Discussion of the Read Paper by A. Kong, P. McCullagh, D. Nicolae, Z. Tan, and X.L. Meng to the Society on December 11th, 2002

Christian P. Robert

CREST, INSEE, and CEREMADE, Université Dauphine, Paris

Past contributions of the authors to the Monte Carlo literature, including the notion of *bridge sampling*, are noteworthy, and I therefore regret that this paper does not possess a similar dimension for our field. Indeed, the "theory of Monte Carlo integration" advertised in the title reduces to a formal justification of bridge sampling, via non-parametric maximum likelihood estimation, while the device of pretending to estimate the dominating measure allows in addition for an asymptotic approximation of the Monte Carlo error through the corresponding Fisher information matrix. While this additional level of interpretation of importance and bridge sampling Monte Carlo approximations (rather than estimations) of integrals is welcomed, and rather exciting as a formal exercise, it seems impossible to derive a *working principle* out of the paper.

A remark of interest related to the supposedly unknown measure is that groups can be seen as acting on the measure rather than on the distribution. This brings much more freedom (but also embarrassment of wealth) in looking for group actions, since

$$\int_\Gamma q_s(x)d\lambda(x) = \int_{\Gamma/\mathcal{G}} \int_{\mathcal{G}} q_s(gx)d\nu(g)\, d\lambda(x) = \int_{\Gamma/\mathcal{G}} |\mathcal{G}|\, \bar{q}_s(x)d\lambda(x)$$

is satisfied for all groups $\mathcal{G}$ such that $d\lambda(gx) = d\lambda(x)$. However, this representation also exposes the confusion central to the paper between variance reduction, which is obvious by a Rao–Blackwell argument, and Monte Carlo improvement, which is unclear since the computing cost is not taken into account. For instance, Figure 1 analyses the example of Section 2 based on a *single realisation* from P_1, reparameterised to $[0,1]^2$, where larger groups acting on $[0,1]^2$ bring better approximations of $c_\sigma/c_1 = 1/\sigma^2$. But this improvement does not directly pertain to Monte Carlo integration, rather to numerical integration (with the curse of dimensionality lurking in the background). While the paper exhibits examples where using the group structure clearly brings substantial argument, it does not bring much light on the comparison of different groups from a Monte Carlo point of view.

The *numerical* nature of the improvement is also clear when one realises that, since the group action is on the measure rather than on the distribution, it often unlikely to be related to the geometry of this distribution, and thus can bring little improvement when $q(gx_i) = 0$ for most g's and generated x_i's. See Figure 2 for the case of a Monte Carlo experiment on a bivariate normal distribution with the same groups as above: the concentration of the logit transform of the $\mathcal{N}((-5,10), \sigma^2 I_2)$ distribution on a small area of the $[0,1]^2$ square *(left)* prevents good evaluations even after 4^{10} computations *(right)*.

A second opening related to the likelihood representation of the weight evaluation is that a Fisher information matrix can be associated with this problem and thus variance-like figures are produced in the paper. While these matrices stand as a formal ground for comparison of Monte Carlo strategies, I have difficulties with these figures, given that they do not necessarily provide a good approximation to the Monte Carlo errors. See for instance the example of Section 2: one could apply Slutsky's lemma with the transform $c_\sigma = \exp(\log(c_\sigma))$ to get the variance of the $\hat{c}_\sigma$'s as

$$\mathrm{diag}(c_\sigma)\mathbf{V}\mathrm{diag}(c_\sigma)\,,$$

but this approximation is invalidated by the absence of variance of the $\hat{c}_\sigma$'s. (See also the normal range confidence lower bound on Figure 2(right) which is completely unrelated with the range of the estimates.) Given that importance sampling estimators are quite prone to suffer from infinite variance (Robert and

