On a Likelihood Approach for Monte Carlo Integration

Zhiqiang TAN

The use of estimating equations has been a common approach for constructing Monte Carlo estimators. Recently, Kong et al. proposed a formulation of Monte Carlo integration as a statistical model, making explicit what information is ignored and what is retained about the baseline measure. From simulated data, the baseline measure is estimated by maximum likelihood, and then integrals of interest are estimated by substituting the estimated measure. For two different situations in which independent observations are simulated from multiple distributions, we show that this likelihood approach achieves the lowest asymptotic variance possible by using estimating equations. In the first situation, the normalizing constants of the design distributions are estimated, and Meng and Wong's bridge sampling estimating equation is considered. In the second situation, the values of the normalizing constants are known, thereby imposing linear constraints on the baseline measure. Estimating equations including Hesterberg's stratified importance sampling estimator, Veach and Guibas's multiple importance sampling estimator, and Owen and Zhou's method of control variates are considered.

KEY WORDS: Bridge sampling; Control variate; Importance sampling; Stratified sampling.

1. INTRODUCTION

Monte Carlo is a useful method for numerical integration. Specifically, let μ_0 be a nonnegative measure on a state space $\mathcal{X}$ and consider evaluating the integral

$$Z = \int_{\mathcal{X}} q(x)\, d\mu_0$$

for a real-valued function $q(x)$. We refer to μ_0 as the baseline measure, typically counting measure or Lebesgue measure. It is helpful to distinguish two different issues of design and estimation in Monte Carlo integration.

First, various sampling designs have been proposed for Monte Carlo integration. Importance sampling involves simulating observations from a single distribution. Generally, bridge sampling or stratified mixture sampling involves simulating observations from multiple distributions (see Geyer 1994; Hesterberg 1995; Meng and Wong 1996; Owen and Zhou 2000). For $1 \le j \le m$, let $q_j(x)$ be a nonnegative function whose integral Z_j is finite and positive with respect to μ_0. Then

$$dP_j = \frac{q_j(x)}{Z_j}\, d\mu_0$$

is a probability distribution, and Z_j is called the normalizing constant for the sampler P_j. For convenience, assume that for each x, $q_j(x) > 0$ for at least one j; otherwise, we can replace $\mathcal{X}$ by the union of the supports of $q_j(x)$. Suppose that a stream of n_j independent observations $\{x_{j1}, \dots, x_{jn_j}\}$ is available from P_j by a simulation technique. Denote by $\{x_1, \dots, x_n\}$ the pooled sample of size $n = \sum_{j=1}^m n_j$, and by P_* the distribution $n^{-1}\sum_{j=1}^m n_j P_j$. In asymptotic considerations, let each n_j tend to infinity such that n_j/n is fixed.

The second issue of estimation is the one we are concerned with in this article. For importance sampling ($m = 1$), the ratio Z/Z_1, or Z relative to Z_1, can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{q(x_i)}{q_1(x_i)}. \tag{1}$$

At first sight, this estimator is constructed via the identity

$$\frac{Z}{Z_1} = E_1\left[\frac{q(x)}{q_1(x)}\right],$$

where E_1 denotes expectation under P_1. However, a statistician would ask the fundamental question "what model underlies the given estimator." In the usual sense, there is no unknown quantity, because the simulated data are generated from a process completely controlled by the statistician.

Recently, the foregoing question was satisfactorily addressed by a statistical formulation, making explicit what information is ignored and what is retained about the baseline measure (Kong, McCullagh, Meng, Nicolae, and Tan 2003). The baseline measure is estimated as a discrete measure by maximum likelihood, and then integrals of interest are estimated as finite sums by substituting the estimated measure. The importance sampling estimator (1) can be derived as the maximum likelihood estimator (MLE) for the foregoing setting.

There appears to be only one estimating equation under importance sampling where a single sampler is involved and no additional analytical information is used. But the situation becomes more complicated if we apply multiple samplers and/or allow extra analytical information. Generally, there is a class of estimating equations to choose from. Whether the likelihood approach achieves the lowest asymptotic variance possible by using estimating equations given the same amount of information remains an important question. We investigate two situations, denoted by (I) and (II).

(I) For any function $\alpha(x)$ such that the integral $\int \alpha(x) \times q_1(x) q_2(x)\, d\mu_0$ is finite and nonzero, we have the identity

$$\frac{Z_2}{Z_1} = \frac{E_1[\alpha(x) q_2(x)]}{E_2[\alpha(x) q_1(x)]}. \tag{2}$$

Given draws $x_{11}, \dots, x_{1n_1}$ from P_1 and $x_{21}, \dots, x_{2n_2}$ from P_2, the ratio Z_2/Z_1 can be estimated by

$$\frac{n_1^{-1}\sum_{i=1}^{n_1} \alpha(x_{1i}) q_2(x_{1i})}{n_2^{-1}\sum_{i=1}^{n_2} \alpha(x_{2i}) q_1(x_{2i})},$$

which was termed "bridge sampling" by Meng and Wong (1996). Note that multiplying $\alpha(x)$ by a constant gives rise to

Zhiqiang Tan is Assistant Professor, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *ztan@jhsph.edu*). An earlier version of this work was part of the author's doctoral thesis at the University of Chicago. The author thanks Peter McCullagh and Xiao-Li Meng for advice and support throughout the process, and also thanks Tom Louis, and the editor, the associate editor, and two referees for helpful comments.

© 2004 American Statistical Association
Journal of the American Statistical Association
December 2004, Vol. 99, No. 468, Theory and Methods
DOI 10.1198/016214504000001664

the same estimator. An optimal choice of $\alpha(x)$ minimizing the asymptotic variance is

$$\frac{n_1 Z_1^{-1}}{n_1 Z_1^{-1} q_1(x) + n_2 Z_2^{-1} q_2(x)},$$

which depends on the unknown ratio Z_2/Z_1. The iterative bridge sampling estimator, defined as the unique limit of the sequence

$$\widehat{Z_2/Z_1}^{(t+1)} = \frac{n_1^{-1}\sum_{i=1}^{n_1} q_2(x_{1i})/[n_1 \widehat{Z_2/Z_1}^{(t)} q_1(x_{1i}) + n_2 q_2(x_{1i})]}{n_2^{-1}\sum_{i=1}^{n_2} q_1(x_{2i})/[n_1 \widehat{Z_2/Z_1}^{(t)} q_1(x_{2i}) + n_2 q_2(x_{2i})]}, \quad t = 0, 1, \ldots, \qquad (3)$$

with a positive starting value $\widehat{Z_2/Z_1}^{(0)}$, achieves the minimum asymptotic variance (Meng and Wong 1996) and is in fact identical to the MLE under the full model of Kong et al. (2003). We generalize these results to the situation where more than two samplers are involved (Sec. 2).

(II) Let $g_1(x), \ldots, g_l(x)$ be real-valued functions whose integrals are known with respect to μ_0. Without loss of generality, let these integrals be 0. For an arbitrary vector $\mathbf{b} = (b_1, \ldots, b_l)^\top$, we have the identity

$$\frac{Z}{Z_1} = E_1\left[\frac{q(x) - \mathbf{b}^\top \mathbf{g}(x)}{q_1(x)}\right],$$

where $\mathbf{g} = (g_1, \ldots, g_l)^\top$. Given draws $x_1, \ldots, x_n$ from P_1, the ratio Z/Z_1 can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i) - \mathbf{b}^\top \mathbf{g}(x_i)}{q_1(x_i)};$$

this is referred to as the method of control variates. The optimal choice of $\mathbf{b}$ minimizing the variance is

$$\boldsymbol{\beta} = \mathrm{var}_1^{-1}\left[\frac{\mathbf{g}}{q_1}\right]\mathrm{cov}_1^\top\left[\frac{q}{q_1}, \frac{\mathbf{g}}{q_1}\right],$$

where var_1 and cov_1 denote variance and covariance under P_1. The minimum variance is achieved asymptotically by the regression estimator (Cochran 1977; Hammersley and Handscomb 1964)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i) - \tilde{\boldsymbol{\beta}}^\top \mathbf{g}(x_i)}{q_1(x_i)}, \qquad (4)$$

where $\boldsymbol{\beta}$ is estimated by

$$\tilde{\boldsymbol{\beta}} = \widetilde{\mathrm{var}}_1^{-1}\left[\frac{\mathbf{g}}{q_1}\right]\widetilde{\mathrm{cov}}_1^\top\left[\frac{q}{q_1}, \frac{\mathbf{g}}{q_1}\right],$$

and $\widetilde{\mathrm{var}}_1$ and $\widetilde{\mathrm{cov}}_1$ denote sample variance and covariance. We show that the regression estimator is a first-order approximation to the constrained MLE under the linear submodel of Kong et al. (2003), and then generalize these results to the situation where more than one sampler is involved and control variates are used (Sec. 3).

2. FULL MODEL

Consider the setting where estimating the normalizing constants is part of the inferential problem even though such estimation is not necessary for simulation. A practical motivation is that Markov chain Monte Carlo (MCMC) algorithms can be applied to simulate approximate observations from a distribution without requiring the value of its normalizing constant. We consider the setting where the values of the normalizing constants are known in Section 3, but refer to Kong et al. (2003) and Tan (2003a,b) for Markov chain schemes.

