BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Free energy biased sampling and mixture modelling

Peter J. Green
University of Bristol, UK
P.J.Green@bristol.ac.uk

Summary

Free energy biased sampling is another advanced Monte Carlo technique borrowed from statistical physics; this is a discussion of its potential use in Bayesian computation via Markov chain Monte Carlo and Sequential Monte Carlo, and application in particular to posterior sampling in mixture models.

Keywords and Phrases: Adaptive biasing; Markov chain Monte Carlo; Sequential Monte Carlo.

1. INTRODUCTION

It is a pleasure to present this discussion of Chopin and Jacob (2010), which has been influenced by reading in parallel the recent paper by Chopin, Lelièvre and Stoltz (2010). This gives more detail on free-energy biasing, and applies it in the context of Markov chain Monte Carlo, and is also illustrated by applications to mixture modelling.

My discussion focuses on the general ideas of free energy biased sampling (FEBS), including estimation of the free energy, and on comparisons of the different impact of FEBS on sequential Monte Carlo and Markov chain Monte Carlo. Turning to the mixtures application, I will give my own views on the label switching issue. Finally, I will comment on the prospects for wider use of FEBS in Monte Carlo methods for Bayesian computation.

1.1. *Further de-mystifying free energy biased sampling*

Statistical science has repeatedly borrowed ideas about Monte Carlo sampling from statistical physics over the years, and this paper is about one more example of this tradition. I want to go a little further than Chopin and Jacob in presenting the basic idea in a language and notation that should be familiar to statisticians.

Peter Green is Professor of Statistics at the University of Bristol. This work was conducted with the partial support of an EPSRC Science and Innovation award for the SuSTaIn programme.

Given a target distribution (e.g. Bayesian posterior) $\pi(\theta)$, $\theta \in \mathcal{R}^d$, we write

$$\pi(\theta) = \pi(\theta_1) \times \pi(\theta_2|\theta_1) \text{ for } \theta_1 \in \mathcal{R}, \theta_2 \in \mathcal{R}^{d-1}$$

and then replace

$$\pi(\theta) = \pi(\theta_1) \times \pi(\theta_2|\theta_1) \text{ by } \widetilde{\pi}(\theta) = \widetilde{\pi}(\theta_1) \times \pi(\theta_2|\theta_1)$$

where $\widetilde{\pi}(\theta_1)$ is 'broader', usually Uniform on $[x_{\min}, x_{\max}]$, but here with added tails.

Finally, we sample from $\widetilde{\pi}(\theta)$, and use importance sampling with weights $\pi(\theta_1)/\widetilde{\pi}(\theta_1)$ to estimate probabilities and expectations under the target distribution.

Of course, it is not really that simple. In fact, first we transform $\theta \leftrightarrow (\xi, \eta)$, then we do as above, with θ_1 replaced by ξ, called the *reaction coordinate*. Thus we will actually sample from

$$\widetilde{\pi}(\theta) = \widetilde{\pi}(\xi(\theta)) \times \pi(\eta(\theta)|\xi(\theta)) \times \left| \frac{\partial(\xi, \eta)}{\partial \theta} \right|$$

Furthermore, we will have to estimate $\pi(\xi)$ online as we do so – this is needed to evaluate $\pi(\eta|\xi)$.

Why should this be a good idea? The point is that with a suitable choice of ξ, $\widetilde{\pi}(\theta)$ may be easier to sample from than $\pi(\theta)$, whether directly, or by MCMC or SMC.

The free energy associated with the distribution $\pi(\theta)$ using reaction coordinate $\xi(\theta)$ is the function $A(x) = -\log \pi(\xi)$, evaluated at $\xi(\theta) = x$. Estimating $A(x)$ (up to an additive constant) is equivalent to estimating $\pi(\xi)$ (up to a multiplicative factor).

1.2. *There's no such thing as free energy*

In the physical chemistry community, there is some institutional effort to eliminate the adjective "free" in "free energy" as it is regarded as redundant, see IUPAC (1990), but this effort seems to be only partially successful to date. But perhaps statistical scientists adopting the method will further preserve the name!

