6th Valencia meeting

Exact sampling for Bayesian inference: towards general purpose algorithms,
by Peter Green and Duncan Murdoch

Authors' reply to Discussion

We are grateful to each of the discussants for their thorough reading of both the present paper and our Scandinavian Journal article, for their generous comments, and for their insightful and stimulating discussions. They are evidently as excited as we are by the apparent opportunities provided by coupling from the past, and share our mixed feelings about whether some tremendous benefits for Bayesian computation are waiting to be uncovered, or whether this will all turn out to be just an intriguing sideline.

Four years gestation should be enough for real applications to emerge, so one measure of how it turns out will be the number of papers at the *next* Valencia that exploit exact sampling. Whether this is zero or a lot, we are pleased that we took the somewhat risky decision to present these ideas at the conference.

Most of the discussants allude to the cumbersome algebra and analysis needed to set up coupling from the past in all but the simplest of models. This is a central issue. There is a very real danger that in pursuing the benefits of "guaranteed convergence", implementors will be throwing away the very freedoms that MCMC bought for them in the first place. Any substantive data analysis that is limited by what can be done exactly turns the clock back to when tractability was a dominant issue in Bayesian analysis.

For these reasons, we see partial automation of CFTP implementation as absolutely crucial to the practical application of these ideas. The methods introduced in the Scandinavian Journal article—the multigamma and rejection couplers, and their partitioned versions—are very special. Their existence does little more than demonstrate their existence!

However, the methods introduced in Sections 3 and 4 of the present paper—bisection coupling and dominated coupling—claim to be much more generic. They are constructed from an analysis of the target distribution, which in Bayesian computation is usually explicit if complicated, rather than the transition mechanism of the chosen Markov chain, which is usually extremely awkward. One price to be paid for their generic character is probable increased coalescence time, as in several places in the derivation of these methods, exact expressions are replaced by bounds, and sets of update states enlarged for convenience of representation. Such conservatism does not affect the exactness of the ultimate draw from the target, but may delay coalescence—perhaps, fatally, to time infinity—and complicate analysis of performance.

Perhaps one value of the earlier methods is to shed some light on the roles of regeneration and renewal in coupling. Professor Tierney exploits this in proposing some very nice modifications of the multigamma method. It is potentially useful to be able to relax from the requirement to organise housekeeping of the random number streams to ensure that $X_{t+1} = \phi(X_t, U_{t+1})$. However, we repeat that we do not see multigamma coupling as a general method, although it may play a role as a building block.

We thank Professor Everson for emphasising that the update function/stochastic recursive sequence representation $X_{t+1} = \phi(X_t, U_{t+1})$ allows much flexibility. We exploited somewhat similar formulations in our implementations of rejection couplers, which are set out in the Scandinavian Journal paper. Such flexibility can be taken much further. Taking a finite state space example for clarity, suppose that $\chi = \{1, 2, \ldots, S\}$, and that the $U_t = (U_t^{(1)}, \ldots, U_t^{(S)})$ are S-dimensional. Then it is legitimate to take $X_{t+1} = \phi(X_t, U_{t+1}) = \tilde{\phi}(s, U_{t+1}^{(s)})$ when $X_t = s$, for an appropriate $\tilde{\phi}$, so that we actually use a *different* random number for each current state X_t! We have nevertheless ensured what we must, namely that the update from X_t at time t is the same on every occasion that this update is considered.

For us, the most substantial reason to doubt the eventual widespread use of CFTP for truly exact sampling in Bayesian analysis is the non-modular nature of all existing CFTP prescriptions. Ordinary MCMC methods are all inherently modular, graphical modelling representations of statistical models are modular, and the practice of model-building for data analysis is modular. We completely agree with Professor Tierney in his analysis of this issue, based on a "divide-and-conquer" interpretation of the usual MCMC recipes. The only place in our methods where we can work in a modular fashion is in constructing the approximate cell bounds described in Section 3.3. We do have some prototype code that can create tables of cell bounds automatically from a BUGS specification of the model, but this is untested in real problems, and the whole strategy of partition by rectangular cells is of course limited by the curse of dimensionality.

The role played by monotonicity in CFTP deserves further study. We have stressed the non-monotone nature of our methods, meaning that we need to use structure functions $\phi(x, u)$ that are not monotonic in x. (Of course, other aspects of monotonicity are being used in establishing the bounds we need; for example, acceptance probability is monotonic in the target density value for the current state, an essential ingredient in Section 3.) Professors Robert and Titterington give a nice example, demonstrating that we were wrong to rule out the possibility of monotonicity with respect to state in continuous-parameter problems, but we are not sure how often such tricks will be available.