In the likelihood approach, we take the functions $q_j(x)$ as given and consider a model for simulated observations (Kong et al. 2003). Specifically, the model assumes that $x_{j1}, \ldots, x_{jn_j}$ are independent and identically distributed as

$$q_j(\cdot)\,d\mu \Big/ \int q_j(x)\,d\mu,$$

where μ is a nonnegative measure on $\mathcal{X}$ such that $\int q_j(x)\,d\mu$ is finite and positive for $1 \le j \le m$. In the language of statistical inference, μ is a parameter and μ_0 is the value by which the data were generated. The parameter space consists of essentially all nonnegative measures on $\mathcal{X}$, and in this sense the model is full. The model is formally equivalent to Vardi's (1985) biased sampling model, except the parameter space is not restricted to probability measures.

Because of independence, the likelihood at μ is the product

$$\prod_{j=1}^{m}\prod_{i=1}^{n_j}\left[q_j(x_{ji})\mu(\{x_{ji}\}) \Big/ \int q_j(x)\,d\mu\right]. \qquad (5)$$

Let $\hat{P}$ be the empirical distribution placing mass n^{-1} at each of the points $x_1, \ldots, x_n$. Under Vardi's (1985) conditions, there exists a unique MLE $\hat{\mu}$ up to a positive multiple. The measure $\hat{\mu}$ is supported on the points $x_1, \ldots, x_n$ and has mass

$$\hat{\mu}(\{x\}) = \frac{n\hat{P}(\{x\})}{\sum_{j=1}^{m} n_j \hat{Z}_j^{-1} q_j(x)},$$

where $\hat{Z}_j$ is the MLE of Z_j and satisfies

$$\hat{Z}_j = \int q_j(x)\,d\hat{\mu} = \sum_{i=1}^{n}\frac{q_j(x_i)}{\sum_{k=1}^{m} n_k \hat{Z}_k^{-1} q_k(x_i)}. \qquad (6)$$

Consequently, the integral Z is estimated up to the same positive multiple by

$$\hat{Z} = \int q(x)\,d\hat{\mu} = \sum_{i=1}^{n}\frac{q(x_i)}{\sum_{k=1}^{m} n_k \hat{Z}_k^{-1} q_k(x_i)}. \qquad (7)$$

Here we use $\hat{\mu}$ rather than μ_0 for computational purposes, even though the true value μ_0 is known. For definiteness, let Z_1 be the reference value. Then we solve (6) with $j = 2, \ldots, m$ for the ratios $(\widehat{Z_2/Z_1}, \ldots, \widehat{Z_m/Z_1})$, and substitute these values in (7) to obtain the ratio $\widehat{Z/Z_1}$. Previously, the point estimators (6) and (7) were suggested by Geyer (1994) and Meng and Wong (1996), using entirely different arguments.

A large sample theory can be established in a similar manner as was done by Gill, Vardi, and Wellner (1988). Consider the graph on the vertices $1, \ldots, m$ such that h and j are connected by an edge if and only if $\mu_0(\{x : q_h(x) > 0\} \cap$

$\{x : q_j(x) > 0\}) > 0$. Assume that every pair of vertices is connected by a path and that $q(x)/q_*(x)$ has finite variance under P_*, where $q_*(x) = n^{-1} \sum_{j=1}^m n_j (Z_j/Z_1)^{-1} q_j(x)$. In the Appendix we show that the asymptotic variance matrix of $\widehat{\mathbf{Z}}_{(1)} = (\widehat{Z_2/Z_1}, \ldots, \widehat{Z_m/Z_1}, \widehat{Z/Z_1})^\top$ is

$$\left(\mathbf{O}_{(1)}^{-1} - \mathbf{\Lambda}_{(1)}\right)^{-1}/n - \mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^\top/n_1, \tag{8}$$

where $\mathbf{Z}_{(1)} = (Z_2/Z_1, \ldots, Z_m/Z_1, Z/Z_1)^\top$, and $\mathbf{\Lambda}_{(1)}$ and $\mathbf{O}_{(1)}$ are defined by the partitions

$$\mathbf{\Lambda} = \begin{pmatrix} n_1/n & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{(1)} \end{pmatrix} \quad \text{and} \quad \mathbf{O} = \begin{pmatrix} o_{11} & \mathbf{o}_1^\top \\ \mathbf{o}_1 & \mathbf{O}_{(1)} \end{pmatrix}.$$

Here $\mathbf{\Lambda}$ is the diagonal matrix with $(n_1/n, n_2/n(Z_2/Z_1)^{-2}, \ldots, n_m/n(Z_m/Z_1)^{-2}, 0)$ on the diagonal and $\mathbf{O}$ is the matrix $(o_{hj})_{1 \le h,j \le m+1}$, where $q_{m+1}(x) = q(x)$ and

$$o_{hj} = \int \frac{q_h(x) q_j(x)}{[n^{-1} \sum_{j=1}^m n_j (Z_k/Z_1)^{-1} q_k(x)]^2} \, dP_*.$$

Note that (8) includes both the ratios $(\widehat{Z_2/Z_1}, \ldots, \widehat{Z_m/Z_1})$ and the general ratio $\widehat{Z/Z_1}$, and that it simplifies the corresponding formula of Gill et al. (1988). The asymptotic variance matrix of $(\widehat{Z_2/Z_1}, \ldots, \widehat{Z_m/Z_1})$ also has the form of (8) but with $\mathbf{Z}_{(1)}$ replaced by $(Z_2/Z_1, \ldots, Z_m/Z_1)^\top$ and $\mathbf{\Lambda}_{(1)}$ and $\mathbf{O}_{(1)}$ replaced by their leading principal submatrices of order $m-1$.

2.1 Bridge Sampling

As an alternative, Meng and Wong's (1996) bridge sampling provides a class of estimating equations described in Section 1(I) for $m = 2$. The iterative bridge sampling estimator (3) solves the fixed-point equation

$$\frac{\hat{Z}_2}{\hat{Z}_1} = \frac{n_1^{-1} \sum_{i=1}^{n_1} q_2(x_{1i})/[n_1 \hat{Z}_1^{-1} q_1(x_{1i}) + n_2 \hat{Z}_2^{-1} q_2(x_{1i})]}{n_2^{-1} \sum_{i=1}^{n_2} q_1(x_{2i})/[n_1 \hat{Z}_1^{-1} q_1(x_{2i}) + n_2 \hat{Z}_2^{-1} q_2(x_{2i})]},$$

which is in fact equivalent to the likelihood equation (6) for $m = 2$. In this case, the likelihood approach is successful in identifying the optimal bridge sampling estimator in an automatic manner.

The basic identity (2) can be used to construct a variety of estimators if more than two normalizing constants are estimated using draws from the corresponding distributions ($m > 2$). For the simple case $m = 3$, there are at least three ways to estimate the ratios $(Z_2/Z_1, Z_3/Z_1)$:

1. Estimate Z_2/Z_1 using draws from P_1 and P_2, and estimate Z_3/Z_1 using draws from P_1 and P_3.
2. Estimate Z_2/Z_1 using draws from P_1 and P_2, estimate Z_3/Z_2 using draws from P_2 and P_3, and estimate Z_3/Z_1 as $(Z_2/Z_1)(Z_3/Z_2)$.
3. Estimate Z_3/Z_1 using draws from P_1 and P_3, estimate Z_2/Z_3 using draws from P_2 and P_3, and estimate Z_2/Z_1 as $(Z_3/Z_1)(Z_2/Z_3)$.

The choice appears to be problem-specific among the three estimators. For example, if P_1, P_2, and P_3 are normal with mean 0, 1, and 2 and unit variance, then the second estimator is best. In fact, when the optimal estimator (3) is used for single ratios, the relative standard errors of estimators 1–3 are (.101, .221), (.101, .175), and (.195, .221) ($n_1 = n_2 = n_3 = 50$). Moreover, these estimators are special cases of the construction

$$\begin{pmatrix} E_2[\alpha_{23} q_3 + \alpha_{21} q_1] & -E_3[\alpha_{23} q_2] \\ -E_2[\alpha_{32} q_3] & E_3[\alpha_{32} q_2 + \alpha_{31} q_1] \end{pmatrix} \begin{pmatrix} Z_2/Z_1 \\ Z_3/Z_1 \end{pmatrix} = \begin{pmatrix} E_1[\alpha_{21} q_2] \\ E_1[\alpha_{31} q_3] \end{pmatrix},$$

where $\alpha_{21}(x)$, $\alpha_{23}(x)$, $\alpha_{31}(x)$, and $\alpha_{32}(x)$ are real-valued functions, by taking

(a) $\alpha_{23}(x) = \alpha_{32}(x) \equiv 0$,
(b) $\alpha_{23}(x) = \alpha_{31}(x) \equiv 0$,
(c) $\alpha_{32}(x) = \alpha_{21}(x) \equiv 0$.

Generally, let $\alpha_{hj}(x)$ be $(m-1)^2$ real-valued functions for $h \ne j$, $2 \le h \le m$, $1 \le j \le m$. The basic identity (2) implies that $\mathbf{B}(Z_2/Z_1, \ldots, Z_m/Z_1)^\top = \mathbf{b}$, where

$$\mathbf{B} = \begin{pmatrix} b_{22} & -b_{23} & \cdots & -b_{2m} \\ -b_{32} & b_{33} & \cdots & -b_{3m} \\ \vdots & \vdots & \ddots & \vdots \\ -b_{m2} & -b_{m3} & \cdots & b_{mm} \end{pmatrix}$$

and

$$\mathbf{b} = \begin{pmatrix} b_{21} \\ b_{31} \\ \vdots \\ b_{m1} \end{pmatrix},$$

with

$$\begin{cases} b_{hh} = \sum_{j=1, j \ne h}^{m} E_h[\alpha_{hj} q_j], & 2 \le h \le m, \\ b_{hj} = E_j[\alpha_{hj} q_h], & h \ne j, 1 \le j \le m. \end{cases}$$

Then an extended bridge sampling estimator is $\tilde{\mathbf{B}}^{-1}\tilde{\mathbf{b}}$, where $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{b}}$ are sample counterparts of $\mathbf{B}$ and $\mathbf{b}$ with the jth sample average $\tilde{E}_j$ in place of E_j (Meng and Wong 1996). It is interesting to ask whether the MLE (6) is asymptotically efficient among extended bridge sampling estimators, regardless of problem-specific details. We give a positive answer in Theorem 1, which is proved in the Appendix. For the earlier example, the relative standard errors of $(\widehat{Z_2/Z_1}, \widehat{Z_3/Z_1})$ are (.093, .168) and are smaller than the corresponding ones of estimators 1–3.

