51

Nested Sampling

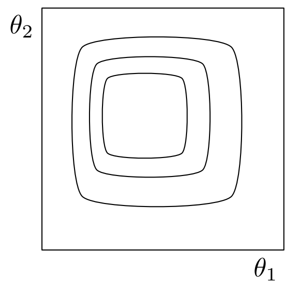


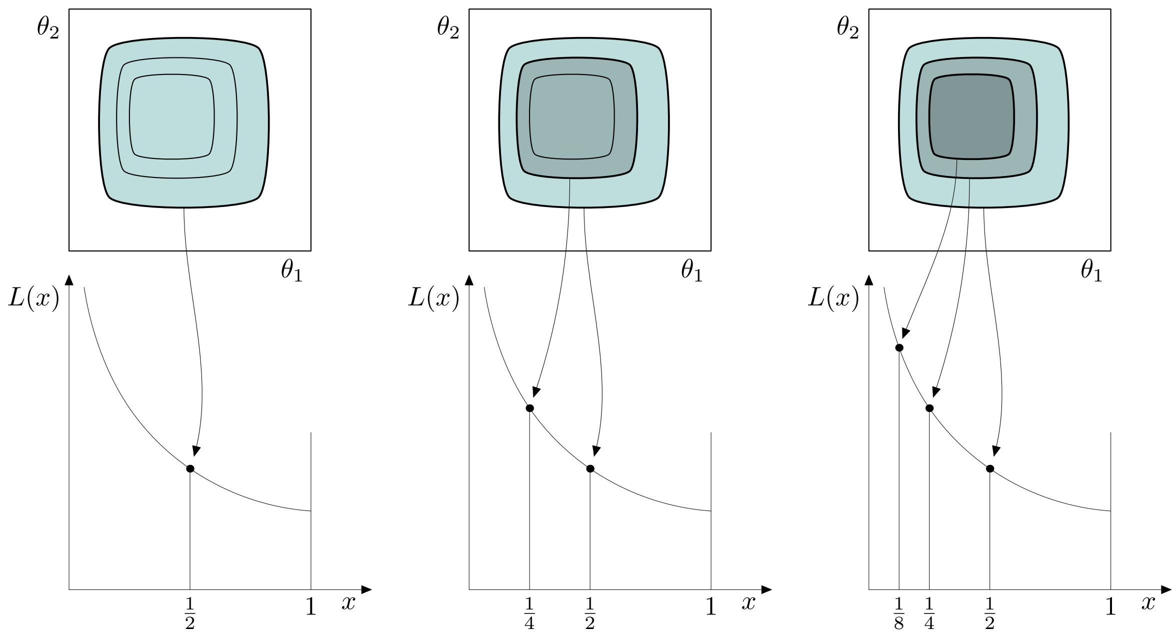
Figure 51.1. Contour plot of a likelihood function $\mathcal{L}(\boldsymbol{\theta})$.

Figures by David MacKay.

John Skilling's way of thinking about the integral $Z = \int d^K\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Let $x(L)$ be the prior mass enclosed within the contour $\mathcal{L}(\boldsymbol{\theta}) = L$, and $L(x)$ be the contour value such that the volume enclosed is x.

$$Z = \int dx\ L(x).$$



An example of $L(x)$

Let $\boldsymbol{\theta}$ be a collection of G unknown binary variables $\theta_g \in \{0, 1\}$, and let our data be a list of G independent noisy observations of them – one observation each. So the likelihood function will have the form

$$\mathcal{L}(\boldsymbol{\theta}) \propto \exp\left(\sum_{g=1}^{G} b_g\theta_g\right), \tag{51.1}$$

where the b_g is the bias for θ_g towards or away from 1 (if b_g is positive or negative respectively). If all the noisy observations have the same noise level then the magnitudes of the b_g will be the same for all g.

Clearly the posterior distribution is separable. This is a very simple inference problem, but it epitomizes some of the issues arising in more realistic problems.

