Teething problems! — Monte Carlo evaluation of Normalizing Constants

David J.C. MacKay
Cavendish Laboratory
mackay@mrao.cam.ac.uk

November 29, 1994— *Draft 1.3*

Abstract

This is a case study of the use of Monte Carlo methods to evaluate normalizing constants. I describe the trials and tribulations of importance sampling and of variational free energy approaches. The results are for a small model with just one latent variable.

More efficient evaluation of the evidence using importance sampling

If we create a sampling distribution $Q_j(\mathbf{x})$ that is similar to the posterior distribution $P(\mathbf{x}|\mathbf{F}_j)$ then the evidence integral can be approximated in terms of $\{\mathbf{x}^{(r)}\}_{r=1}^{R}$, which are random samples from $Q(\mathbf{x})$.:

$$\begin{aligned} L_j(\mathbf{w}) &= \log \int d^H\mathbf{x}\ \exp(G_j(\mathbf{x};\mathbf{w}))P(\mathbf{x}) \\ &\simeq \log \left[\frac{1}{R} \sum_r \exp(G_j(\mathbf{x};\mathbf{w})) \frac{P(\mathbf{x})}{Q(\mathbf{x})} \right] \end{aligned}$$

Later, I use this expression to evaluate accurately the evidence for a model that has been adapted by the simple Monte Carlo method above. The sampling distribution $Q_j(\mathbf{x})$ is set to a Gaussian with mean $\bar{\mathbf{x}}_j$ and diagonal covariance matrix Σ_j obtained from statistics returned by the simple algorithm.

The simple Monte Carlo algorithm gave the results illustrated in figure 4, as H and R were varied. The graphs show the evidence as a function of R. Notice that for R greater than 10 or so, the evidence value settles down, and increasing R makes negligible difference.

In the case of data TOY 1, as H is increased beyond 1, the evidence does not become either substantially larger or substantially smaller, even when the hidden vector has a dimensionality bigger than the dimensionality of the output space. This means that the model is finding a density of effective dimensionality about 1. There is apparently no overfitting problem.

	TOY 1												TOY 2
	i	1	2	3	4	5	i	1	2	3	4	5	
	j						j						
Data	1	5	2	0	0	0	1	5	2	0	0	1	
	2	2	3	1	0	0	2	2	3	1	0	0	
	3	0	5	3	0	0	3	0	5	3	0	0	
	4	0	1	2	4	1	4	0	1	2	4	1	
	5	0	0	1	3	4	5	0	0	1	3	4	
	6	1	1	1	1	1	6	2	0	0	2	3	

Table 1: **Parameters of models for the TOY problems**

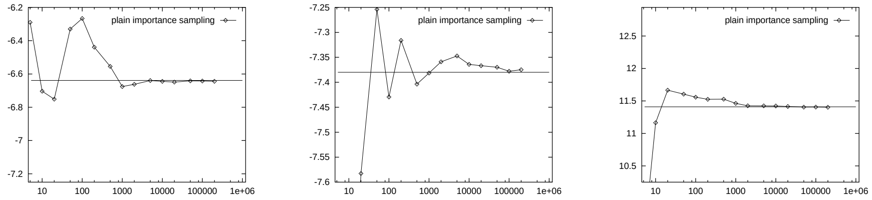

Figure 1: **Toy example. Individual evidences (cols 1 and 2), and sum for all 6 data (col 3).** Log evidence (y axis) is shown as a function of R (number of Monte Carlo samples, x axis). Top line = plain importance sampling results.

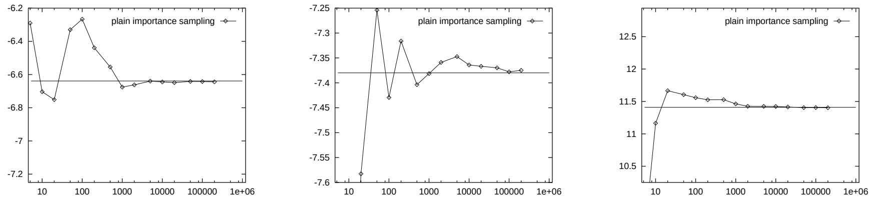

Figure 2: **Toy example. CAUCHY importance sampler. Individual evidences (cols 1 and 2), and sum for all 6 data (col 3).**

Log evidence (y axis) is shown as a function of R (number of Monte Carlo samples, x axis). Top line = plain importance sampling results.