Theorem 1. Assume that $\mathbf{B}$ is nonsingular and that $\mathrm{var}_h[\alpha_{hj} q_j]$ and $\mathrm{var}_j[\alpha_{hj} q_h]$ are finite for $h \ne j$, $2 \le h \le m$, $1 \le j \le m$. Then the bridge sampling estimator $\tilde{\mathbf{B}}^{-1}\tilde{\mathbf{b}}$ is consistent and asymptotically normal. The asymptotic variance matrix has a minimum (in the order on positive definite matrices) at

$$\alpha_{hj}(x) = \frac{n_j Z_j^{-1}}{\sum_{k=1}^m n_k Z_k^{-1} q_k(x)}.$$

The MLE (6) achieves the minimum asymptotic variance.

In the Appendix we also prove that the MLE $\widehat{Z/Z_1}$ from (7) has no greater asymptotic variance than not only the estimator

$$\sum_{i=1}^{n} \frac{q(x_i)}{\sum_{k=1}^m n_k (\widetilde{Z_k/Z_1})^{-1} q_k(x_i)}, \tag{9}$$

but also any estimator of the form

$$\sum_{j=1}^{m} \frac{\widetilde{Z_j/Z_1}}{n_j} \sum_{i=1}^{n_j} \lambda_j(x_{ji}) \frac{q(x_{ji})}{q_j(x_{ji})}, \tag{10}$$

where $(\widetilde{Z_2/Z_1}, \ldots, \widetilde{Z_m/Z_1})$ is a bridge sampling estimator, and $\lambda_1(x), \ldots, \lambda_m(x)$ are real-valued functions such that $\lambda_j(x) = 0$ if $q_j(x) = 0$ and $\sum_{j=1}^m \lambda_j(x) \equiv 1$ on $\mathcal{X}$. These optimality results lend strong support to the appropriateness of Kong et al.'s formulation. The MLEs (6) and (7) use draws from multiple distributions in an efficient manner, and so we do not need to worry about choices such as estimators 1–3.

3. LINEAR SUBMODEL

First, consider the setting in Section 1(II). Recall that μ is a point in the parameter space and μ_0 is the true value. Because the integral of $g_j(x)$ is 0 with respect to μ_0 for $1 \le j \le l$, or, equivalently, μ_0 satisfies

$$\int g_j(x)\, d\mu_0 = 0,$$

we constrain the parameter space to those measures μ satisfying the similar equation

$$\int g_j(x)\, d\mu = 0.$$

The submodel with this reduced parameter space is called a linear submodel (Kong et al. 2003). The baseline measure is then estimated by maximum likelihood subject to the linear constraints (Thm. 2). The effect of variance reduction is such that the resulting estimator (14) has zero variance if $q(x)$ is a linear combination of $g_1(x), \ldots, g_l(x)$ and $q_1(x)$ with arbitrary combination coefficients. We show that the classical regression estimator (4) is a first-order approximation to the likelihood estimator (14) in Theorem 3.

Next, consider the setting where the values of the normalizing constants $Z_1, \ldots, Z_m$ are known for multiple distributions $P_1, \ldots, P_m$. By rescaling, assume that these values are all equal. Accordingly, we consider the submodel in which $\int q_j(x)\, d\mu$ are equal for $1 \le j \le m$, and solve the corresponding maximum likelihood problem (Thm. 4). A substantial variance reduction can be achieved if $q(x)$ is matched sufficiently well by some linear combination of $q_1(x), \ldots, q_m(x)$. Theorem 6 implies that the resulting estimator (16) is more efficient than not only Hesterberg's (1995) stratified importance sampling estimator

$$\sum_{i=1}^{n} \frac{q(x_i)}{\sum_{j=1}^m n_j q_j(x_i)}, \tag{11}$$

but also Veach and Guibas's (1995) multiple importance sampling estimator

$$\sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} \lambda_j(x_{ji}) \frac{q(x_{ji})}{q_j(x_{ji})}, \tag{12}$$

where $\lambda_1(x), \ldots, \lambda_m(x)$ are real-valued functions such that $\lambda_j(x) = 0$ if $q_j(x) = 0$ and $\sum_{j=1} \lambda_j(x) \equiv 1$ on $\mathcal{X}$. By the method of control variates, Owen and Zhou (2000) derived the regression estimator (17) under unstratified sampling, where the mixture proportions are random, and extended it directly to current stratified sampling. They raised the question of whether an improved regression estimator exists due to stratification. We show that the regression estimator (17) is a first-order approximation to the likelihood estimator (16) in Theorem 5, and give a negative answer to their question in Theorem 6(a).

3.1 Illustration

Before developing the main results, we illustrate different designs and estimators by the following example. The state space is $\mathcal{R}^{10}$, and the baseline measure is Lebesgue measure. The integrand is

$$q(x) = .8 \prod_{j=1}^{10} \phi(x^j) + .2 \prod_{j=1}^{10} \psi(x^j; 4),$$

where $\phi(\cdot)$ is the standard normal density and $\psi(\cdot; 4)$ is the t density with 4 degrees of freedom. Let

$$q_1(x) = \prod_{j=1}^{10} \psi(x^j; 1)$$

and

$$q_2(x) = \prod_{j=1}^{10} \phi(x^j),$$

so that P_1 is a product of univariate Cauchy distributions and P_2 is a product of univariate normal distributions. The importance sampling estimator using the design density $q_2(x)$ has infinite variance, even though $q_2(x)$ is nearly proportional to $q(x)$ in the center. As a remedy, we consider

(a) design density $q_1(x)$ with n draws, or

(b) two design densities $q_1(x)$ and $q_2(x)$, each with $n/2$ draws.

The fact that the integral of $g_1(x) = q_2(x) - q_1(x)$ is 0 can be used for variance reduction. The results are summarized in Table 1. The likelihood estimator has mean squared error (MSE) reduced by a factor of $(.162/.00931)^2 \approx 303$ compared with the importance sampling (IS) estimator under the design (a), and by a factor of $(.0175/.00881)^2 \approx 4$ compared with the stratified IS estimator under the design (b). The regression estimator yields similar MSE as the likelihood estimator under each design. The two-sampler design leads to more accurate estimates than the one-sampler design.

3.2 Importance Sampling

For importance sampling ($m = 1$), all observations $x_1, \ldots, x_n$ are simulated from P_1. Let $g_1(x), \ldots, g_l(x)$ be real-valued functions whose integrals are 0 with respect to μ_0. We consider the submodel with parameter space

$$\left\{\mu : \int g_j(x)\, d\mu = 0 \text{ for } 1 \le j \le l\right\}.$$

The likelihood (5) becomes

$$\prod_{i=1}^{n} \left[q_1(x_i)\mu(\{x_i\}) \Big/ \int q_1(x)\, d\mu \right].$$

Unlike for the full model, the measure maximizing the likelihood for the submodel may not exist, or it may place mass

Table 1. Comparison of Designs and Estimators

	One-sampler design			Two-sampler design		
	IS	Regression	Likelihood	IS	Regression	Likelihood
Sqrt MSE	.162	.00942	.00931	.0175	.00881	.00881
Std Err	.162	.00919	.00920	.0174	.00885	.00884

NOTE: Sqrt MSE is $\sqrt{\text{mean squared error}}$ of the point estimates and Std Err is $\sqrt{\text{mean of the variance estimates}}$ from 10,000 repeated simulations of size $n = 500$.

outside the sample. The Appendix provides two examples to illustrate such possibilities. Alternatively, we maximize the likelihood with restriction to measures supported on the sample. Owen (2001, sec. 2.4) presented a related argument in the construction of the profile empirical likelihood.

Specifically, let $w_i = \mu(\{x_i\})/\int q_1(x)\,d\mu$, and restrict our attention to measures μ placing zero mass outside the sample and belonging to the reduced parameter space. The constrained maximum likelihood problem becomes

$$\max \sum_{i=1}^{n} \log w_i \tag{13}$$

for $(w_1, \dots, w_n)$ in the constraint set A_n such that $w_i \ge 0$ for $1 \le i \le n$, $\sum_{i=1}^{n} w_i q_1(x_i) = 1$, and $\sum_{i=1}^{n} w_i g_j(x_i) = 0$ for $1 \le j \le l$. Recall that $\mathbf{g}$ is the column vector $(g_1, \dots, g_l)^\top$. Theorem 2 says that the constrained maximum likelihood problem (13) can be solved by maximizing the concave function

$$\ell_n(\boldsymbol{\zeta}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[q_1(x_i) + \boldsymbol{\zeta}^\top \mathbf{g}(x_i)\right]$$

on the set Ξ_n such that $q_1(x_i) + \boldsymbol{\zeta}^\top \mathbf{g}(x_i)$ is positive for $1 \le i \le n$. Compared with a Lagrange multiplier argument, this result is more complete in providing a necessary and sufficient condition for solving the problem (13). A proof is given in the Appendix.