To connect to my chapter on sex, we can note that if all the b_g happen to be $+b$ then the log-likelihood is proportional to the fitness $F \equiv \sum_{g=1}^{G} \theta_g$ that I assumed there.

So, what does $L(x)$ look like? The volume fraction $x = 1/2^G$, is associated with the unique maximum likelihood state. Moving away from that corner of the hypercube, the log-likelihood increases in proportion to the Hamming distance from that corner, and the number of states at Hamming distance d is $\binom{G}{d}$. Or, in terms of the fitness F, which is $G - d$, the number of states is $\binom{G}{F}$.

Figure 51.2 shows $L(x)$ from various points of view, for the case where the number of independent variables is $G = 30$. Of these graphs, 51.2(b) is perhaps the easiest to relate to: flipping the two axes round, this graph is almost exactly the cumulative normal distribution function, shifted and scaled.

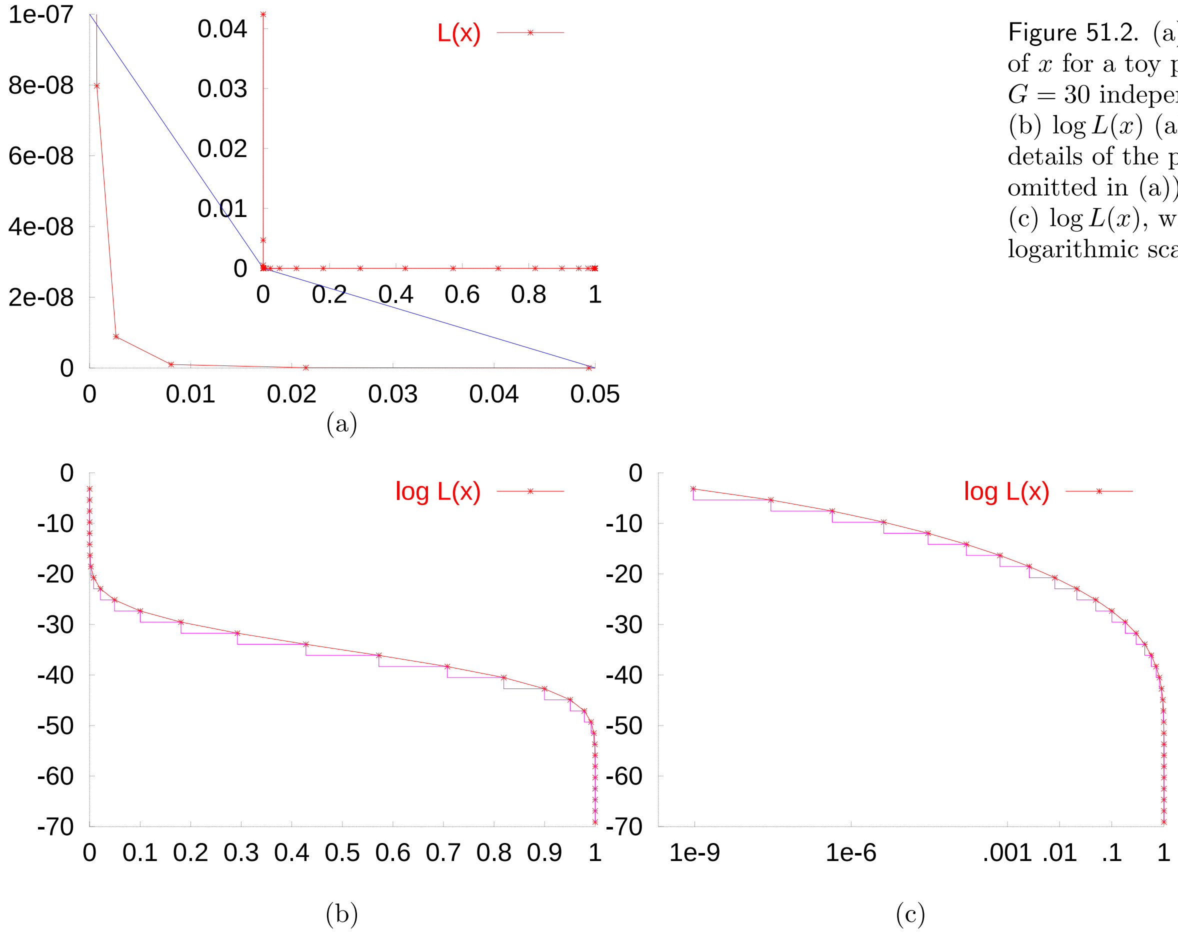


Figure 51.2. (a) $L(x)$ as a function of x for a toy problem with $G = 30$ independent variables. (b) $\log L(x)$ (also showing the details of the plateaus of L, omitted in (a)). (c) $\log L(x)$, with x shown on a logarithmic scale.

Notice that $L(x)$ is a very sharply increasing function as $x \to 0$. $\log L(x)$ is locally a roughly linear function of $\log x$ (if we neglect the plateaus of L, so locally we can think of L as behaving like a power law $L(x) \simeq x^{-p}$, for some p. For this example, a crude but useful description of the situation is that halving the volume x increases $\log L(x)$ by a constant of order 1.