1.3. *Visualising FEBS in simple cases*

For motivation into whether and how FEBS might be effective, let us look at some 2-component bivariate normal mixtures for $\pi(\theta)$, as did Nicolas Chopin in his oral presentation – here $A(x)$ is of course known. It is clear from Figure 1 that while free energy biasing achieves its goal of making one component of the target distribution have a uniform distribution, where before it was bimodal, it is not necessarily true that the resulting modified joint target $\widetilde{\pi}$ is unimodal – that would require appropriate alignment of the tails of the modes in the target. More careful choice of ξ does not help in this case (Figure 2).

These are very simple examples – consider how much more complicated the picture can be in a general situation of high dimensionality, with several modes of different shapes and sizes, and without easy visualisation to aid the choice of reaction coordinate!

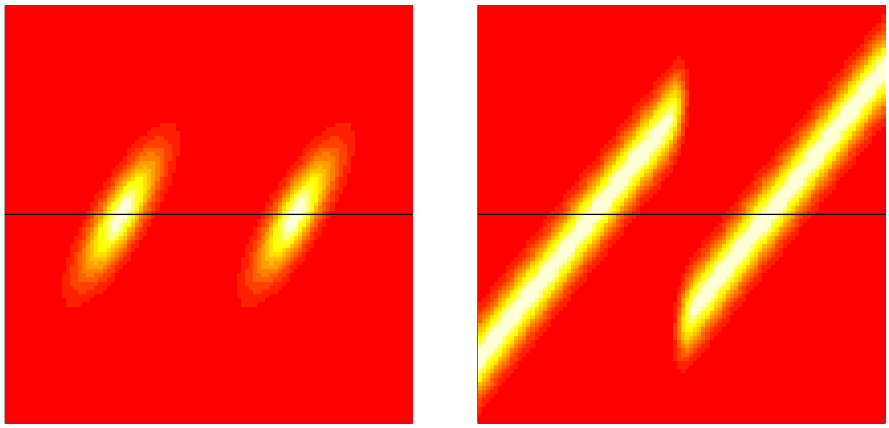

Figure 1: *A 2-component bivariate normal mixture, before and after free energy biasing.*

2. LEARNING THE FREE ENERGY TO EMPLOY FEBS

To use FEBS, we need to evaluate $A(x)$; but $\exp(-A(x)) = \pi(\xi)$ is a marginal of a complex target distribution $\pi(\theta)$, so this evaluation implicitly involves integration, which is accomplished by online updating using the empirical history of the simulation. Both Chopin and Jacob (2010) and Chopin, Lelièvre and Stoltz (2010) consider the adaptive biasing potential (ABP), and adaptive biasing force (ABF) methods. Methods similar to ABP are somewhat familiar to statisticians already, even going back to Geyer and Thompson (1992).

These standard approaches involves discretisation of the $x = \xi(\theta)$ scale – is it worth doing something smoother?

In ABP, we use

$$\exp(-\widehat{A}_1(x)) = \frac{\sum_{n=1}^N w_n I\{\xi(\theta_n) \in [x_i, x_{i+1}]\}}{\sum_{n=1}^N w_n} I[x \in [x_i, x_{i+1}]],$$

which can be viewed as using a density estimate based on a histogram, and this might be replaced by

$$\exp(-\widehat{A}_1(x)) = \frac{\sum_{n=1}^N w_n h^{-1} K((\xi(\theta_n) - x)/h)}{\sum_{n=1}^N w_n}.$$

for some kernel function $K(\cdot)$.

Similarly, in ABF,

$$\widehat{A}'_2(x) = \frac{\sum_{n=1}^N w_n I\{\xi(\theta_n) \in [x_i, x_{i+1}]\} f(\theta_n)}{\sum_{n=1}^N w_n}, \quad \text{for } x \in [x_i, x_{i+1}]$$