Theorem 2. Assume that $q_1(x_1), \dots, q_1(x_n)$ are positive and that the matrix with columns $\mathbf{g}(x_1), \dots, \mathbf{g}(x_n)$ has rank l. Then the following statements are equivalent:

(a) The set Ξ_n is bounded.

(b) The function ℓ_n has a maximum on Ξ_n.

(c) The constraint set A_n contains at least one point such that $w_i > 0$ for all i.

(d) The problem (13) has a solution such that $w_i > 0$ for all i.

If $\hat{\boldsymbol{\zeta}}$ is a maximizer of ℓ_n on Ξ_n, then the constrained MLE is

$$\hat{\mu}(\{x\}) \propto \frac{\hat{P}(\{x\})}{q_1(x) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x)}.$$

The foregoing computational recipe has an interesting interpretation. For example, if $g_j(x) = q_{j+1}(x) - q_1(x)$, where $q_{j+1}(x)$ is a nonnegative function whose integral equals Z_1 with respect to μ_0 for $1 \le j \le l$, consider the mixture model with components $q_1(x), q_2(x), \dots, q_{l+1}(x)$. Then ℓ_n is the log-likelihood function of the data $x_1, \dots, x_n$, and $q_1(x) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x)$ is the estimated density by maximum likelihood. It is not necessary that the mixture coefficients lie between 0 and 1, as long as $q_1(x_i) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x_i)$ is positive for all $1 \le i \le n$.

After the baseline measure is estimated, the ratio Z/Z_1 is estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{q(x_i)}{q_1(x_i) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x_i)}. \tag{14}$$

In comparison with the estimator (1), the design density $q_1(x)$ is adjusted to be $q_1(x) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x)$, estimated from the data. The submodel estimator has zero variance if $q(x)$ is a linear combination of $g_1(x), \dots, g_l(x)$ and $q_1(x)$, because

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g_j(x_i)}{q_1(x_i) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x_i)} = 0$$

by the fact that $\hat{\mu}$ belongs to the reduced parameter space. We give the large sample properties in Theorem 3, which is proved in the Appendix. Although this result is not the most general one, it is sufficient for many importance sampling applications where the design density $q_1(x)$ dominates all of the functions $g_j(x)$ on $\mathcal{X}$.

Theorem 3. Assume that $g_1(x), \dots, g_l(x)$ are linearly independent, $g_j(x)/q_1(x)$ is bounded on $\mathcal{X}$ for $1 \le j \le l$, and $q(x)/q_1(x)$ has finite variance under P_1. Then the estimator (14) is consistent and asymptotically normal with variance

$$n^{-1}\left\{\mathrm{var}_1\left[\frac{q}{q_1}\right] - \mathrm{cov}_1\left[\frac{q}{q_1}, \frac{\mathbf{g}}{q_1}\right]\mathrm{var}_1^{-1}\left[\frac{\mathbf{g}}{q_1}\right]\mathrm{cov}_1^{\top}\left[\frac{q}{q_1}, \frac{\mathbf{g}}{q_1}\right]\right\}$$
$$= n^{-1}\,\mathrm{var}_1\left[\frac{q(x) - \boldsymbol{\beta}^\top \mathbf{g}(x)}{q_1(x)}\right].$$

The difference between the regression estimator (4) and the likelihood estimator (14) is $o_p(n^{-1/2})$.

Glynn and Szechtman (2000) also noted that the regression estimator is equivalent to the constrained MLE to first order, and gave a proof under the weaker condition that $g_j(x)/q_1(x)$ has finite fourth moment under P_1 for $1 \le j \le l$. Although the basic ideas are similar, they are interested in estimating expected values on a probability space. In comparison, our work is motivated by estimating integrals with respect to a baseline measure, say counting measure or Lebesgue measure. We now generalize our development to multiple samplers.

3.3 Stratified Sampling

Consider the setting where observations are simulated from multiple distributions $P_1, \dots, P_m$ and the values of the normalizing constants $Z_1, \dots, Z_m$ are known. Assume that these values are equal to Z_*, typically 1. In the previous notation,

$q_*(x)$ is the function $n^{-1}\sum_{j=1}^m n_j q_j(x)$ and P_* is the corresponding distribution $n^{-1}\sum_{j=1}^m n_j P_j$.

Instead of the full model, we consider the submodel with parameter space

$$\left\{\mu : \int q_j(x)\,d\mu \text{ are equal for } 1 \le j \le m\right\},$$

which can be rewritten as

$$\left\{\mu : \int g_j(x)\,d\mu = 0 \text{ for } 1 \le j \le m-1\right\},$$

where $g_j(x) = q_{j+1}(x) - q_1(x)$. For μ in the reduced parameter space, the likelihood (5) is proportional to

$$\prod_{i=1}^n \mu(\{x_i\}) \Big/ \int q_*(x)\,d\mu.$$

Theorem 2 can be used to find the constrained MLE. Specifically, let $w_i = \mu(\{x_i\})/\int q_*(x)\,d\mu$, and consider the problem

$$\max \sum_{i=1}^n \log w_i \tag{15}$$

for $(w_1, \ldots, w_n)$ in the constraint set A_n such that $w_i \ge 0$ for $1 \le i \le n$, $\sum_{i=1}^n w_i q_*(x_i) = 1$, and $\sum_{i=1}^n w_i g_j(x_i) = 0$ for $1 \le l \le m-1$. Define

$$\ell_n(\boldsymbol{\zeta}) = \frac{1}{n}\sum_{i=1}^n \log\big[q_*(x_i) + \boldsymbol{\zeta}^\top \mathbf{g}(x_i)\big]$$

on the set Ξ_n such that $q_*(x_i) + \boldsymbol{\zeta}^\top \mathbf{g}(x_i)$ is positive for $1 \le i \le n$.

Theorem 4. Assume that $q_*(x_1), \ldots, q_*(x_n)$ are positive and that the matrix with columns $\mathbf{g}(x_1), \ldots, \mathbf{g}(x_n)$ has rank $m-1$. Then the following statements are equivalent:

(a) The set Ξ_n is bounded.

(b) The function ℓ_n has a maximum on Ξ_n.

(c) The constraint set A_n contains at least one point such that $w_i > 0$ for all i.

(d) The problem (15) has a solution such that $w_i > 0$ for all i.

If $\hat{\boldsymbol{\zeta}}$ is a maximizer of ℓ_n on Ξ_n, then the constrained MLE is

$$\hat\mu(\{x\}) \propto \frac{\hat P(\{x\})}{q_*(x) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x)}.$$

It appears that the likelihood approach fits the mixture model with components $q_1(x), q_2(x), \ldots, q_m(x)$ to the data $x_1, \ldots, x_n$ and then uses the estimated density $q_*(x) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x)$ with coefficients $n_1/n - \sum_{j=1}^{m-1}\hat\zeta_j$, $n_2/n + \hat\zeta_2, \ldots$, and $n_m/n + \hat\zeta_{m-1}$. After the baseline measure is estimated, the ratio Z/Z_* is estimated by

$$\frac{1}{n}\sum_{i=1}^n \frac{q(x_i)}{q_*(x_i) + \hat{\boldsymbol{\zeta}}^\top \mathbf{g}(x_i)}, \tag{16}$$

which has zero variance if $q(x)$ is a linear combination of $q_1(x), \ldots, q_m(x)$. Owen and Zhou's (2000) regression estimator is

$$\frac{1}{n}\sum_{i=1}^n \frac{q(x_i) - \tilde{\boldsymbol{\beta}}^\top \mathbf{g}(x_i)}{q_*(x_i)}, \tag{17}$$

where $\tilde{\boldsymbol{\beta}} = \widetilde{\mathrm{var}}_*^{-1}[\frac{\mathbf{g}}{q_*}]\widetilde{\mathrm{cov}}_*^\top[\frac{q}{q_*}, \frac{\mathbf{g}}{q_*}]$, and $\widetilde{\mathrm{var}}_*$ and $\widetilde{\mathrm{cov}}_*$ denote pooled-sample variance and covariance under $\hat P$. Note that $g_j(x)/q_*(x)$ is automatically bounded on $\mathcal{X}$ for $1 \le j \le m-1$. The proof of the following theorem is similar to that of Theorem 3, even though here $x_1, \ldots, x_n$ are not identically distributed.

Theorem 5. Assume that $q_1(x), \ldots, q_m(x)$ are linearly independent and that $q(x)/q_*(x)$ has finite variance under P_*. Then the estimator (16) is consistent and asymptotically normal with variance

$$n^{-1}\sum_{j=1}^m \frac{n_j}{n}\mathrm{var}_j\left[\frac{q(x) - \boldsymbol{\beta}^\top \mathbf{g}(x)}{q_*(x)}\right]$$

$$= n^{-1}\,\mathrm{var}_*\left[\frac{q(x) - \boldsymbol{\beta}^\top \mathbf{g}(x)}{q_*(x)}\right], \tag{18}$$

where $\boldsymbol{\beta} = \mathrm{var}_*^{-1}[\frac{\mathbf{g}}{q_*}]\mathrm{cov}_*^\top[\frac{q}{q_*}, \frac{\mathbf{g}}{q_*}]$, and var_* and cov_* denote variance and covariance under P_*. The difference between the regression estimator (17) and the likelihood estimator (16) is $o_p(n^{-1/2})$.