In this connection, we should correct an impression given in Section 1.3. Häggström and Nelander (1997b) do consider non-monotonic couplers for Markov random fields. They include an example where use of a monotonic coupler is vastly superior to a non-monotonic one in terms of coalescence time, and further research on such comparisons would be interesting. But, as with Robert and Titterington's example, it is not clear how much of the gain can be attributed to monotonicity (allowing the tracing of fewer paths, 2 instead of $n + 1$ in the mixture example), as the coupling methods are entirely different (one giving intrinsically faster coalescence).

Professor Damien is right to note that the key to creating coalescing paths in a continuous parameter problem is to look for ways of inducing discreteness; this was indeed our motivation. This idea has appeared previously in Nummelin's splitting technique, translated into practical MCMC methodology by Mykland, Tierney and Yu (1995).

We have already begun thinking about exact sampling for variable dimension problems such as mixture analysis, as suggested by Professor Damien. (In fact, we do not necessarily buy his motivation for doing this; we feel that the breakdown of existing diagnostic methods comes not in going from finite to infinite or variable dimensions, but in going from small to moderate dimensionality. In all cases, you will apply diagnostics to some carefully-chosen scalar functionals; all that changes with dimension is the capacity of a few such functionals to capture the behaviour of the entire chain.) It is not too far-fetched to surmise that it might be possible to couple MCMC methods for mixtures. These problems do have some of the character of point-process models—here the points would be marked—and several such models have already been addressed by CFTP. Robert and Titterington's implementation of CFTP for a particular one-parameter mixture problem is a start, although we are not sure how far this example can be extended.

On the question of how to increase the trial values of M in coupling from the past, raised by Professors Robert and Titterington: of course, any increasing sequence is valid, and the matter is simply a question of computing time. Propp and Wilson give an argument to support a geometric increase. Besides that argument, there is the heuristic reasoning that computing a failure is usually much more expensive than computing a success, because many more paths need to be followed, so you want to reduce the number of failures. Since the distribution of T often has a heavy tail, you need a rapidly increasing sequence of trial values.

For the small Beta problem discussed in Section 3.4, we have given the histogram of the successful M values in 1000 independent replications of the random walk Metropolis method. This says something about the distribution of T, since $T \in (M/2, M]$. Clearly a linear increase in M would tie T down to be a little smaller, but at the cost of many more failed attempts to achieve coalescence. The approximately symmetric shape of the distribution of $\log T$ also seems to support the geometric schedule. Empirically, for this example, the schedules (16,32,64,...), (8,32,128,...) and (32,128,512,...) all take about the same time on average, (4,8,16,...) takes 30% longer, but (8,16,24,...) takes twice as long, and (8,10,12,...) seven times longer.

Some further analysis of the running time implications of various details of CFTP implementation and use will be found in Murdoch and Rosenthal (1998).

These remarks perhaps also answer Professor Damien's question about "how far back should we go to get coalescence by time 0". The point is, of course, that we do not know, initially, but that the CFTP protocol provides a self-tuning mechanism for finding out. Incidentally, if CFTP is independently replicated, it is legitimate to change the M schedule for later runs in the light of experience about when coalescence typically occurs, gained from the early runs.

We now turn to questions about use of CFTP in an approximate mode. It may well be that in moderate- to high-dimensional Bayesian problems, "exact exact sampling" is too hard, and that some less aesthetically pleasing but still useful role for CFTP ideas will be the best we can do, as mentioned at the end of our conclusions. Thus we agree with Professor Damien that combining our methods with other strategies, like Johnson's, is a good idea. Even if exact sampling is impractical the couplers we present in this paper are fairly easy to implement and could be used as diagnostics instead. Non-exact uses of couplers are also discussed in Murdoch and Rosenthal (1998). Regarding Professor Tierney's ideas for a use for non-coalesced endpoints: unfortunately, we do not have a coupler that is both general purpose and easy to apply, where it is straightforward to come up with a finite list of candidates from an easily identifiable time. The rejection and multigamma couplers are too hard to apply, and the random walk couplers can have infinite sets of states that take a long time to move.

Finally, we know this is a Bayesian conference and a Bayesian volume, but we still think it is entirely appropriate to take a frequentist perspective in judging the quality of a Monte Carlo sample!

References

Murdoch, D. J. and Rosenthal, J. S. (1998). Efficient use of exact samples. This volume.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. J. Amer. Statist. Assoc., 90, 233–241.