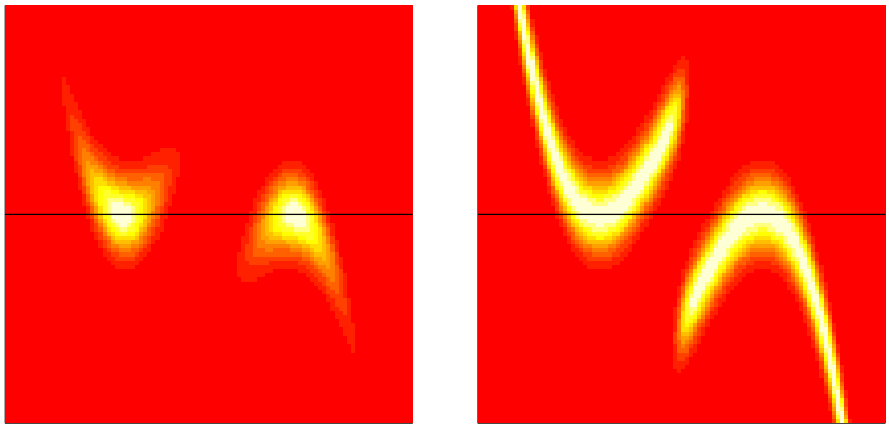

Figure 2: *The same 2-component bivariate normal mixture as in Figure 1, before and after free energy biasing, but with a different $x \leftrightarrow (\xi, \eta)$ transformation.*

might be replaced by

$$\widehat{A}_2'(x) = \frac{\sum_{n=1}^N w_n h^{-1} K((\xi(\theta_n) - x)/h) f(\theta_n)}{\sum_{n=1}^N w_n},$$

and followed perhaps by a more sophisticated partial integration method than $\widehat{A}_2(x) = \sum_{j:x_j \le x} \widehat{A}_2'(x_j)(x_{j+1} - x_j)$. Increments in $\widehat{A}(x)$, evaluated at observed $\xi(\theta_{t,n})$, are needed in the free energy SMC algorithm, and $\widehat{A}(x)$ itself in the final 'debiasing' (importance sampling) step – and these can be fast computations with care in implementation of the kernel methods.

Discretisation using a histogram approximation will be slightly cheaper, but perhaps involves a more critical choice of bin-width than that of the bandwidth h, and smoothness is cited as desirable (and is generally true of the theoretical $A(x)$).

3. ADAPTIVE BIASING IN MCMC AND SMC

In Chopin, Lelièvre and Stoltz (2010), the authors modify the target distribution every MCMC sweep, using the whole cumulative history, each single $\xi(\theta_t)$ being appropriately weighted. However, here in Chopin and Jacob (2010), the authors use the current SMC particle cloud to update the target distribution at each full iteration. From these studies, the authors conclude that, at least in these gaussian mixture applications, satisfactory posterior simulation results are obtained at lower cost with free energy SMC than free energy MCMC.

Is this saying anything deeper than that FEBS has only limited power to make a target distribution easier to simulate, and that SMC is less sensitive than MCMC to the (reduced but still substantial) multi-modality we are left with?

4. MIXTURE POSTERIORS, USING FEBS

Bayesian mixture analysis is a good choice of a test application for FEBS for this audience – mixtures are well-known to give difficult posterior surfaces.

In both papers, the same alternatives for reaction coordinate $\xi(\theta)$ are considered: (a) a component mean, (b) a component weight, (c) the scale hyperparameter of the prior on component precisions, and (d) the log unnormalised posterior density. Choice (d) is taken in both papers. This process is far from automatic – listing these choices demands insight about model and resulting posterior, and selection from the list is partly empirical.

There are not many hints on how to choose a good reaction coordinate in general. Several principles are suggested: (i) "a direction in which the target density is multimodal" (Chopin, Lelièvre and Stoltz, 2010) (ii) "conditional on $\xi(\theta) = x$, multimodality is much less severe, at least for some x" (Chopin and Jacob, 2010) (cf. tempering) or (iii) "time-scale for dynamics on ξ larger" [than for η]: "ξ is a slow variable" (Lelièvre, Rousset and Stoltz, 2009) We need methods for evaluating, steering or even automating choice of $\xi(\theta)$.

5. LABEL SWITCHING

Rather a lot of ink has been spilt on the vexed question of label-switching, and what to do about it.

Label-switching refers to (a) the fact that in a standard mixture model, $y \sim \sum_{j=1}^{k} w_j f(\cdot;\theta_j)$, the model is invariant to permutation of the component labels j, and to (b) the consequences for sample-based computation.