It is incorrect to say that the asymptotic variance of the regression estimator (17) is smaller than (18) by invoking stratification. The equality (18) follows from the fact that the stratum means $E_j[(q(x) - \boldsymbol{\beta}^\top \mathbf{g}(x))/q_*(x)]$ are equal to each other, because

$$\int \frac{q(x) - \boldsymbol{\beta}^\top \mathbf{g}(x)}{q_*(x)}\big(q_j(x) - q_1(x)\big)\,d\mu_0 = 0.$$

The asymptotic variance can be estimated by

$$n^{-1}\sum_{j=1}^m \frac{n_j}{n}\widetilde{\mathrm{var}}_j\left[\frac{q - \tilde{\boldsymbol{\beta}}^\top \mathbf{g}}{q_*}\right],$$

where $\widetilde{\mathrm{var}}_j$ denotes jth sample variance, or by

$$n^{-1}\widetilde{\mathrm{var}}_*\left[\frac{q - \tilde{\boldsymbol{\beta}}^\top \mathbf{g}}{q_*}\right].$$

The latter variance estimate is larger unless the sample means of $(q(x) - \tilde{\boldsymbol{\beta}}^\top \mathbf{g}(x))/q_*(x)$ are equal. But such a difference is asymptotically negligible.

We conclude this section with the results that the likelihood estimator (16) or, equivalently, the regression estimator (17) achieves asymptotic efficiency among two classes of estimators constructed by different arguments. Special cases are Hesterberg's (1995) stratified importance sampling estimator and Veach and Guibas's (1995) multiple importance sampling estimator. Theorem 6(a) says, somehow surprisingly, that the optimal choice of $\mathbf{b}$ is always $\boldsymbol{\beta}$ whether the draws are identically distributed from P_* or are stratified. A proof is given in the Appendix.

Theorem 6. (a) For an arbitrary vector $\mathbf{b}$, the estimator

$$\frac{1}{n}\sum_{j=1}^m \sum_{i=1}^{n_j} \frac{q(x_{ji}) - \mathbf{b}^\top \mathbf{g}(x_{ji})}{q_*(x_{ji})}$$

is unbiased. The variance has a minimum at $\mathbf{b} = \boldsymbol{\beta}$. The likelihood estimator (16) achieves the minimum variance asymptotically.

(b) Let $\lambda_1(x), \ldots, \lambda_m(x)$ be real-valued functions and let $\mathbf{c}_1(x), \ldots, \mathbf{c}_m(x)$ be vector-valued functions such that $\lambda_j(x) = 0$ if $q_j(x) = 0$, $\sum_{j=1}^m \lambda_j(x) \equiv 1$, and $\sum_{j=1}^m \lambda_j(x) \times \mathbf{c}_j(x) \equiv \mathbf{b}$ on $\mathcal{X}$ for an arbitrary vector $\mathbf{b}$. Then the estimator

$$\sum_{j=1}^m \frac{1}{n_j} \sum_{i=1}^{n_j} \lambda_j(x_{ji}) \frac{q(x_{ji}) - \mathbf{c}_j^\top(x_{ji})\mathbf{g}(x_{ji})}{q_j(x_{ji})}$$

is unbiased. The variance has a minimum at $\lambda_j(x) = n_j q_j(x)/(nq_*(x))$ and $\mathbf{c}_j(x) \equiv \boldsymbol{\beta}$. The likelihood estimator (16) achieves the minimum variance asymptotically.

4. SUMMARY

For two different situations where independent observations are simulated from multiple distributions, we show that the likelihood approach of Kong et al. (2003) achieves the lowest asymptotic variance possible by using estimating equations for Monte Carlo integration. In the first situation, the normalizing constants of the design distributions are analytically intractable and must be estimated. In the second situation, the values of the normalizing constants are known, thereby imposing linear constraints on the baseline measure.

Our results deal with optimal estimation using available draws. There remains the design issue of choosing samplers. For importance sampling, a good sampler is such that its density is approximately proportional to $|q(x)|$ and the required simulation is fast. It is important to find a balance between these conflicting criteria by exploiting the structure of a problem in practice. Similar considerations hold when searching for multiple samplers, but it becomes possible to choose individual samplers to meet different needs. For example, a heavy-tailed sampler and a sampler that approximates the integrand in the center can be applied. These ideas require further formalization and investigation.

APPENDIX: PROOFS

Proof of Formula (8)

Let $Z_{m+1} = Z$ and recall that $q_{m+1}(x) = q(x)$. The likelihood equations (6) and (7) can be written as $\mathbf{T}(\mathbf{z}) = \mathbf{0}$, where $\mathbf{z} = (z_1, \ldots, z_m, z_{m+1})^\top$, $\mathbf{T} = (T_1, \ldots, T_m, T_{m+1})^\top$, and

$$T_j(\mathbf{z}) = -z_j + \sum_{i=1}^n \frac{q_j(x_i)}{\sum_{k=1}^m n_k z_k^{-1} q_k(x_i)}.$$

Note that $\mathbf{T}(c\mathbf{z}) = c\mathbf{T}(\mathbf{z})$ for an arbitrary constant c (>0). For definiteness, fix $z_1 = 1$. The MLE $\widehat{\mathbf{Z}}_{(1)} = (\widehat{Z_2/Z_1}, \ldots, \widehat{Z_m/Z_1}, \widehat{Z_{m+1}/Z_1})^\top$ is a solution to the m equations $\mathbf{T}_{(1)}(\mathbf{z}_{(1)}) = \mathbf{0}$, where $\mathbf{z}_{(1)} = (z_2, \ldots, z_m, z_{m+1})^\top$ and $\mathbf{T}_{(1)} = (T_2, \ldots, T_m, T_{m+1})^\top$. By similar arguments as those of Gill et al. (1988), $\widehat{\mathbf{Z}}_{(1)}$ is consistent and asymptotically normal with variance $n^{-1}\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{\top -1}$, where

$$\mathbf{G} = n \operatorname{var}[\mathbf{T}_{(1)}(\mathbf{Z}_{(1)})] = \mathbf{O}_{(1)} - \mathbf{O}_{(1)}\boldsymbol{\Lambda}_{(1)}\mathbf{O}_{(1)} - \frac{\mathbf{o}_1 n_1 \mathbf{o}_1^\top}{n},$$

$$\mathbf{H} = E\left[\frac{\partial}{\partial \mathbf{z}_{(1)}} \mathbf{T}_{(1)}(\mathbf{Z}_{(1)})\right] = \mathbf{O}_{(1)}\boldsymbol{\Lambda}_{(1)} - \mathbf{I}_m,$$

and $\mathbf{I}_m$ is the identity matrix of order m. Now it is straightforward to check that $(\mathbf{I}_{m+1} - \mathbf{O}\boldsymbol{\Lambda})(1, Z_2/Z_1, \ldots, Z_m/Z_1, Z_{m+1}/Z_1)^\top = \mathbf{0}$ and thus $(\mathbf{I}_m - \mathbf{O}_{(1)}\boldsymbol{\Lambda}_{(1)})\mathbf{Z}_{(1)} = n_1\mathbf{o}_1/n$. Using this fact twice, we first obtain

$$\mathbf{H}^{-1}\mathbf{G} = -(\mathbf{O}_{(1)} - \mathbf{Z}_{(1)}\mathbf{o}_1^\top), \tag{19}$$

and then obtain formula (8). The asymptotic variance matrix of $(\widetilde{Z_2/Z_1}, \ldots, \widetilde{Z_m/Z_1})$ is of the same form as (8), by similar calculations as above.

Proof of Theorem 1

We consider the ratios $(Z_2/Z_1, \ldots, Z_m/Z_1)$ and the general ratio Z/Z_1 simultaneously. For convenience, assume that $q(x) \ge 0$ and $Z > 0$; otherwise, we can replace $q(x)$ by the vector $[\max(q(x), 0), \max(-q(x), 0)]$ and extend the following argument to allow a vector-valued $q(x)$. By the law of large numbers, $(\widetilde{Z_2/Z_1}, \ldots, \widetilde{Z_m/Z_1})$ is consistent and so is the estimator (10). By the inequality

$$\left|\sum_{i=1}^n \frac{q(x_i)}{\sum_{k=1}^m n_k (\widetilde{Z_k/Z_1})^{-1} q_k(x_i)} - \sum_{i=1}^n \frac{q(x_i)}{\sum_{k=1}^m n_k (Z_k/Z_1)^{-1} q_k(x_i)}\right|$$
$$\le \left\{\sum_{i=1}^n \frac{|q(x_i)|}{\sum_{k=1}^m n_k (Z_k/Z_1)^{-1} q_k(x_i)}\right\} \times \left\{\sum_{j=1}^m \frac{|(\widetilde{Z_j/Z_1})^{-1} - (Z_j/Z_1)^{-1}|}{(\widetilde{Z_j/Z_1})^{-1}}\right\},$$

it follows that the estimator (9) is also consistent. For $2 \le h \le m$, let $\alpha_{h,m+1}(x) \equiv 0$. For $1 \le j \le m$, let

$$\alpha_{m+1,j}(x) = \frac{n_j z_j^{-1}}{\sum_{k=1}^m n_k z_k^{-1} q_k(x)} \qquad \text{or} \qquad \frac{\lambda_j(x)}{q_j(x)},$$

according to which estimator, (9) or (10), is treated as $\widetilde{Z_{m+1}/Z_1}$. In either case, the bridge sampling estimator $\widetilde{\mathbf{Z}}_{(1)} = (\widetilde{Z_2/Z_1}, \ldots, \widetilde{Z_m/Z_1}, \widetilde{Z_{m+1}/Z_1})^\top$ is a solution to the estimating equation $\tilde{\mathbf{B}}(\mathbf{z}_{(1)})\mathbf{z}_{(1)} - \tilde{\mathbf{b}}(\mathbf{z}_{(1)}) = \mathbf{0}$, where $z_1 = 1$, $\mathbf{z}_{(1)} = (z_2, \ldots, z_m, z_{m+1})^\top$, and $\tilde{\mathbf{B}}(\mathbf{z}_{(1)})$ and $\tilde{\mathbf{b}}(\mathbf{z}_{(1)})$ are sample counterparts of $\mathbf{B}(\mathbf{z}_{(1)})$ and $\mathbf{b}(\mathbf{z}_{(1)})$, now defined as

$$\mathbf{B} = \begin{pmatrix} b_{22} & \cdots & -b_{2m} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ -b_{m2} & \cdots & b_{mm} & 0 \\ -b_{m+1,2} & \cdots & -b_{m+1,m} & 1 \end{pmatrix}$$

and

$$\mathbf{b} = \begin{pmatrix} b_{21} \\ \vdots \\ b_{m1} \\ b_{m+1,1} \end{pmatrix},$$

with

$$\begin{cases} b_{hh} = \sum_{j=1, j\ne h}^{m+1} E_h[\alpha_{hj} q_j], & 2 \le h \le m+1, \\ b_{hj} = E_j[\alpha_{hj} q_h], & h \ne j, 1 \le j \le m+1. \end{cases}$$

