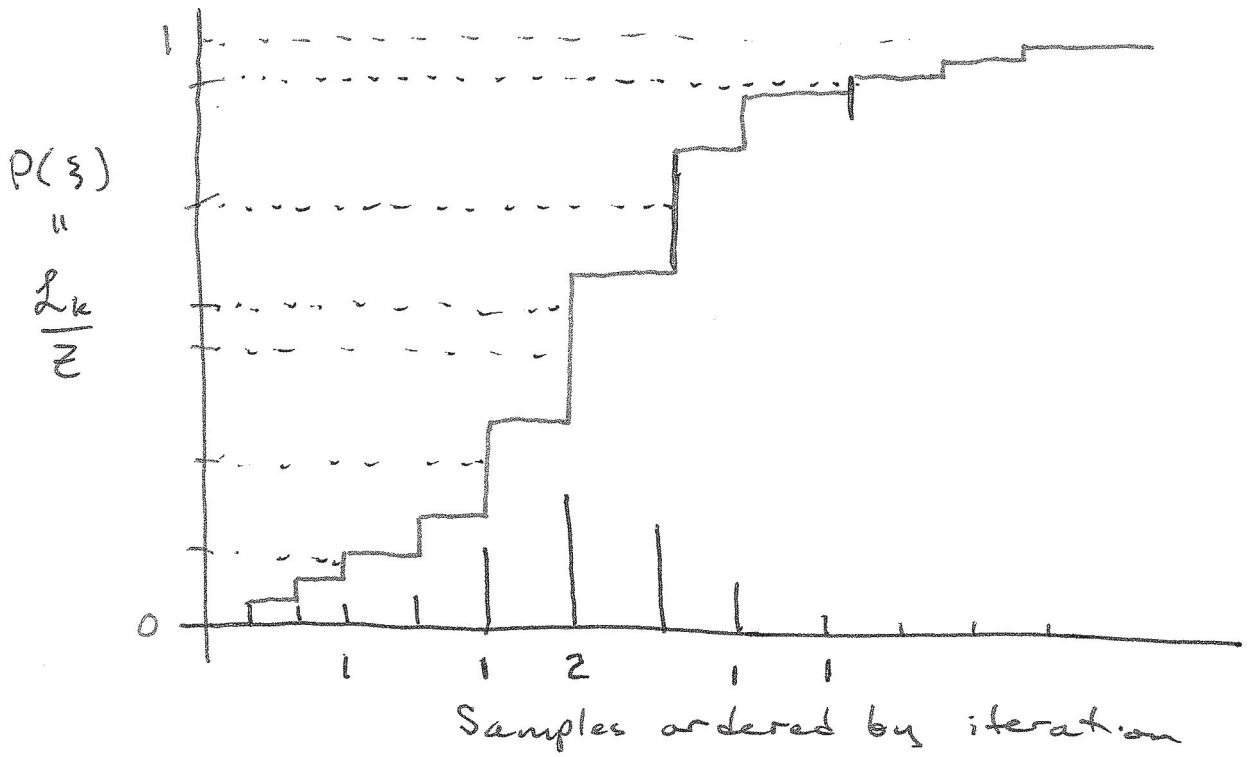
- The samples must be selected from the prior probability. In this implementation, it is assumed that the prior was uniform.

Often one can perform a change of variables so that your priors are uniform.

- The algorithm is focused on integrating to get the evidence

- The set of thrown-away samples DO NOT constitute a set of samples distributed according to the posterior probability.