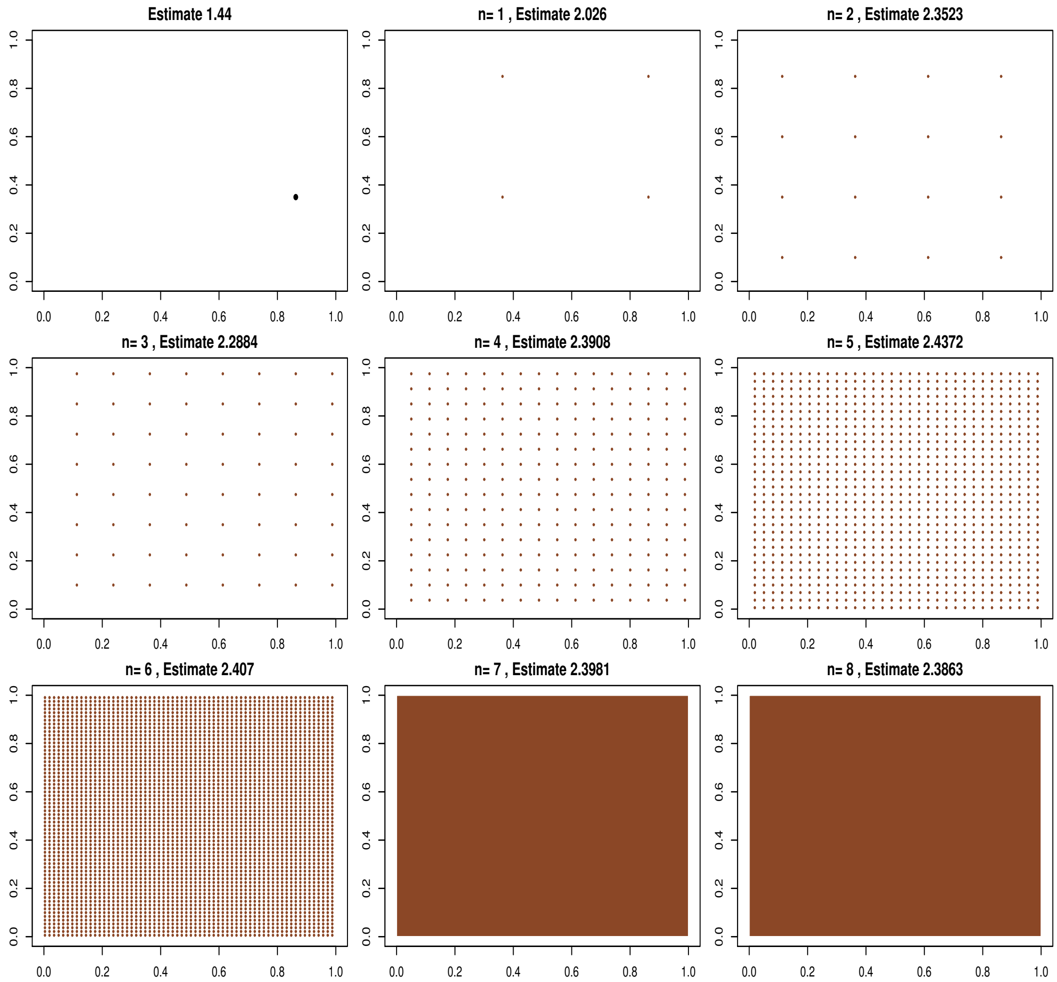

Fig. 1. Successive actions of the groups $\mathfrak{G}_n$ on the simulated value $(x_1, x_2) \sim P_1$ (upper left corner), where $\mathfrak{G}_n$ operates on $(z_1, z_2) = (\exp(x_1)/(1 + \exp(x_1)), x_2/(1 + x_2)) \in [0, 1]^2$ by changing some of the first n bits in the binary representation of the decimal part of z_1 and z_2. The estimates of $1/\sigma^2 = 2.38$ are given on top of each graph. The size of the orbit of $\mathfrak{G}_n$ is 4^n.

