it. As a CV to be used jointly with the stratification, one may consider $H = A - E[A|B]$, with a coefficient $\beta(B)$ that depends on the value of B. The optimal coefficient is $\beta^*(B) = E[HL|B]/E[H^2|B]$ if the goal is to estimate $E[L]$. To estimate $\beta^*(b)$ as a function of b, one could estimate the two functions $q_1(b) = E[HL|B = b]$ and $q_2(b) = E[H^2|B = b]$ from the sample $\{(B_i, H_i, L_i),\ i = 1, \ldots, n\}$ of n values of (B, H, L), for example, using least-squares approximation to fit a curve $\hat{q}_1$ to the points $(B_i, H_i L_i)$ and another curve $\hat{q}_2$ to the points (B_i, H_i^2). The ratio will estimate the function $\beta^*(b)$. In the situations where this function is far from being a constant, this could make a significant difference compared with using the same β for all values of B.

CVS FOR FUNCTIONS OF SEVERAL EXPECTATIONS

The authors have considered a setting where linear CVs are used to correct the estimator of a *single* mathematical expectation estimated by a sample average. This could be generalized to the estimation of a function of several expectations, say, $g(\mu) = g(\mu_1, \ldots, \mu_d)$ by

$$g(\hat{X}_1, \ldots, \hat{X}_d) - \beta^T(\hat{H} - \theta),$$

where g is continuously differentiable at $(\mu_1, \ldots, \mu_d)$ and $\sqrt{n}(\hat{X}_1 - \mu_1, \ldots, \hat{X}_d - \mu_d)$ converges to a multinormal with mean zero when $n \to \infty$ (as in Glynn, 1994, e.g.). The asymptotically optimal β in this case is $\beta_{\mathrm{mc}} = (\mathrm{Cov}[\hat{H}])^{-1}\,\mathrm{Cov}[\hat{H}, \hat{X}]\nabla g(\mu)$, and similarly for RQMC, where $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_d)$. In other words, in the generalization it suffices to replace $\mathrm{Cov}[\hat{H}, \hat{I}]$ with $\mathrm{Cov}[\hat{H}, \hat{X}]\nabla g(\mu)$ in (15). One simple useful example of this is the estimation of a ratio of expectations, where $g(\mu_1, \mu_2) = \mu_1/\mu_2$.

ACKNOWLEDGMENTS

The author's work was supported by NSERC-Canada Grant ODGP0110050, NATEQ-Québec Grant 02ER3218, a Killam Research Fellowship and the Canada Research Chair in Stochastic Simulation and Optimization.

Comment: Computation, Survey and Inference

Xiao-Li Meng

1. THE SURVEY CONNECTION

1.1 Anticipating the "Surprises"

As someone who has benefited greatly from the sample survey literature, I am particularly pleased to see Hickernell, Lemieux and Owen's (HLO) emphasis on the equivalence between the control variates in Monte Carlo estimation and regression estimators in the sample survey literature. Indeed, the "surprises" described in HLO can be anticipated from similar phenomena in sample survey. For example, suppose that we, as a marketing firm, want to estimate the average household consumption of a certain product for the first six months of this year, based on a simple random sample (SRS) of a well-defined population of households (SRS is too simplistic for most practices, but adequate for the current discussion). Suppose a previous year's population counterpart is available (e.g., from a census source) for covariance adjustment (i.e., as a control variate). Let Y be the variable for the current semiannual consumption and let X represent the same period of the previous year. Given an SRS $\{(x_i, y_i),\ i = 1, \ldots, n\}$, asymptotically our best estimator is the well-known regression estimator

$$\hat{\mu}_y = \bar{y}_n - \hat{\beta}_{y,x}(\bar{x}_n - \mu_x), \tag{1.1}$$

where μ_x and μ_y are population averages, and $\hat{\beta}_{y,x}$ is the usual least-squares estimator from regressing Y on X.

Suppose, however, that we discover that the population average consumption for the first quarter, denoted by $\mu_{y^{(F)}}$, can be treated as known (e.g., there was a much larger survey for the first quarter by a different marketing firm). Then we can estimate μ_y by

Xiao-Li Meng is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: meng@stat.harvard.edu).

$\hat{\mu}_y^* = \mu_{y^{(F)}} + \hat{\mu}_{y^{(S)}}$, where $\mu_{y^{(S)}}$ denotes the population average for the second quarter, assuming $\{y_i^{(S)},\ i = 1, \ldots, n\}$ were available (e.g., we collected monthly consumption for the first six months). This setting mimics HLO's setting with $f(x) = f_{\mathrm{G}}(x) + f_{\mathrm{B}}(x)$, where the integration of f_{G} is done with no error by design, so all the estimation or integration errors come from the second component. [The analogy, of course, is not perfect because in HLO the choice of f_{G} depends on the design and f_{G} approaches f (in L^2) as the data size increases. In sample surveys, the estimand rarely depends on the choice of designs, including the sample size. Fortunately, these differences are immaterial for our current discussion because the use of control variates is postdesign and with a given finite sample size.]