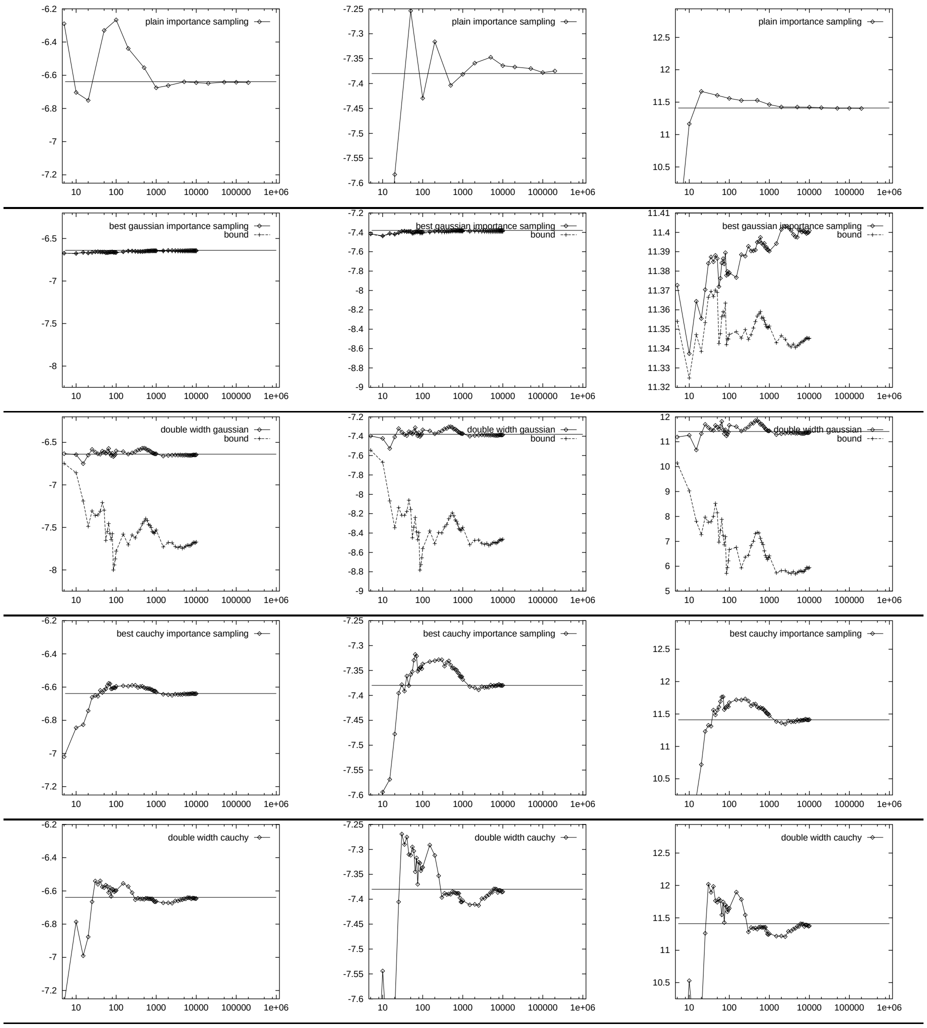

Figure 3: **Toy example. Various samplers, well optimized. Individual evidences (cols 1 and 2), and sum for all 6 data (col 3).**

Top line: plain importance sampling results. 2: Optimized gaussian. 3: Gaussian of double width. 4: Cauchy. 5: Cauchy of double width.