In truth, in such models, the model parameter is not a vector, but a (marked) point process – in the absence of prior information distinguishing the components, we are not entitled to make inference about individual components, e.g. $P\{\theta_2 > 4.65|\text{data}\}$, but only about components simultaneously, e.g. $E(\#\{j : \theta_j > 4.65\}|\text{data})$. This fact provides a constraint on the information that it is legitimate to extract from the posterior simulation. Taking this perspective, it is not true that there are $k!$ modes (but, as observed by Jasra at al (2005), there may still be more than 1).

That observation does not make the issue of mixing unimportant, since most samplers will represent the point process using vectors, and updates may be sensitive to the current ordering.

However, in my view, apparently mixing successfully between the $k!$ equivalent representations is neither necessary nor sufficient for confidence that the sampler is reliable. It is a surrogate diagnostic that we do not fully know how to use.

To make inference about specific numbered components, there has (unusually) to be prior information distinguishing them, then they are intrinsically labelled. This raises one or two interesting new issues in mixing (is *this* cluster of data or *that* one the cluster fitted by component labelled 2?), but the permutation invariance issue no longer applies.

6. CONCLUSIONS

6.1. *Prospects: using FEBS in general?*

The authors are pretty cautious, in not making extravagant claims for FEBS beyond their current studies. I think this is wise.

Firstly, for widespread and routine application, we would need methods for evaluating, steering or even automating choice of $\xi(\theta)$, and current knowledge seems

to be thin on this. One of the major differences between Monte Carlo methodology in practice in statistical science compared to statistical physics, is that the latter tends to concentrate on a relatively limited range of standard but hard problems, and so it is worth spending effort on extensive tuning, whereas statisticians face a different posterior surface every time they perform an analysis. One possibility for progress might use ideas of projection pursuit based on pilot runs? However, as we see even in the mixture example it may be sections rather than projections of the target distribution that suggest promising reaction coordinates.

Secondly, recalling our simple examples earlier in this discussion, we have to question how well can FEBS perform even with optimal choice of $\xi(\theta)$, and perfect learning of $A(x)$.

Where FEBS is effective, given the efficiency loss in importance sampling, is there possibly an advantage in not trying too hard to optimise estimates of $A(x)$, but rather to compromise between efficiency and mode-bridging? (There is some discussion of efficieny issues in Chopin, Lelièvre and Stoltz.)

The principle of FEBS applies also to higher-dimensional $\xi(\theta)$; what are the advantages and disadvantages? Presumably, such ξ are harder to choose, and it is more difficult to get stable density estimates, but methods can be more powerful at mode-bridging, while the importance sampling less efficient? One route to selecting higher-dimensional reaction coordinates might be to proceed sequentially, choosing reaction coordinates $\xi_1(\theta), \xi_2(\theta), \ldots$ in turn.

6.2. *Summary*

It seems difficult at present to understand the potential future impact of FEBS in computational Bayesian analysis. Will this be another apparently good idea from statistical physics (Swendsen-Wang, CBMC,...) that does not fulfil optimistic expectations? Are there more ideas in modern molecular dynamics computation that we can use?

However, the empirical results here are impressive. They do support the conclusions that for these examples, FEBS helps, and that SMC+FEBS beats MCMC+FEBS. Many thanks to Nicolas and Pierre for introducing us to this topic!

REFERENCES

Chopin, N. A. and Jacob, P. (2010). Free energy Sequential Monte Carlo, application to mixture modelling. *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press.

Chopin, N. A., Lelièvre, T. and Stoltz, G. (2010). Free energy methods for efficient exploration of mixture posterior densities. ArXiv: 1003.0428.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. B* **5**4, 657–699.

IUPAC, International Union of Pure and Applied Chemistry, Commission on Atmospheric Chemistry (1990). "Glossary of Atmospheric Chemistry Terms (Recommendations 1990)". Pure Appl. Chem. 62: 2167ñ-2219. doi:10.1351/pac199062112167

Jasra, A. Holmes, C. and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statist. Science* **2**0, 50–67.

Lelièvre, T., Rousset, M. and Stoltz, G. (2009). Long-time convergence of an adaptive biasing force method. *Nonlinearity*, **21**, 1155-81.