This hypothetical survey example makes it clearer that as far as the estimation of $\mu_{y^{(S)}}$ goes, neither X nor $\beta_{y,x}$ is necessarily the best choice, even if they are for (1.1). It is likely that a better covariance adjustment for $Y^{(S)}$ is $X^{(S)}$, the second quarter consumption for the same previous year, perhaps due to the seasonality of the product. This is analogous to HLO's discussion in Section 4 with $f = f_{\mathrm{G}} + f_{\mathrm{B}}$ and $h = h_{\mathrm{G}} + h_{\mathrm{B}}$; since f_{G} and h_{G} do not contribute to the variance calculation, the goal is not to have h correlated with f, but rather h_{B} correlated with f_{B}. Furthermore, even if the semiannual consumption X is still a better covariance adjustment for $Y^{(S)}$ because $\mathrm{Corr}^2(X, Y^{(S)}) > \mathrm{Corr}^2(X^{(S)}, Y^{(S)})$, the regression slope in (1.1) will need to be changed from $\beta_{y,x}$ to $\beta_{y^{(S)},x}$. Therefore, unless $\mathrm{Corr}^2(X, Y^{(S)}) > \mathrm{Corr}^2(X^{(S)}, Y^{(S)})$ and $\beta_{y,x} = \beta_{y^{(S)},x}$, using $\hat{\beta}_{y,x}(\bar{x}_n - \mu_y)$ to adjust $\bar{y}_n^{(S)}$ will not produce an optimal estimator. This is in agreement with HLO's summary discussion at the beginning of Section 4.

1.2 When Does the Wrong Optimality Hurt?

Indeed, it is also well known in the survey literature that using a nonoptimal adjustment may actually do some harm compared to no adjustment, for example, in the context of comparing ratio estimators with SRS estimators (e.g., Cochran, 1977, Chapter 6). The same survey literature inspires the following general result regarding when it becomes harmful to use a wrong optimal regression adjustment compared to making no adjustment.

LEMMA 1. *Let*

$$\hat{\theta}_{\mathrm{opt}}^{(i)} = \hat{\theta}^{(i)} - \beta_{\mathrm{opt}}^{(i)}(\hat{\psi}^{(i)} - \psi^{(i)}), \quad i = 1, 2, \tag{1.2}$$

be two regression estimators for the same estimand θ, where $\beta_{\mathrm{opt}}^{(i)} = \mathrm{Cov}(\hat{\theta}^{(i)}, \hat{\psi}^{(i)})/\mathrm{Var}(\hat{\psi}^{(i)}) > 0$ is treated as known. Let

$$\hat{\theta}^{(1,2)} = \hat{\theta}^{(1)} - \beta_{\mathrm{opt}}^{(2)}(\hat{\psi}^{(1)} - \psi^{(1)}) \tag{1.3}$$

be the "wrong" regression estimator, that is, it uses $\hat{\psi}^{(1)} - \psi^{(1)}$ to adjust $\hat{\theta}^{(1)}$, but with the regression slope from the other estimator. Then $\mathrm{Var}(\hat{\theta}^{(1,2)}) > \mathrm{Var}(\hat{\theta}^{(1)})$ if and only if

$$\left|\frac{\beta_{\mathrm{opt}}^{(2)}}{\beta_{\mathrm{opt}}^{(1)}} - 1\right| > 1, \quad \textit{that is,} \quad \frac{\beta_{\mathrm{opt}}^{(2)}}{\beta_{\mathrm{opt}}^{(1)}} > 2 \quad \textit{or} \quad \frac{\beta_{\mathrm{opt}}^{(2)}}{\beta_{\mathrm{opt}}^{(1)}} < 0. \tag{1.4}$$

The proof of this lemma follows directly from the fact that

$$\mathrm{Var}(\hat{\theta}^{(1,2)}) = \mathrm{Var}(\hat{\theta}^{(1)}) - [\beta_{\mathrm{opt}}^{(1)}]^2 \mathrm{Var}(\hat{\psi}^{(1)}) + [\beta_{\mathrm{opt}}^{(2)} - \beta_{\mathrm{opt}}^{(1)}]^2 \mathrm{Var}(\hat{\psi}^{(1)}).$$

This result provides theoretical support of HLO's empirical finding that the use of β_{MC} still often leads to useful improvement with QMC, because it assures us that unless the regression slope changes substantially, that is, either it changes the sign or it is at least twice as large in magnitude, the use of the wrong regression slope is still beneficial compared to not making any adjustment, regardless of whether or not we use the same control covariate. For HLO's "cautionary example" (Section 4.1), $\beta_{\mathrm{MC}} = 1 - 2M^{-2} > 0$, but $\beta_{\mathrm{RQMC}} = -1$, so there is a switching of the sign of the regression slope. Consequently, using β_{MC} in place of β_{RQMC} will lead to an estimator with larger variance than the RQMC estimator without adjusting for the control variate. Note that in HLO's example, $\hat{\psi}^{(1)} = \hat{\psi}^{(2)}$; indeed Lemma 1 can be recast with only one regression class estimator, $\hat{\theta}_\beta = \hat{\theta} - \beta(\hat{\psi} - \psi)$, and then using a nonoptimal β becomes harmful if and only if $|(\beta/\beta_{\mathrm{opt}}) - 1| > 1$. Also note that in real applications the regression slope is seldom known and will be replaced by its least-squares estimator. This replacement, however, does not affect the conclusion of Lemma 1 asymptotically because of the forgiving nature of the regression estimators to the error in the slope, as discussed toward the end of Section 3 of HLO.