In the case where the estimator (9) is treated, the derivative matrix

$$\frac{\partial}{\partial \mathbf{z}_{(1)}}[\tilde{\mathbf{B}}(\mathbf{z}_{(1)})\mathbf{z}_{(1)} - \tilde{\mathbf{b}}(\mathbf{z}_{(1)})]$$
$$= \tilde{\mathbf{B}}(\mathbf{z}_{(1)}) - \begin{pmatrix} \mathbf{0}_{(m-1)\times m} \\ \sum_{j=1}^m \tilde{E}_j[\frac{\partial \alpha_{m+1,j}}{\partial \mathbf{z}_{(1)}} q_{m+1}] z_j \end{pmatrix}.$$

is such that its supremum norm for $\mathbf{z}_{(1)}$ in a neighborhood of $\mathbf{Z}_{(1)}$ is square integrable by the assumption that $o_{m+1,m+1}$ is finite, and its expectation at $\mathbf{z}_{(1)} = \mathbf{Z}_{(1)}$ equals $\mathbf{B}(\mathbf{Z}_{(1)})$ because the extra term vanishes by (2) and $\sum_{j=1}^{m} \alpha_{m+1,j}(x) q_j(x) \equiv 1$:

$$\sum_{j=1}^{m} E_j\left[\frac{\partial \alpha_{m+1,j}}{\partial \mathbf{z}_{(1)}} q_{m+1}\right]\frac{Z_j}{Z_1} = \sum_{j=1}^{m} E_{m+1}\left[\frac{\partial \alpha_{m+1,j}}{\partial \mathbf{z}_{(1)}} q_j\right]\frac{Z_{m+1}}{Z_1} = 0.$$

By the asymptotic theory of M-estimators (van der Vaart 1998), $\widetilde{\mathbf{Z}}_{(1)}$ is asymptotically normal with variance $\mathbf{B}^{-1}\operatorname{var}[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}]\mathbf{B}^{\top -1}$, where $\mathbf{B} = \mathbf{B}(\mathbf{Z}_{(1)})$, $\tilde{\mathbf{B}} = \tilde{\mathbf{B}}(\mathbf{Z}_{(1)})$, and $\tilde{\mathbf{b}} = \tilde{\mathbf{b}}(\mathbf{Z}_{(1)})$. In the case where the estimator (10) is treated, $\tilde{\mathbf{B}}(\mathbf{z}_{(1)})$ and $\tilde{\mathbf{b}}(\mathbf{z}_{(1)})$ are in fact free of $\mathbf{z}_{(1)}$, and this result follows trivially. Applying the matrix version of the Cauchy–Schwartz inequality

$$\operatorname{var}(\mathbf{X}) \geq \operatorname{cov}(\mathbf{X}, \mathbf{Y}) \operatorname{var}^{-1}(\mathbf{Y}) \operatorname{cov}^{\top}(\mathbf{X}, \mathbf{Y})$$

with $\mathbf{X} = \tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}$ and $\mathbf{Y} = \mathbf{T}_{(1)}(\mathbf{Z}_{(1)})$, we obtain

$$\begin{aligned}\operatorname{var}\big[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}\big] \geq \operatorname{cov}\big[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}, \mathbf{T}_{(1)}(\mathbf{Z}_{(1)})\big] \times n\mathbf{G}^{-1} \\ \times \operatorname{cov}^{\top}\big[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}, \mathbf{T}_{(1)}(\mathbf{Z}_{(1)})\big].\end{aligned}$$

Now for $2 \leq h$, $j \leq m + 1$, the $(h - 1, j - 1)$th element of $n\operatorname{cov}[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}, \mathbf{T}_{(1)}(\mathbf{Z}_{(1)})]$ is

$$\begin{aligned}&\sum_{\nu=1,\nu\neq h}^{m} \operatorname{cov}_h\left[\alpha_{h\nu} q_\nu \frac{Z_h}{Z_1}, \frac{n q_j/Z_1}{\sum_{k=1}^{m} n_k q_k/Z_k}\right] \\ &\quad - \sum_{\nu=1,\nu\neq h}^{m} \operatorname{cov}_\nu\left[\alpha_{h\nu} q_h \frac{Z_\nu}{Z_1}, \frac{n q_j/Z_1}{\sum_{k=1}^{m} n_k q_k/Z_k}\right] \\ &= \sum_{\nu=1,\nu\neq h}^{m} E_h[\alpha_{h\nu} q_\nu]\frac{Z_h}{Z_1} E_h\left[\frac{n q_j/Z_1}{\sum_{k=1}^{m} n_k q_k/Z_k}\right] \\ &\quad - \sum_{\nu=1,\nu\neq h}^{m} E_\nu[\alpha_{h\nu} q_h]\frac{Z_\nu}{Z_1} E_\nu\left[\frac{n q_j/Z_1}{\sum_{k=1}^{m} n_k q_k/Z_k}\right] \\ &= \sum_{\nu=1,\nu\neq h}^{m} E_h[\alpha_{h\nu} q_\nu] o_{hj} - \sum_{\nu=1,\nu\neq h}^{m} E_\nu[\alpha_{h\nu} q_h] o_{\nu j},\end{aligned}$$

and the $(h - 1, j - 1)$th element of $\mathbf{B}(\mathbf{H}^{-1}\mathbf{G})$, due to (19), is

$$\begin{aligned}&-\sum_{\nu=1,\nu\neq h}^{m} E_h[\alpha_{h\nu} q_\nu]\left(o_{hj} - \frac{Z_h}{Z_1} o_{j1}\right) \\ &\quad + \sum_{\nu=2,\nu\neq h}^{m} E_\nu[\alpha_{h\nu} q_h]\left(o_{\nu j} - \frac{Z_\nu}{Z_1} o_{j1}\right) \\ &= -E_h[\alpha_{h1} q_1]\left(o_{hj} - \frac{Z_h}{Z_1} o_{j1}\right) - \sum_{\nu=2,\nu\neq h}^{m} E_h[\alpha_{h\nu} q_\nu] o_{hj} \\ &\quad + \sum_{\nu=2,\nu\neq h}^{m} E_\nu[\alpha_{h\nu} q_h] o_{\nu j} \\ &= -\sum_{\nu=1,\nu\neq h}^{m} E_h[\alpha_{h\nu} q_\nu] o_{hj} + \sum_{\nu=1,\nu\neq h}^{m} E_\nu[\alpha_{h\nu} q_h] o_{\nu j}.\end{aligned}$$

The corresponding elements of $\operatorname{cov}[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}, \mathbf{T}_{(1)}(\mathbf{Z}_{(1)})]$ and $-n^{-1}\mathbf{B}(\mathbf{H}^{-1}\mathbf{G})$ are equal. Thus $\mathbf{B}^{-1}\operatorname{cov}[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}, \mathbf{T}_{(1)}(\mathbf{Z}_{(1)})] = -n^{-1}\mathbf{H}^{-1}\mathbf{G}$. Consequently, we have

$$\mathbf{B}^{-1}\operatorname{var}\big[\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}}\big]\mathbf{B}^{\top -1} \geq n^{-1}\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{\top -1}.$$

The right side is exactly the asymptotic variance of $\widehat{\mathbf{Z}}_{(1)}$; see the proof of (8). The equality holds if $\alpha_{hj}(x) = n_j Z_j^{-1}/\sum_{k=1}^{m} n_k Z_k^{-1} q_k(x)$ for $h \neq j$, $2 \leq h \leq m + 1$, $1 \leq j \leq m + 1$, because then $\tilde{\mathbf{B}}\mathbf{Z}_{(1)} - \tilde{\mathbf{b}} = -\mathbf{T}_{(1)}(\mathbf{Z}_{(1)})$.

Two Examples. In the first example, there does not exist a measure that maximizes the likelihood in the reduced parameter space. In the second example, the maximizing measure places mass outside the sample.