Nested sampling

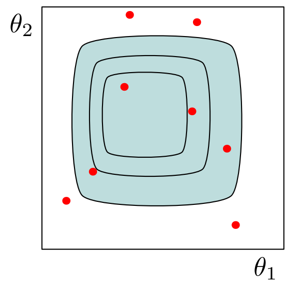
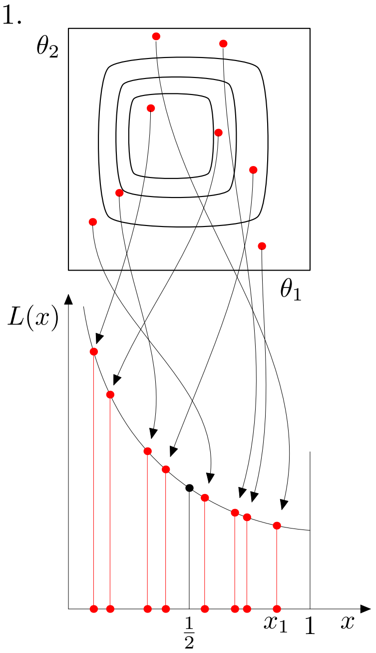


Figure 51.3. $N = 8$ points drawn uniformly from the prior.

We start by drawing N points uniformly from the prior. Let $N = 8$, say. Roughly half of the points fall inside the shaded region corresponding to the contour with $x = 1/2$. Roughly one quarter of them are inside the contour associated with $x = 1/4$. Roughly one eighth of them are inside the contour associated with $x = 1/8$.

We can associate each point $\boldsymbol{\theta}_i$ with an x-value, namely the volume that would be enclosed by the contour $\mathcal{L}(\boldsymbol{\theta}_i)$. Since the points are uniformly distributed under the prior, the N x-values are uniformly distributed between 0 and 1.



Let x_1 be the largest x-value. The typical value of x_1 is something like $N/(N+1)$ or $e^{-1/N}$. (The former is its arithmetic expected value, the latter its geometric mean.) We introduce a contour associated with this point.

Nested sampling now draws a new point, uniformly distributed in the region satisfying $\mathcal{L} \geq L(x_1)$. (We assume that this operation can be done, perhaps by a Markov chain method, just as annealing methods assume that a point can be drawn from the distribution $\propto \mathcal{L}^{\beta}$.) The new point is shown by the big purple dot.