It is also known from the survey literature that the use of regression estimators tends to have diminishing gains for stratified sample designs relative to SRS, be-

cause covariance/regression adjustment is essentially a form of (deep) stratification. Consequently, unless the two stratifying variables are uncorrelated with each other, the stratified design has already "achieved" a part of gain in efficiency intended by the regression adjustment. The degree of the "achievement" depends on how deep the original stratification is in the sampling design. Since QMC designs, especially the more advanced ones as reviewed in HLO, are often very deep stratifications (compared to the types of stratifications in sample surveys), it comes as no surprise that the gains of using control variates tend to be noticeably less pronounced for QMC than for MC, as summarized in Section 10 of HLO.

1.3 Why Do We Need to Go beyond the Design-Based Perspective?

The sampling survey, or more generally the design-based perspective, however, does not explain everything. Consider the following question/comparison. In the semiannual consumption example in Section 1.1 we had

$$\hat{\mu}_y = h(\hat{\mu}_{y^{(F)}}, \hat{\mu}_{y^{(S)}}) \equiv \hat{\mu}_{y^{(F)}} + \hat{\mu}_{y^{(S)}}. \tag{1.5}$$

When the true value of $\mu_{y^{(F)}}$ is known, it is almost impossible to resist the temptation to replace $\hat{\mu}_{y^{(F)}}$ with its true value in $h(\hat{\mu}_{y^{(F)}}, \hat{\mu}_{y^{(S)}})$ to form $\hat{\mu}_y^* = h(\mu_{y^{(F)}}, \hat{\mu}_{y^{(S)}}) = \mu_{y^{(F)}} + \hat{\mu}_{y^{(S)}}$ to estimate μ_y. Indeed, why not? How could we get hurt, as far as efficiency/variance goes, by taking advantage of as much truth as we know?

Now consider the regression estimator given in (1.1), which can also be written as

$$\hat{\mu}_y = g(\bar{y}_n, \bar{x}_n, \hat{\beta}_{y,x}) = \bar{y}_n - \hat{\beta}_{y,x}(\bar{x}_n - \mu_x). \tag{1.6}$$

It is legitimate to consider (1.1) as a function of $\bar{y}_n, \bar{x}_n$ and $\hat{\beta}_{y,x}$ only, because only these quantities depend on the sample. Putting it differently, we can give a user a "black-box" software routine that computes the value of $\hat{\mu}_y$, with $\bar{y}_n, \bar{x}_n$ and $\hat{\beta}_{y,x}$ as input, calculated from the user's particular sample. Suppose that the user accidentally discovered that the population true value of μ_x was actually available from a census source, just as we (hypothetically) discovered that the true value of $\mu_{y^{(F)}}$ was available. Now if the user adopts the same reasoning/intuition as we did with h, then she or he would surely input μ_x in g in place of her or his sample average $\bar{x}_n$. However, this action will completely wipe out the regression adjustment. See Liu, Rubin and Wu (1998) for a similar discussion in the context of viewing the PX–EM algorithm as a covariance adjusted EM algorithm.

One may argue that the problem occurred simply because the user did not understand the actual form of the estimator, but this is exactly the issue: For a general estimation procedure, which can be of arbitrary complexity, how can we tell when it is and when it is not beneficial to substitute a part of our estimation procedure by a more precise estimator (including its true value)? This question is particularly relevant for Monte Carlo estimators, be they quasi or not, because in a simulation setting, nothing is *unknown*, in its original sense. Consequently, the formulation of optimal estimators based on simulated data will depend intricately on how we model what we *ignore*, not what we know—a question that is beyond the realm of any design-based perspective. A different perspective is therefore needed, which is the subject of the next section. In particular, we shall see how the new perspective leads to a new interpretation of control variates and, more importantly, leads to a new control-variate estimator that appears to be difficult to anticipate from the traditional design-based perspective of Monte Carlo integration or of sample survey.

2. THE INFERENCE CONNECTION

2.1 Why Does Likelihood Inference Appear to Be Useless with Simulated Data?

To define optimality meaningfully, we first need to quantify what data and model assumptions we permit ourselves to use. In a real-data analysis, once the data are collected or provided, the central challenge typically is to postulate a suitable set of reasonable assumptions, parametric or nonparametric, to link our data with our estimand of interest. Once the model is posited and a measure of efficiency is chosen (e.g., variance), the corresponding optimality can then be quantified theoretically, at least asymptotically (e.g., via Fisher information).