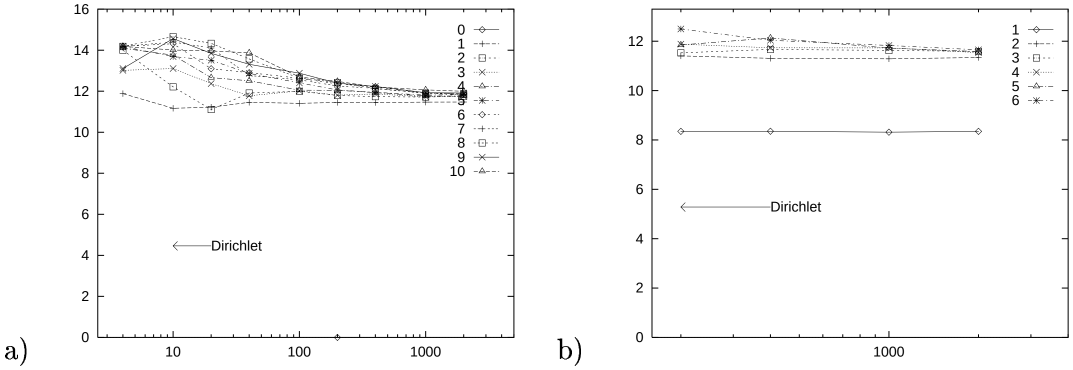

Figure 4: **Toy examples. Estimated evidence.**
Log evidence (y axis) is shown as a function of R (number of Monte Carlo samples, x axis), for models with different numbers of hidden components (H between 0 to 7).
The evidence for the optimized Dirichlet model is also marked. All values are log evidences relative to the null model $\mathcal{H}_0$.
a) Toy example number 1. b) Toy example number 2.

In the case of data TOY 2, the results are similar, except that the model with a two-dimensional componential representation is significantly more probable than the one-dimensional density network.

One way to understand what a model is doing is to look at its parameters (at least for small H). Table 1 shows the parameters for the nets with $H = 1$ and $H = 2$, ordered from $i = 1$ to 5 vertically (c.f. horizontal in the data table earlier). Notice that the weights from the inputs in the TOY 1 cases capture the one dimension apparent to the human eye. When there are two inputs, the weight vectors for those inputs are not orthogonal; they are virtually identical (except for a change of sign). This similarity of the vectors of weights from the two inputs produces a low effective dimensionality in the output space.

When it is adapted to the TOY 2 data set, the parameters of the density network with two hidden components are very different. The two vectors over i are here virtually orthogonal, so that a fully two-dimensional distribution is produced in the output space.

Amino acid probabilities in aligned protein families

Figure 5 shows the estimated evidence, for $J = 60$ examples, each with a count of $F_j \simeq 177$. Clearly many Monte Carlo samples are needed for a convergent estimate of the evidence.

The evidence for the Dirichlet model is also displayed. According to these results, a componential model with 13 components is more probable than the Dirichlet model.

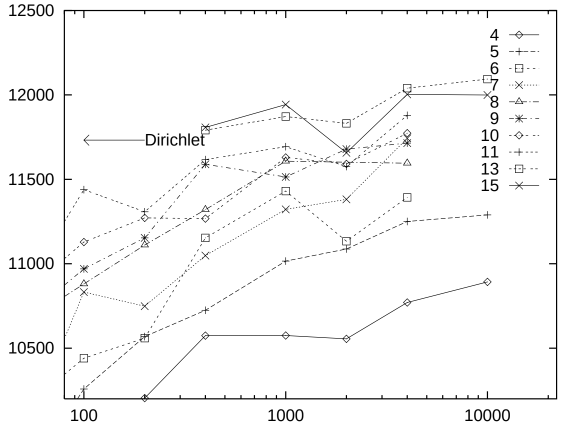

Figure 5: **Amino acid modelling.**
Estimated evidence, as a function of R (number of Monte Carlo samples, x axis), for models with different numbers of hidden components ($H = 3$ to 15).
The evidence for the optimized Dirichlet model is also marked. The evidence for other traditional Dirichlet models can also be reported: $\log P(D|\mathbf{u} = (1, 1, \ldots, 1)) = 10894.5$; $\log P(D|\mathbf{u} = (.05, .05, \ldots, .05)) = 11356.7$.
All values are log evidences relative to the null model $\mathcal{H}_0$.