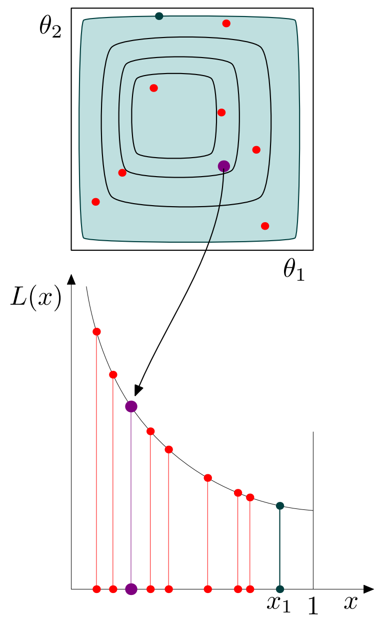


Figure 51.4. Replace the point at x_1 by a new point uniformly distributed between 0 and x_1.

We insert this new point and find among the N live points the biggest x-value, x_2. (Remember there's a chance of roughly $1/N$ that the new point might have landed between the second-biggest x and x_1.)

These x-values are uniformly distributed between 0 and x_1.

We don't know the values of the volumes x_i, but we do know their order, since we know the values of $L(x_i) = \mathcal{L}(\boldsymbol{\theta}_i)$.

At each iteration, the volume shrinks roughly by a factor of $e^{-1/N}$.

▶ 51.1 What is a typical sequence $\{x_i\}$ like?

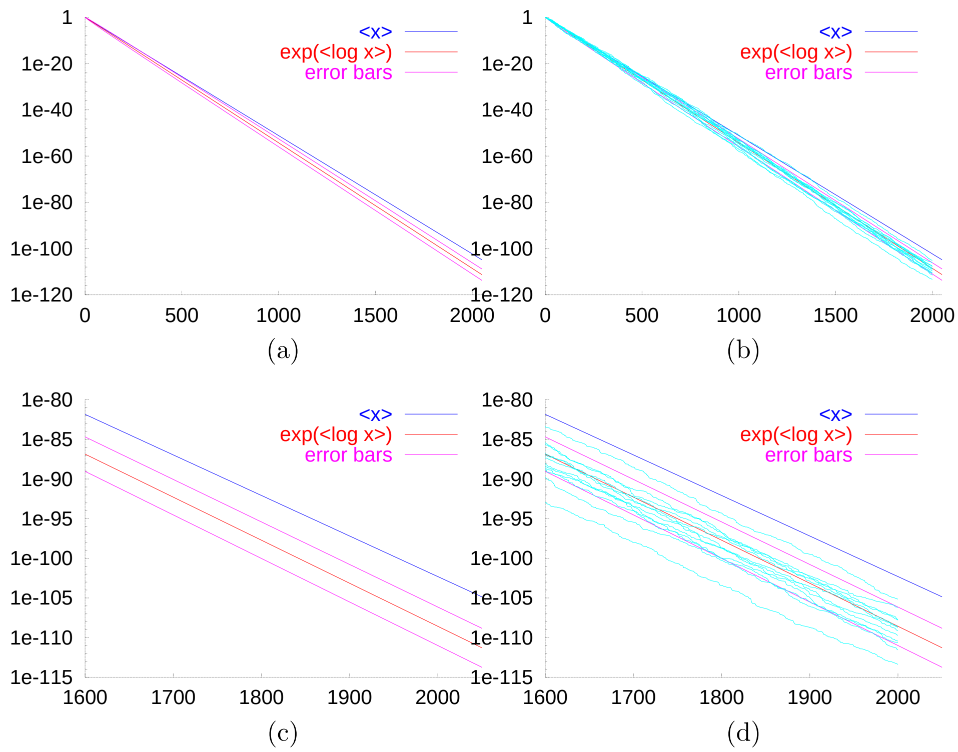


Figure 51.5. (a) The arithmetic and geometric means of x_i for the case $N = 8$; also, error bars on the geometric mean,

$$\exp(-i/N \pm \sqrt{i}/N).$$

(b) A dozen samples from the distribution of $\{x_i\}$, for runs of duration 2000 steps.
(c,d) Detail of (a,b).