The above discussion might lead us to believe that quantifying optimality with simulated data is an easier task, because there is no issue of model uncertainty, for we are the one who generated all the data (or design points). Ironically, the issue turns out to be far more complicated, precisely because we know too much. To illustrate, consider importance sampling, as discussed in HLO. We are interested in the value of $c_1 = \int_\Omega q_1(x)\mu(dx)$, where $q_1(x)$ is our known integrand and μ is the baseline measure, typically Lebesgue or counting. We have draws from a trial density $p_2 = q_2/c_2$, denoted by $\{X_{i2},\ i = 1, \ldots, n_2\}$. Then

the well-known importance sampling identity

$$r \equiv \frac{c_1}{c_2} = E_2\left[\frac{q_1(X)}{q_2(X)}\right], \tag{2.1}$$

where E_2 is the expectation with respect to p_2, provides us with an estimation equation from which we arrive at the well-known importance sampling (IS) estimator

$$\hat{r} = \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{q_1(X_{i2})}{q_2(X_{i2})}. \tag{2.2}$$

Note that in common IS settings, as in HLO, c_2 is chosen to be 1 and thus $r = c_1$, but in more general settings ratios are of interest; see Meng and Schilling (2002) for a recent discussion of this issue.

So on what basis can we claim (2.2) is optimal? How do we know there is no other estimation equation that can deliver a more efficient estimator than (2.1) can? Since asymptotically the maximum likelihood estimator is most efficient (under standard regularity conditions) and since asymptotic arguments are more relevant for simulated data because the size of data is under our control, we naturally wonder what the well established likelihood theory can tell us for such questions. For simplicity, let us assume that the draws from $p_2 = q_2/c_2$ are i.i.d. Then the density of our "data" $\{X_{i2},\ i = 1, \ldots, n_2\}$ is given by

$$p(X_{12}, \ldots, X_{n_2 2}) = \prod_{i=1}^{n_2}\frac{q_2(X_{i2})}{c_2}. \tag{2.3}$$

The above expression immediately suggests that something is quite amiss. On one hand, our estimand c_1 does not even appear in our "likelihood function" (2.3). On the other hand, it is clear that without $\{X_{i2},\ i = 1, \ldots, n_2\}$, we do not even have the IS estimator (2.2). So could this be an obvious counterexample to the likelihood principle?

Take bridge sampling as another example. Bridge sampling is a generalization of importance sampling, as described by Meng and Wong (1996). Here our goal is still to estimate $r = c_1/c_2$, as in the IS setting. The difference is that we now have draws from both $p_1 = q_1/c_1$ and $p_2 = q_2/c_2$, denoted by $\{X_{ij},\ i = 1, \ldots, n_j\},\ j = 1, 2$. Since q_1 and q_2 are assumed to be known, under the assumption of independent draws, the "likelihood" for c_1 and c_2 becomes

$$\begin{aligned} &L(c_1, c_2 | \{X_{ij},\ i = 1, \ldots, n_j\},\ j = 1, 2) \\ &\quad = \prod_{j=1}^{2}\prod_{i=1}^{n_j}\frac{q_j(X_{ij})}{c_j} \propto c_1^{-n_1}c_2^{-n_2}, \end{aligned} \tag{2.4}$$

which is free of any data! So once again, the likelihood method seems to fail, whereas estimators based on the estimation equation approach abound (see Meng and Wong, 1996).

One answer to the above paradoxes is simply that likelihood methods are not applicable to simulated data. Whereas logically this is an admissible answer, if it were true, it certainly would be the most disturbing puzzle lying in the foundation of likelihood inference, at least to some of us. How could it be? How could an inferential method so powerful with an uncertain data-generating mechanism becomes completely useless when the mechanism is completely known?

2.2 The Answer: Because We Were Looking at the Wrong Parameter!

An astute reader may have already seen a hidden problem with the "likelihood" as given in (2.4). The normalizing constant c_j is deterministically related to q_j via

$$c_j = \int_\Omega q_j(x)\mu(dx), \quad j = 1, 2. \tag{2.5}$$

So when we ignore $q_j(X_{ij})$ from (2.4) because they are known, we actually have also effectively ignored a part of the "parameter" that our likelihood intends to infer. A closer inspection of (2.5) reveals that the problem is far more serious than just appropriately sorting out the connection between c_j and $q_j(X_{ij})$. The problem is that it is impossible to treat c_j as an unknown parameter when we treat q_j as known, unless we can treat the baseline measure μ as unknown. In other words, when we treat both q_j and μ as known, there is no statistical inference problem for c_j to speak of, since c_j is completely determined by q_j and μ. Putting it differently, although c_j's or their ratios are what we are after, they cannot be the *only unknown* model parameters for any meaningful statistical modeling.