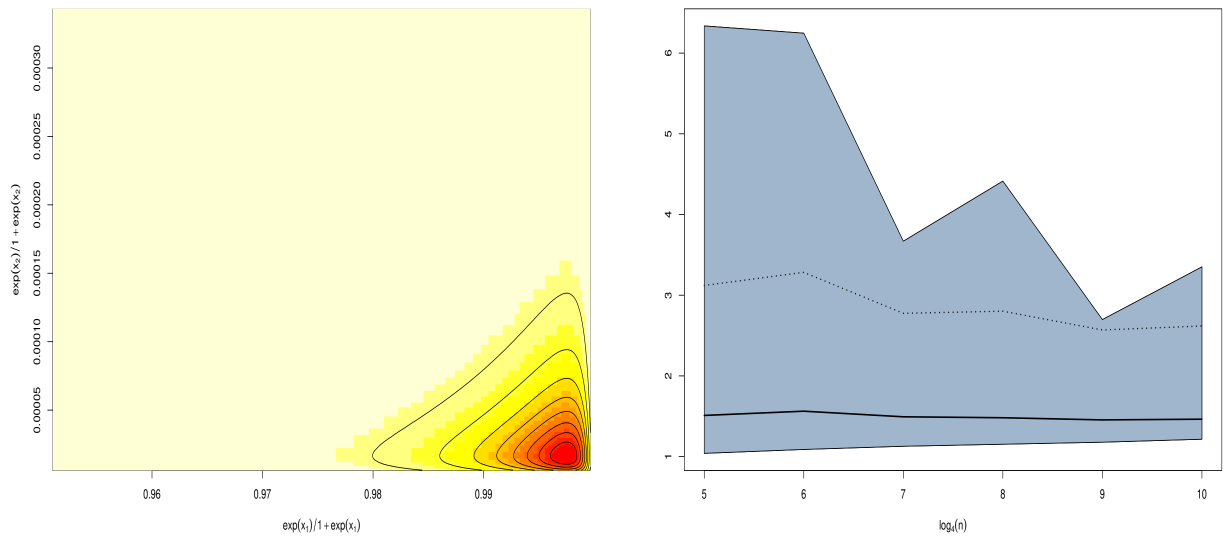

Fig. 2. *(left)* Contour plot of the density of the logit transform of the $\mathcal{N}((-5,10),\sigma^2 I_2)$ distribution *(right)* Summary of the successive evaluations of $c_\sigma/c_1 = \sigma^2 = 1.5$ by averages over the groups $\mathfrak{G}_n$, for the logit transform of a single simulated value $(x_1,x_2) \sim \mathcal{N}((-5,10),\sigma^2 I_2)$: the curve in bold is the average over 100 replications, the zone in grey is the range of the evaluations, and the dotted line is the two standard deviation upper bound.

Casella, 1999, Section 3.3.2), the appeal of using Fisher information matrices is somehow spurious as it gives a false confidence in estimators that should not be used.

The extension to MCMC settings is another interesting point in the paper, in that the estimator (3.10) reduces to

$$\hat{c}_r = \sum_{t=1}^{n} q_r(x_t) \Big/ \sum_{j=1}^{n} q_1(x_t|x_j)$$

if one uses a single chain based on a transition q_1 (with known constant). While this estimator only seems to apply for the Gibbs sampler, given that the general Metropolis–Hastings algorithm has no density against the Lebesgue measure, it is a special case of Rao–Blackwellisation (Gelfand and Smith, 1990; Casella and Robert, 1996) applied to importance sampling. As noted in Robert (1995), the naïve importance sampling alternative

$$\tilde{c}_r = \frac{1}{n} \sum_{t=1}^{n} q_r(x_t)/q_1(x_t|x_{t-1}),$$

where the "true" distribution of x_t is used instead, is a poor choice, since it most often suffers from infinite variance. (The notation in Section 3.6 is mildly confusing in that $c_r(x)$ is unrelated to c_r and is also most often known, on contrast to c_r. It thus seems inefficient to estimate the $c_r(x_t)$'s.)

Except for the misplaced statements of the last paragraph about "orthodox Bayesians", which simply bring to light the fundamental difference between designed Monte Carlo experiments and statistical inference, I enjoyed working on the authors' paper and thus unreservedly second the vote of thanks!

References

Casella, G. and Robert, C.P. (1996) Rao-Blackwellisation of sampling schemes. *Biometrika* **83**(1), 81–94.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling based approaches to calculating marginal densities. *J. Amer. Stat. Ass.* **85**, 398–409.

Robert, C.P. (1995) Convergence control techniques for MCMC algorithms. *Statis. Science* **10**(3), 231–253.

Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods.* Springer–Verlag, New York.