Let the state space be the unit interval $(0, 1)$ and the baseline measure be Lebesgue measure. Let $q_1(x) \equiv 1$ and $q_2(x) = 3(x^{-1/4} - 1)$. Then the integral of $g_1(x) = q_2(x) - q_1(x)$ is 0. Suppose that the observations are x_1, x_2, and x_3, for which $q_2(x)$ equals 1/5, 1, and 6/5. For simplicity, consider only measures μ such that both $\int q_1(x)\,d\mu$ and $\int q_2(x)\,d\mu$ are 1. The log-likelihood is $\sum_{i=1}^{3} \log \mu(\{x_i\})$ up to an additive constant. For measures supported on the sample, the log-likelihood has maximum -3.74, which is achieved by the measure with mass 2/15 at x_1, 1/3 at x_2, and 8/15 at x_3. For each $0 < \delta < 1/3$, the measure with mass $1/3 - \delta$ on x_1, 1/3 on x_2, 1/3 on x_3, and δ on x_4, where $q_2(x_4) = (1 + 1/\delta)/5$, satisfies the constraint, and the log-likelihood is $-3\log 3 + \log(1 - 3\delta)$. This sequence can be arbitrarily close to the unconstrained maximum $-3\log 3$ (≈ -3.30). But the limit measure does not satisfy the constraint. Thus there does not exist an MLE for this example.

For the second example, $q_2(x)$ is changed to the beta(20, 20) density function. The observations are changed such that $q_2(x_1)$, $q_2(x_2)$, and $q_2(x_3)$ remain as 1/5, 1, and 6/5. As before, the log-likelihood, subject to the constraint $\int g_1(x)\,d\mu = 0$, has maximum -3.74 over measures supported on the sample. But the global maximum is -3.43 at the measure with mass .28 at x_1, .33 at x_2, .35 at x_3, and .04 at x_4, where $q_2(x_4)$ is the maximum q^* ($= 5.01$) of $q_2(x)$ on $(0, 1)$. This measure can be found by maximizing $\sum_{i=1}^{3} \log w_i$ over (w_1, w_2, w_3) such that w_1, w_2, and w_3 are nonnegative, $w_1/5 + w_2 + 6w_3/5 \leq 1$, and $w_1/5 + w_2 + 6w_3/5 + q^*(1 - \sum_{i=1}^{3} w_i) \leq 1$.

Proof of Theorem 2

The set Ξ_n contains a neighborhood of 0 because $q_1(x_i) > 0$ for $1 \leq i \leq n$. Further, it is an open and convex set. The function ℓ_n is twice continuously differentiable with derivatives

$$\frac{\partial \ell_n}{\partial \zeta_j} = \frac{1}{n}\sum_{i=1}^{n} \frac{g_j(x_i)}{q_1(x_i) + \boldsymbol{\zeta}^{\top}\mathbf{g}(x_i)}$$

and

$$\frac{\partial^2 \ell_n}{\partial \zeta_h\,\partial \zeta_j} = -\frac{1}{n}\sum_{i=1}^{n} \frac{g_h(x_i) g_j(x_i)}{[q_1(x_i) + \boldsymbol{\zeta}^{\top}\mathbf{g}(x_i)]^2}.$$

It follows that ℓ_n is strictly concave on Ξ_n because the matrix with columns $\mathbf{g}(x_1), \ldots, \mathbf{g}(x_n)$ has full rank l.

(a) ⇒ (b): Suppose that Ξ_n is bounded. Then ℓ_n is bounded from above on Ξ_n and approaches $-\infty$ at the boundary. By strict concavity, ℓ_n achieves a unique maximum.

(b) ⇒ (a): Suppose that ℓ_n has a maximum on Ξ_n. Then ℓ_n is bounded from above on Ξ_n. It follows that Ξ_n is bounded. Otherwise, there exists a sequence of pairs $(c_k, \boldsymbol{\zeta}_k)$, where c_k is a positive number and $\boldsymbol{\zeta}_k$ is a unit vector, such that $c_k \to \infty$ as $k \to \infty$ and $q_1(x_i) + c_k\boldsymbol{\zeta}_k^{\top}\mathbf{g}(x_i) > 0$ for $1 \leq i \leq n$. By compactness of the unit ball, there exists a unit vector $\boldsymbol{\zeta}_0$ such that $\boldsymbol{\zeta}_k \to \boldsymbol{\zeta}_0$ as $k \to \infty$. Letting $k \to \infty$ in $\boldsymbol{\zeta}_k^{\top}\mathbf{g}(x_i) > -q_1(x_i)/c_k$, we obtain $\boldsymbol{\zeta}_0^{\top}\mathbf{g}(x_i) \geq 0$ for $1 \leq i \leq n$. The inequality holds strictly for some i because $\mathbf{g}(x_1), \ldots, \mathbf{g}(x_n)$ has full rank l. Then $c\boldsymbol{\zeta}_0$ belongs to Ξ_n for each positive number c, and $\ell_n(c\boldsymbol{\zeta}_0)$ can be arbitrarily large, which is a contradiction.

(b) ⇒ (c): Suppose that ℓ_n is maximized at $\hat{\boldsymbol\zeta}$. Then the derivatives of ℓ_n are 0 at $\hat{\boldsymbol\zeta}$, because the set Ξ_n is open. From the identity

$$\frac{1}{n}\sum_{i=1}^{n}\frac{q_1(x_i)}{q_1(x_i)+\boldsymbol\zeta^\top\mathbf{g}(x_i)} = 1-\sum_{j=1}^{l}\zeta_j\frac{\partial\ell_n}{\partial\zeta_j},$$

it follows that the positive weights

$$\hat{w}_i = \frac{n^{-1}}{q_1(x_i)+\hat{\boldsymbol\zeta}^\top\mathbf{g}(x_i)}$$

are positive and satisfy the constraints that define A_n.

(c) ⇒ (b) + (d): For any $(w_1,\dots,w_n)\in A_n$ and any $\zeta\in\Xi_n$, Jensen's inequality implies that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}&\log\{w_i[q_1(x_i)+\boldsymbol\zeta^\top\mathbf{g}(x_i)]\}\\ &\le\log\left\{\frac{1}{n}\sum_{i=1}^{n}w_i[q_1(x_i)+\boldsymbol\zeta^\top\mathbf{g}(x_i)]\right\}\\ &=-\log(n),\end{aligned}$$

which can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\log w_i\le-\frac{1}{n}\sum_{i=1}^{n}\log[q_1(x_i)+\boldsymbol\zeta^\top\mathbf{g}(x_i)]-\log(n).$$

Suppose that the constraint set A_n contains at least one point such that $w_i>0$ for all i. Then ℓ_n is bounded from above on Ξ_n. By the proof of (b) ⇒ (a), Ξ_n is bounded. Thus ℓ_n achieves a unique maximum on Ξ_n. Let $\hat{\boldsymbol\zeta}$ be the maximizer of ℓ_n. Then the foregoing $(\hat{w}_1,\dots,\hat{w}_n)$ is a unique solution to the problem (13).

(d) ⇒ (c): It is trivially true.

Proof of Theorem 3

Consider the criterion function

$$\ell(\boldsymbol\zeta)=\int\log\left[1+\boldsymbol\zeta^\top\frac{\mathbf{g}(x)}{q_1(x)}\right]dP_1,$$

where the log of 0 or a negative number is taken to be $-\infty$. It is finite in at least a neighborhood of $\boldsymbol\zeta_0=\mathbf{0}$, because $g(x)/q_1(x)$ is bounded on $\mathcal{X}$. For each fixed x, $\log[1+\boldsymbol\zeta^\top\mathbf{g}(x)/q_1(x)]$ is concave in $\boldsymbol\zeta$ and so is $\ell(\boldsymbol\zeta)$. By Jensen's inequality, $\ell(\boldsymbol\zeta)\le\log E_1[1+\boldsymbol\zeta^\top\mathbf{g}(x)/q_1(x)]=\log(1)=0$, and the equality holds only at $\boldsymbol\zeta_0$ because the functions in $\mathbf{g}$ are linearly independent on $\mathcal{X}$. Thus $\ell(\boldsymbol\zeta)$ has a unique maximum at $\boldsymbol\zeta_0$. Now $\hat{\boldsymbol\zeta}$ is defined by maximizing the sample counterpart $\ell_n(\boldsymbol\zeta)$. Note that $\partial^2\ell_n/\partial\boldsymbol\zeta^2$ is uniformly bounded for $\boldsymbol\zeta$ over a neighborhood of $\boldsymbol\zeta_0$, and $-E[\partial^2\ell_n/\partial\boldsymbol\zeta^2(\boldsymbol\zeta_0)]=\mathrm{var}_1[\mathbf{g}/q_1]$. By the asymptotic theory of M-estimators from convex minimization (Niemiro 1992), $\hat{\boldsymbol\zeta}$ converges to $\boldsymbol\zeta_0$ with probability 1, and $\sqrt{n}(\hat{\boldsymbol\zeta}-\boldsymbol\zeta_0)$ has the expansion

$$\hat{\boldsymbol\zeta}-\boldsymbol\zeta_0=\mathrm{var}_1^{-1}\left[\frac{\mathbf{g}}{q_1}\right]\cdot\frac{\partial\ell_n}{\partial\boldsymbol\zeta}(\boldsymbol\zeta_0)+o_p(n^{-1/2}).$$

Then $1+\hat{\boldsymbol\zeta}^\top\mathbf{g}(x)/q_1(x)$ converges to 1 uniformly on $\mathcal{X}$, and the right side of the inequality

$$\begin{aligned}&\left|\widehat{Z/Z_1}-\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i)}{q_1(x_i)}\right|\\ &\qquad\le|\hat{\boldsymbol\zeta}|^\top\left\|\frac{\mathbf{g}(x)}{q_1(x)+\hat{\boldsymbol\zeta}^\top\mathbf{g}(x)}\right\|_{\sup}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{q(x_i)}{q_1(x_i)}\right|\right\}\end{aligned}$$

converges to 0, because $\mathbf{g}(x)/q_1(x)$ is bounded on $\mathcal{X}$. Thus $\widehat{Z/Z_1}$ converges to Z/Z_1 with probability 1. For $\hat{\boldsymbol\zeta}$ about $\boldsymbol\zeta_0$, a Taylor expansion of $\widehat{Z/Z_1}$ yields

$$\begin{aligned}\widehat{Z/Z_1}&=\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i)}{q_1(x_i)}\left(1-\hat{\boldsymbol\zeta}^\top\frac{\mathbf{g}(x_i)}{q_1(x_i)}\right)+o_p(n^{-1/2})\\ &=\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i)}{q_1(x_i)}\\ &\quad-\left(\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i)}{q_1(x_i)}\frac{\mathbf{g}^{*\top}(x_i)}{q_1(x_i)}\right)\mathrm{var}_1^{-1}\left[\frac{\mathbf{g}}{q_1}\right]\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{g}(x_i)}{q_1(x_i)}\right)\\ &\quad+o_p(n^{-1/2}).\end{aligned}$$