To resolve this problem, Kong, McCullagh, Meng, Nicolae and Tan (2003) proposed to conduct the likelihood inference by treating the baseline measure μ as the unknown parameter and then to estimate c_j as a linear functional of μ via (2.5). With this approach, (2.3) becomes a well-defined and meaningful likelihood in the form of

$$\begin{aligned} &L(\mu | X_{12}, \ldots, X_{n_2 2}) \\ &\quad = \prod_{i=1}^{n_2}\frac{q_2(X_{i2})\mu(X_{i2})}{\int q_2(x)\mu(dx)} \propto \frac{\prod_{i=1}^{n_2}\mu(X_{i2})}{[\int q_2(x)\mu(dx)]^{n_2}}, \end{aligned} \tag{2.6}$$

where $\mu(X) = \mu(\{X\})$ or $\mu(\{dX\})$. The maximum likelihood estimator of μ, among all possible nonnegative measures, is given by $\hat{\mu}(x) \propto P_{n_2}(x)/q_2(x)$, where $P_{n_2}(x)$ is the usual empirical measure, with n_2^{-1} mass at each observed X_{i2}. Clearly from (2.6), μ (and thus c_j's) can only be estimated up to a multiplicative constant. Substituting μ in (2.5) with $\hat{\mu}$ shows that $\hat{r}$ of (2.2) is indeed the (nonparametric) maximum likelihood estimator (MLE) of r under the likelihood (2.6). This suggests that, without employing any other information, $\hat{r}$ of (2.2) is indeed (asymptotically) the best possible estimator of r given $\{X_{i2},\ i = 1, \ldots, n_2\}$. Similarly, Kong et al. (2003) have shown that the optimal bridge sampling estimator given in Meng and Wong (1996) is the same as the MLE when we have $\{X_{ij},\ i = 1, \ldots, n_j;\ j = 1, 2\}$ as our data.

The reason why this likelihood perspective can easily resolve these paradoxes is that it captures the real inference structure of Monte Carlo integration. Specifically, Monte Carlo simulation means that we use *samples* to represent, and therefore effectively *estimate*, the underlying population $q_j(x)\mu(dx)$, and hence *estimate* μ since q_j is known. One may find the phrase "estimate" puzzling because we invariably know what μ is (e.g., Lebesgue or counting). However, our knowledge of μ is never used in any way, for example, in forming (2.2). This can be best seen by considering that there are two individuals: a simulator and an analyst. The simulator provides the simulated data $\{X_{i2},\ i = 1, \ldots, n_2\}$ to the analyst, who has the task of estimating r. The analyst is also given both q_1 and q_2, but is never told about the actual μ used in simulation. Nevertheless, the analyst can consistently estimate r, which obviously depends on μ, as long as the support of q_1 does not exceed that of q_2. (This well-known condition on the supports can also be clearly seen from the likelihood perspective, because we can only make inference about μ on a support that is identifiable from the data $\{X_{i2},\ i = 1, \ldots, n_2\}$.) Consequently, as far as (2.2) goes, μ is completely unknown; more precisely, no knowledge of μ is used in (2.2) and thus it is legitimate (and actually necessary) to treat μ as the unknown model parameter.

The above discussion also suggests that we can use partial knowledge of μ to improve upon (2.2), as long as the resulting MLE for r is still easy to compute. Clearly we should not use our full knowledge about μ, which will lead us back to the infeasible analytic calculation required by (2.5). For example, since Lebesgue measure is invariant to reflection with respect to the origin, we can restrict our parameter space to all nonnegative measures that satisfy this invariance property, if the true μ is indeed Lebesgue. The resulting MLE of r is

$$\hat{r}^* = \frac{1}{n_2}\sum_{i=1}^{n_2} \frac{q_1(X_{i2}) + q_1(-X_{i2})}{q_2(X_{i2}) + q_2(-X_{i2})}, \tag{2.7}$$

which is the Rao–Blackwellization treatment of $\hat{r}$ by averaging over the orbit of the reflection group $\{I, -I\}$, and hence its variance never exceeds that of $\hat{r}$ (under the assumption of i.i.d. draws). See Kong et al. (2003) for a general formulation of using group invariance to restrict the parameter space for μ and hence to improve Monte Carlo efficiency. Also see Casella (1996) for a detailed discussion of the use of Rao–Blackwellization methods in Monte Carlo simulation and, more generally, the interrelationship between statistical inference theory and computational algorithms.

2.3 Indeed a Surprise: An Unexpected Control-Variate Estimator and Insight

Another fundamental advantage of this likelihood approach is that it provides a unified framework for investigating variance reduction techniques, including control variates. In the importance sampling context, when we use a g with

$$\int_\Omega g(x)\mu(dx) = 0 \tag{2.8}$$

as a control variate, we effectively put a constraint on the unrestricted parameter space $\Theta_\mu = \{\mu : \text{all nonnegative measures on } \Omega\}$. Consequently, the MLE under this submodel will be more efficient than the MLE under the full model. The resulting MLE for r under this constraint, however, is not the usual regression estimator, albeit asymptotically they are equivalent, as they should be.