Sampling from the Posterior using Nested Sampling

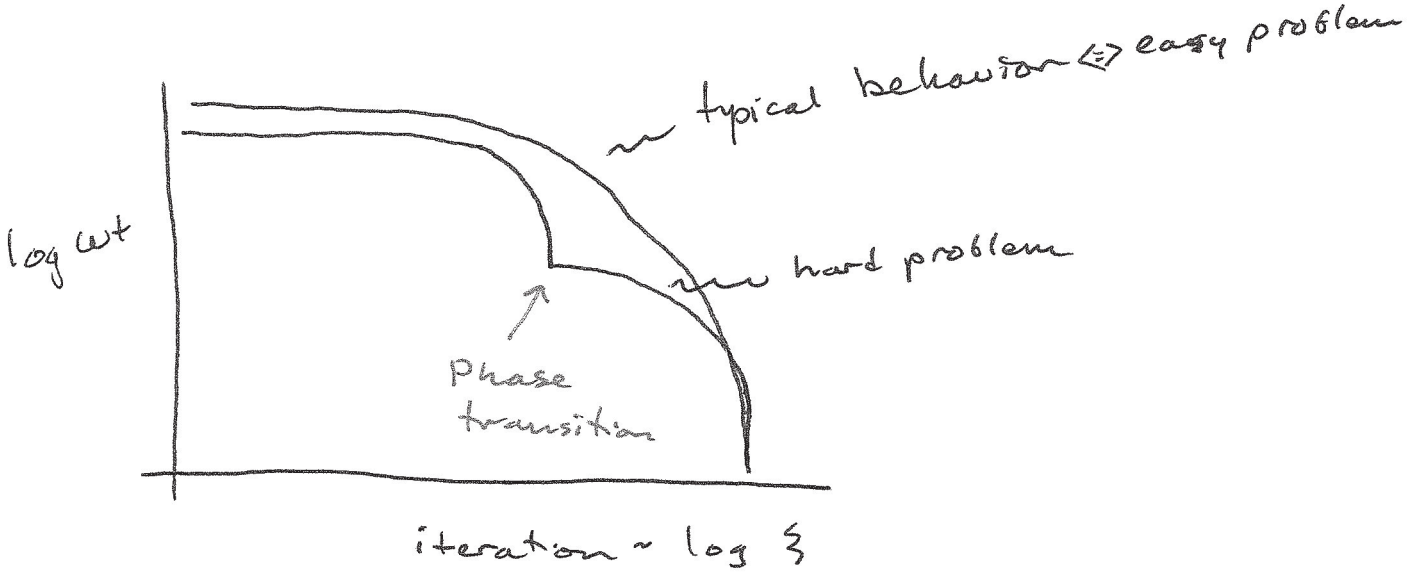


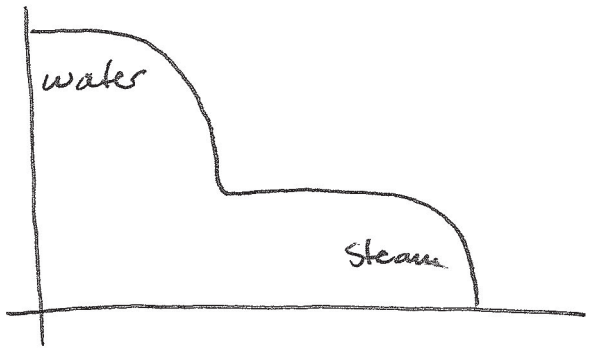
1. Imagine the cumulative distribution function, which is the integral of the pdf from $-\infty$ to x. Note that our pdf is represented by discrete samples.

2. If you want N samples, select N samples from [0, 1]. Mark them on the y-axis.

3 Project N samples onto cdf to select samples along x-axis. The resulting set of samples (including replications) is distributed according to the posterior.

Density of States and Phase Transitions.





If you do not iterate long enough, it looks like you have converged But no!

The water phase is a highly probable region with small volume sitting on top of the steam phase.

Phase Transitions <=> AHAs !!

Inspiration

Interestingly, one can often look at the behavior of the algorithm and identify which specific concepts are being learned at each phase transition.

Simulated Annealing

Turn data on slowly.

Priors are smooth and easy to navigate.

Likelihoods can be rough.

$$P(M|D,I;\lambda) \propto \frac{P(M|I)\,P(D|M,I)^{\lambda}}{\cancel{P(D|I)}}$$

let λ go from 0 to 1

This "turns on" the data term.

λ is the "temperature"

$\lambda = 0$ cold

$\lambda = 1$ hot

$$\langle \log \mathcal{L} \rangle_\lambda = \int \log \mathcal{L}\, dP_\lambda = \frac{\int \mathcal{L}^\lambda \log \mathcal{L}\, d\xi}{\int \mathcal{L}^\lambda d\xi} = \frac{d}{d\lambda}\left[\log\left(\int \mathcal{L}^\lambda d\xi\right)\right.$$

If we integrate from $\lambda = 0$ to 1

$$\int_0^1 \langle \log \mathcal{L} \rangle_\lambda\, d\lambda = \log\left(\int \mathcal{L}^\lambda d\xi\right)\Big|_0^1 = \log\left(\int \mathcal{L}\, d\xi\right) - \log\left(\int d\xi\right)$$

$$= \log \frac{\left(\int \mathcal{L}\, d\xi\right)}{\left(\int d\xi\right)}$$

This methodology is another way to find Z.

$$= \log Z \quad !!!$$ the evidence

Thermodynamic Integration

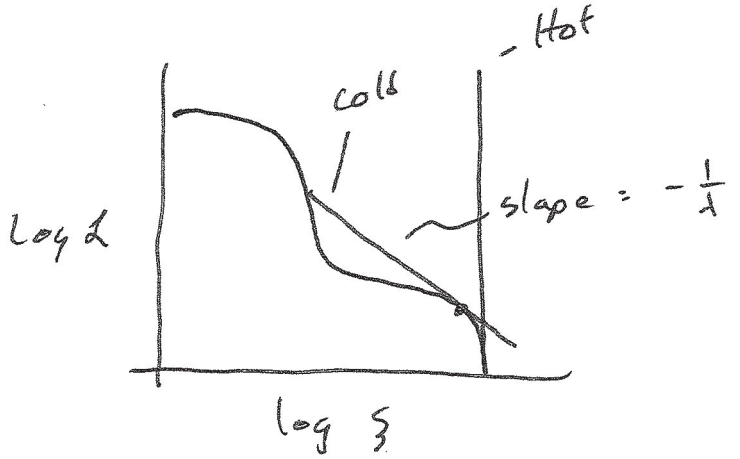
Has problems at phase transitions since it tracks $\frac{d \log \mathcal{L}}{d \log \xi} = -\infty$

Simulated Annealing Tracks Slope

$$dP_\lambda \propto \mathcal{L}^\lambda \, d\xi$$

The greatest probability mass is most likely to be found under the maximum at $\mathcal{L}^\lambda \xi$.

This maximum occurs at $\frac{d \log \mathcal{L}}{d \log \xi} = -\frac{1}{\lambda}$



As one cools from $\lambda = 0$ (infinite slope) to $\lambda = 1$ (45°), simulated annealing will eventually require a jump to get over the phase transition.

The algorithm get stuck here.