The remainder term in the first equation is

$$\hat{\boldsymbol\zeta}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\frac{q_1(x)\mathbf{g}(x_i)\mathbf{g}^\top(x_i)}{(q_1(x_i)+\boldsymbol\zeta^{*\top}\mathbf{g}(x_i))^3}\right)\hat{\boldsymbol\zeta}=o_p(n^{-1/2}),$$

where $\boldsymbol\zeta^*$ lies between $\boldsymbol\zeta_0$ and $\hat{\boldsymbol\zeta}$. The first-order term is a regression estimator with a slightly different regression coefficient than $\tilde{\boldsymbol\beta}$. We conclude the proof because two consistent estimators of $\boldsymbol\beta$ yield equivalent regression estimators to first order (Glynn and Szechtman 2000, thm. 1).

Proof of Theorem 6

(a) The vector that minimizes the variance is

$$\begin{aligned}\mathbf{b}=\mathrm{cov}&\left[\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{q(x_{ji})}{q_*(x_{ji})},\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{\mathbf{g}(x_{ji})}{q_*(x_{ji})}\right]\\ &\times\mathrm{var}^{-1}\left[\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{\mathbf{g}(x_{ji})}{q_*(x_{ji})}\right].\end{aligned}$$

It remains to show that this vector equals $\boldsymbol\beta$. For $2\le k\le m$, we have

$$\begin{aligned}\frac{1}{n}&\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathrm{cov}\left[\frac{q(x_{ji})}{q_*(x_{ji})},\frac{q_k(x_{ji})-q_1(x_{ji})}{q_*(x_{ji})}\right]\\ &=o_{m+1,k}-o_{m+1,1}-\frac{1}{n}\sum_{j=1}^{m}n_jo_{m+1,j}(o_{kj}-o_{1j})\\ &=o_{m+1,k}-o_{m+1,1}-\frac{1}{n}\sum_{j=2}^{m}n_j(o_{m+1,j}-o_{m+1,1})(o_{kj}-o_{1j}),\end{aligned}$$

because $n_1(o_{k1}-o_{11})+\sum_{j=2}^{m}n_j(o_{kj}-o_{1j})=0$. Writing in matrix notation, we obtain

$$\begin{aligned}\frac{1}{n}&\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathrm{cov}\left[\frac{q(x_{ji})}{q_*(x_{ji})},\frac{\mathbf{g}(x_{ji})}{q_*(x_{ji})}\right]\\ &=(\mathbf{o}_{m+1}^\top-o_{m+1,1}\mathbf{1}_{m-1}^\top)[\mathbf{I}_{m-1}-\boldsymbol\Lambda_{(1)}(\mathbf{O}_{(1)}-\mathbf{o}_1\mathbf{1}_{m-1}^\top)],\end{aligned}$$

where $\mathbf{o}_{m+1}=(o_{2,m+1},\dots,o_{m,m+1})^\top$. Similarly, we have

$$\begin{aligned}\frac{1}{n}&\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathrm{var}\left[\frac{\mathbf{g}(x_{ji})}{q_*(x_{ji})}\right]\\ &=(\mathbf{O}_{(1)}-\mathbf{1}_{m-1}\mathbf{o}_1^\top-\mathbf{o}_1\mathbf{1}_{m-1}^\top+o_{11}\mathbf{1}_{m-1}\mathbf{1}_{m-1}^\top)\\ &\quad\times[\mathbf{I}_{m-1}-\boldsymbol\Lambda_{(1)}(\mathbf{O}_{(1)}-\mathbf{o}_1\mathbf{1}_{m-1}^\top)].\end{aligned}$$

Thus the optimal regression coefficient is

$$\begin{aligned}&(\mathbf{O}_{(1)}-\mathbf{1}_{m-1}\mathbf{o}_1^\top-\mathbf{o}_1\mathbf{1}_{m-1}^\top+o_{11}\mathbf{1}_{m-1}\mathbf{1}_{m-1}^\top)^{-1}\\ &\qquad\times(\mathbf{o}_{m+1}-o_{m+1,1}\mathbf{1}_{m-1}),\end{aligned}$$

which is exactly $\boldsymbol{\beta}$.

(b) For $1 \le j \le m$, let $\eta_j = \int \lambda_j(x)(q(x) - \mathbf{c}_j^\top(x)\mathbf{g}(x))\,d\mu_0$. Then $\sum_{j=1}^m \eta_j = Z$. The estimator is unbiased and has variance

$$\begin{aligned}
&\sum_{j=1}^m \frac{1}{n_j} \int \left(\lambda_j(x) \frac{q(x) - \mathbf{c}_j^\top(x)\mathbf{g}(x)}{q_j(x)} - \frac{\eta_j}{Z_*} \right)^2 \frac{q_j(x)}{Z_j}\, d\mu_0 \\
&\quad = \frac{Z_*^{-1}}{n} \int \sum_{j=1}^m \frac{\left(\lambda_j(x)(q(x) - \mathbf{c}_j^\top(x)\mathbf{g}(x)) - \eta_j q_j(x)/Z_*\right)^2}{n_j q_j(x)/n}\, d\mu_0 \\
&\quad \ge \frac{Z_*^{-1}}{n} \int \frac{\left(\sum_{j=1}^m \lambda_j(x)(q(x) - \mathbf{c}_j^\top(x)\mathbf{g}(x)) - \eta_j q_j(x)/Z_*\right)^2}{q_*(x)}\, d\mu_0 \\
&\quad = \frac{Z_*^{-1}}{n} \int \frac{\left(q(x) - \mathbf{b}^\top\mathbf{g}(x) - \sum_{j=1}^m \eta_j q_j(x)/Z_*\right)^2}{q_*(x)}\, d\mu_0 \\
&\quad = \frac{1}{n} \int \left(\frac{q(x) - \mathbf{b}^\top\mathbf{g}(x) - \sum_{j=1}^{m-1} (\eta_{j+1} - Z n_{j+1}/n) g_j(x)/Z_*}{q_*(x)} \right. \\
&\qquad\qquad \left. - \frac{Z}{Z_*} \right)^2 \frac{q_*(x)}{Z_*}\, d\mu_0.
\end{aligned}$$

The last line is the variance of the estimator considered in (a), with $\mathbf{b}$ replaced by $\mathbf{b} + (\eta_2 - Zn_2/n, \ldots, \eta_m - Zn_m/n)^\top / Z_*$, under unstratified sampling from P_*, and thus is no smaller than that of the regression estimator. The equality holds in the foregoing Cauchy–Schwartz inequality if $\lambda_j(x) = n_j q_j(x)/(nq_*(x))$ and $\mathbf{c}_j(x) \equiv \boldsymbol{\beta}$.

[Received November 2002. Revised March 2004.]

REFERENCES

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.

Geyer, C. J. (1994), "Estimating Normalizing Constants and Reweighing Mixtures in Markov Chain Monte Carlo," technical report, University of Minnesota, School of Statistics.

Gill, R., Vardi, Y., and Wellner, J. (1988), "Large Sample Theory of Empirical Distributions in Biased Sampling Models," *The Annals of Statistics*, 16, 1069–1112.

Glynn, P. W., and Szechtman, R. (2000), "Some New Perspectives on the Method of Control Variates," in *Monte Carlo and Quasi-Monte Carlo Methods*, eds. K.-T. Fang, F. J. Hickernell, and H. Niederreiter, New York: Springer-Verlag, pp. 27–49.

Hammersley, J. M., and Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Methuen.

Hesterberg, T. (1995), "Weighted Average Importance Sampling and Defensive Mixture Distributions," *Technometrics*, 37, 185–194.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003), "A Theory of Statistical Models for Monte Carlo Integration" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 65, 585–618.

Meng, X.-L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Explanation," *Statistica Sinica*, 6, 831–860.

Niemiro, W. (1992), "Asymptotics for M-Estimators Defined by Convex Minimization," *The Annals of Statistics*, 20, 1514–1533.

Owen, A. (2001), *Empirical Likelihood*, New York: Chapman & Hall.

Owen, A., and Zhou, Y. (2000), "Safe and Effective Importance Sampling," *Journal of the American Statistical Association*, 95, 135–143.

Tan, Z. (2003a), "Monte Carlo Integration With Markov Chain," working paper, Johns Hopkins University, Department of Biostatistics.

——— (2003b), "Monte Carlo Integration With Acceptance-Rejection," working paper, Johns Hopkins University, Department of Biostatistics.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

Vardi, Y. (1985), "Empirical Distributions in Selection Bias Models," *The Annals of Statistics*, 25, 178–203.

Veach, E. and Guibas, L. (1995), "Optimally Combining Sampling Techniques for Monte Carlo Rendering," in *SIGGRAPH'95 Conference Proceedings*, Reading, MA: Addison-Wesley, pp. 419–428.