Specifically, because any measure with zero mass at any single observation will lead to a zero likelihood in (2.6), the maximization of (2.6) under constraint (2.8) is effectively discrete, as is typical with nonparametric or empirical MLE (e.g., Owen, 2001). The discrete problem we need to solve is

$$\max_{\mu \in \Theta_{n_2}^{(g)}} \left\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\left[\sum_{i=1}^{n_2} q_{2i}\mu_i \right] \right\}, \tag{2.9}$$

where, for simplicity, we have let $\mu_i = \mu(X_{i2})$,

$q_{2i} = q_2(X_{i2})$, $g_i = g(X_{i2})$ and

$$\Theta_{n_2}^{(g)} = \Bigg\{(\mu_1, \ldots, \mu_{n_2}) : \mu_i > 0,\ i = 1, \ldots, n_2; \text{ and } \sum_{i=1}^{n_2} g_i \mu_i = 0\Bigg\}. \tag{2.10}$$

Tan (2003) presented an elegant solution to this maximization problem under the more general setting with multiple control variates. The following is a slightly more elementary recast of Tan's (2003) derivation.

We start by assuming condition (A): $\min_i g_i < 0$ and $\max_i g_i > 0$. This is not a real restriction in view of (2.8) and relatively large n_2 in practice, but technically it is a necessary and sufficient condition for (2.9) to have a solution. Clearly it is necessary, because without it, $\Theta_{n_2}^{(g)}$ will be empty. The sufficiency is established by the following argument, which shows that (2.9) has the unique maximizer when condition (A) holds.

First, because $\sum_{i=1}^{n_2} g_i \mu_i = 0$, (2.9) is the same as

$$\max_{\mu \in \Theta_{n_2}^{(g)}} \Bigg\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\Bigg[\frac{1}{n_2} \sum_{i=1}^{n_2} (q_{2i} + \lambda g_i)\mu_i \Bigg] - n_2 \log n_2 \Bigg\} \tag{2.11}$$

for any $\lambda \in \Lambda_{n_2} = \{\lambda : q_{2i} + \lambda g_i > 0,\ i = 1, \ldots, n_2\}$, which is nonempty because it contains at least $\lambda = 0$ since all $q_{2i} > 0$ by our sample design. Consequently, by Jensen's inequality applied to the second log expression in (2.11), we obtain

$$\begin{aligned} &\max_{\mu \in \Theta_{n_2}^{(g)}} \Bigg\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\Bigg[\sum_{i=1}^{n_2} q_{2i} \mu_i \Bigg] \Bigg\} \\ &\quad \le - \sum_{i=1}^{n_2} \log(q_{2i} + \lambda g_i) - n_2 \log n_2, \end{aligned} \tag{2.12}$$

where the equality holds if and only if

$$\mu_i \propto \frac{1}{q_{2i} + \lambda g_i} \quad \text{and} \quad \sum_{i=1}^{n_2} g_i \mu_i = 0. \tag{2.13}$$

Since (2.12) holds for any $\lambda \in \Lambda_{n_2}$, we can minimize the right-hand side over λ, which leads to

$$\begin{aligned} &\max_{\Theta_{n_2}^{(g)}} \Bigg\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\Bigg[\sum_{i=1}^{n_2} q_{2i} \mu_i \Bigg] \Bigg\} \\ &\quad \le - \max_{\lambda \in \Lambda_{n_2}} \sum_{i=1}^{n_2} \log(q_{2i} + \lambda g_i) - n_2 \log n_2. \end{aligned} \tag{2.14}$$

Second, we can show that the inequality in (2.14) actually is an equality. This is because, under condition (A), Λ_{n_2} is a finite open interval containing zero and

$$\ell(\lambda) \equiv \sum_{i=1}^{n_2} \log(q_{2i} + \lambda g_i) \tag{2.15}$$

is a strict concave and differentiable function on Λ_{n_2}. Consequently, $\ell(\lambda)$ has the unique maximum $\hat{\lambda} \in \Lambda_{n_2}$, which satisfies

$$\frac{d\ell(\hat{\lambda})}{d\lambda} = \sum_{i=1}^{n_2} \frac{g_i}{q_{2i} + \hat{\lambda} g_i} = 0. \tag{2.16}$$

In other words, when we let $\lambda = \hat{\lambda}$ in (2.13), the resulting $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_{n_2})$ indeed satisfies the constraint in (2.13), and therefore this, and only this, choice of μ equates the two sides of (2.14). Consequently,

$$\hat{\mu}(x) \propto \frac{P_{n_2}(x)}{q_2(x) + \hat{\lambda} g(x)} \tag{2.17}$$

is the unique solution to (2.9), where $P_{n_2}(x)$ is the standard empirical measure based on $\{X_1, \ldots, X_{n_2}\}$. The corresponding MLE of r is given by

$$\hat{r}_{\mathrm{MLE}} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{q_1(X_{i2})}{q_2(X_{i2}) + \hat{\lambda} g(X_{i2})}. \tag{2.18}$$

The form of this MLE is rather intriguing. First, unlike the standard regression estimator, which takes a linear form for adjustment, $\hat{r}_{\mathrm{MLE}}$ retains a ratio form. The advantage of the ratio form is that it ensures the nonnegativity of $\hat{r}_{\mathrm{MLE}}$ whenever the integrand q_1 is nonnegative. This is, of course, expected because $\hat{r}_{\mathrm{MLE}}$ is an MLE and hence it must be within the original allowable space of r (as determined by our usable knowledge of q_1). In contrast, the regression estimator does not have this property. Asymptotically, however, linear adjustment is all one needs, and thus $\hat{r}_{\mathrm{MLE}}$ is equivalent to the regression estimator by a Taylor expansion argument, as given in Tan (2003).

Second, $\hat{r}_{\mathrm{MLE}}$ has the same form as the IS estimator (2.2), but with $q_2(x) + \hat{\lambda} g(x)$ as the "trial" density. This can be seen more clearly when our control variate is introduced by using an unnormalized density q_3 such that $\int q_2(x)\mu(dx) = \int q_3(x)\mu(dx)$ (see Kong et al., 2003, for an illustration), that is, $g(x) = q_3(x) - q_2(x)$. Then the function in the denominators in (2.18) becomes a mixture of q_2 and q_3, $(1 - \hat{\lambda})q_2 + \hat{\lambda} q_3$, where $\hat{\lambda}$ is the MLE of the mixture weight λ from fitting the mixture model $(1 - \lambda)q_2 + \lambda q_3$ to the simulated data

$\{X_{i2},\ i = 1, \ldots, n_2\}$. (Note that here λ is not restricted to the unit interval, as long as it is inside $\Theta_{n_2}^{(g)}$.)

This fitting aspect is the most intriguing part of the MLE approach because the true value of λ is known to be zero, since all the data were drawn from q_2. However, with any finite sample, the best fitted $\hat{\lambda}$ under the mixture model will almost surely deviate from the true value $\lambda = 0$, indicating an "imperfection" of the sample to represent the intended population q_2. The MLE approach uses this deviation to adjust for the imperfection via the known relationship (2.8), in the same spirit as the regression estimator uses $\bar{x}_n - \mu_x$ to adjust. Specifically, just as the regression estimator (1.1) effectively treats an "imperfect" sample $\{y_1, \ldots, y_n\}$ with mean μ_y as a "perfect" sample with mean $\mu_y + \beta_{y,x}(\bar{x}_n - \mu_x)$, the MLE treats an imperfect sample from q_2 as a perfect sample from $(1 - \hat{\lambda})q_2 + \hat{\lambda}q_3$: It is perfect as far as estimating $\int_\Omega g(x)\mu(dx) = 0$ goes because of (2.16). The MLE then uses this "perfect" model/sample to perform the usual importance sampling, as in (2.18). This construction appears to be difficult to conceive from a purely design-based perspective, which inevitably would only call for inverse-probability weight $1/q_2(X)$, since X was drawn from q_2. In particular, this is another example where the use of the fitted value is better than using the truth, as discussed in Section 1.3.

2.4 Possible Applications to QMC and Surveys

The discussion so far centers on MC designs, where there is a natural sampling distribution and hence a natural likelihood. The central issue there is to recognize what the correct model parameter is. For deterministic QMC, this approach is not directly applicable since there is no sampling distribution in the design. However, when randomness is reintroduced into QMC, as with the RQMC methods discussed in HLO, the likelihood method appears to be applicable, albeit the implementation could be more complicated in view of the more stratified nature of the design compared to i.i.d. or even the more general MCMC designs, which are typically without stratification. In addition, there appear to be more constraints on μ such as $\int f_G(x) f_B(x)\mu(dx) = 0$ with the QMC methods (Section 2.1 of HLO). It would be interesting to see the form of the resulting MLE for $\int [f_G(x) + f_B(x)]\mu(dx)$ under the likelihood approach.

For deterministic QMC, although the likelihood approach is not directly applicable (and this time there is no paradox, because there is no random data-generating mechanism to start with), the inference perspective is still very fruitful. This was, for example, discussed by Diaconis (1988), where a Bayesian approach, which does not necessarily require a sampling scheme or a likelihood, was investigated. This approach is to put a prior model—a stochastic process—on the integrand q, with q's values at the design points as the observations. The inference is then carried out by computing the conditional distribution of the process, and hence the integration, given the observations. The advantage of this class of methods is that, by choosing appropriate stochastic models, one can take into account known properties of the integrand q. In contrast, our likelihood approach takes advantage of usable known properties of the baseline measure, either via group restrictions or other constraints such as control variates. As a result, the Bayesian approach can produce much more efficient results for specific integrands. Indeed, many well-known numerical integration methods can be rederived from this perspective, as shown by Diaconis (1988) and the references therein. On the other hand, the MLEs obtained under the likelihood approach are much more generally applicable, but they can be made more efficient if specific knowledge of the integrand (e.g., differentiability) can be utilized. So the two approaches complement each other and, ideally, we would like to have a combined inference method that will model the usable knowledge of both the baseline measure and the integrand. Research in this direction is very much needed, and HLO's investigation of using control variates with RQMC methods can be viewed as an important step in this direction because it takes into account both the properties of the integrand and the restriction on the baseline measure via the use of the control variates.

Finally, to complete the circle, the new ratio-type control-variate estimator also suggests a possible corresponding counterpart for sample survey applications, where the two standard estimators for covariance adjustments have been the direct ratio estimator [i.e., $\hat{\mu}_y = (\bar{y}_n/\bar{x}_n)\bar{\mu}_x$] and the regression estimator (1.1). Such a counterpart, if it exists, would be of direct practical value, because it retains important advantages of both the ratio estimator and the regression estimator, as we discussed in Section 2.3, especially considering that many survey estimands are positive by nature.

3. FURTHER CONNECTIONS BETWEEN MCMC AND QMC

As HLO correctly pointed out in their Section 2.5, both MCMC and QMC have a long history and both

have grown rapidly in recent years, yet there is very little overlap between the two fields. This is certainly a very unfortunate and ironic situation, considering that both fields share exactly the same goal. HLO's paper is certainly a very timely contribution to changing this situation—a change that is much needed, because the two fields can learn a great deal from each other, as HLO's paper clearly demonstrates. Here I want to add two topics from recent work that I was involved in to demonstrate the great benefit of using techniques and ideas from both fields.

The first topic is path sampling, which is a generalization of bridge sampling with infinitely many bridges, as well as a general formulation of thermodynamic integration in statistical physics, as shown by Gelman and Meng (1998). The method is particularly suited for handling some very high-dimensional integrations, as discussed by Ogata (1989). The key identity that underlies path sampling expresses $\log r$, where r is the same as in (2.1), as a low-dimensional integration over a prior parameter of a high-dimensional expectation that is conditional on the parameter. This presents an ideal situation to use both MCMC methods and QMC methods, with the former applied to estimate the high-dimensional expectation and the latter applied to numerically estimate the outside low-dimensional integration. The effectiveness of such a hybrid approach was demonstrated by Gelman and Meng (1998), where very basic numerical approaches (e.g., trapezoidal rule; rectangular lattices) were used for the low-dimensional integrations. It is likely that the effectiveness will be even more impressive if the more advanced QMC methods, such as those reviewed in HLO, are used for these low-dimensional integrations.

The second topic is multiprocess parallel antithetic coupling for backward and forward MCMC (Craiu and Meng, 2005). Using antithetic variates is a very old variance reduction technique in the Monte Carlo literature (e.g., Hammersley and Morton, 1956). However, in the standard MCMC literature, typically only a pair of antithetic variables is used (e.g., Frigessi, Gåsemyr and Rue, 2000). Viewing antithetic variates as a form of stratification, employing more than two strata becomes an obvious next step. However, unlike the case of using a pair, generating a set of $k > 2$ antithetic variates is not a trivial task. This is because there is no unique way to generate $k > 2$ antithetic variates that are *negatively associated* (i.e., preserve negative correlation under monotone transformation) and *extremely antithetical* (i.e., as negatively correlated as possible). Nevertheless, we (Craiu and Meng, 2005) found that Latin hypercube sampling, as mentioned in Section 6 of HLO, as well as an iterative extension of it, serves as an effective general-purpose scheme. The advantages of running multiprocess antithetically coupled MCMC, for both the standard forward implementation and the backward perfect-sampling implementation (see Casella, Lavine and Robert, 2001, for an introduction), include not only further reduction of Monte Carlo variances compared to using $k = 2$, but also reduction of biases due to slow mixing, because antithetically coupled chains can search a state space more thoroughly compared with using k independent chains, which is the current common recommendation (e.g., Gelman and Rubin, 1992).

In conclusion, I thank HLO for writing this timely and inspiring article and the Editor for inviting me to discuss it. Given the clear benefit of cross-fertilization between MCMC and QMC, I hope this set of discussion articles can serve as a successful matchmaker for a long, happy and (re)productive marriage between QMC and MCMC!

ACKNOWLEDGMENTS

I thank Radu Craiu, Andrew Gelman, Martin Romero and Zhiqiang Tan for helpful comments and exchanges. The research was supported in part by NSF Grant DMS-02-04552.

Rejoinder

Fred J. Hickernell, Christiane Lemieux and Art B. Owen

We thank Professor L'Ecuyer and Professor Meng for their thoughtful remarks. We particularly liked L'Ecuyer's concise summary of combining variance reduction techniques, and Meng's references to combinations of antithetic and Latin hypercube sampling with MCMC. Our reply is organized by topic.