Brazilian Journal of Probability and Statistics
2011, Vol. 25, No. 3, 437–443
DOI: 10.1214/11-BJPS155

On improved estimation for importance sampling

David Firth

University of Warwick, UK

Abstract. The standard estimator used in conjunction with importance sampling in Monte Carlo integration is unbiased but inefficient. An alternative estimator is discussed, based on the idea of a difference estimator, which is asymptotically optimal. The improved estimator uses the importance weight as a control variate, as previously studied by Hesterberg (Ph.D. Dissertation, Stanford University (1988); *Technometrics* **37** (1995) 185–194; *Statistics and Computing* **6** (1996) 147–157); it is routinely available and can deliver substantial additional variance reduction. Finite-sample performance is illustrated in a sequential testing example. Connections are made with methods from the survey-sampling literature.

1 Introduction

Importance sampling is one of the most effective and commonly used techniques of variance reduction in Monte Carlo simulation, and is described in numerous texts (e.g., Evans and Swartz, 2000; Hammersley and Handscomb, 1964; Ripley, 1987; Robert and Casella, 2004). If the aim is to estimate $\theta = E_f\{\phi(X)\} = \int \phi(x) f(x)\,\mathrm{d}x$, where $f(x)$ is a density function, the method of importance sampling is to generate variates $X_1, \ldots, X_n$ from density $g(x)$ and then to estimate θ by

$$\hat{\theta}_0 = n^{-1} \sum_{i=1}^{n} \phi(X_i) f(X_i)/g(X_i).$$

This estimator is unbiased, and has variance

$$\operatorname{var}(\hat{\theta}_0) = n^{-1} \int \left\{ \phi(x) \frac{f(x)}{g(x)} - \theta \right\}^2 g(x)\,\mathrm{d}x.$$

The density g is chosen to be easily simulated from, and to be such that $\phi(x) f(x)/g(x)$ is nearly constant so that $\operatorname{var}(\hat{\theta}_0)$ is small.

In the following, g is taken to be fixed, and alternatives to $\hat{\theta}_0$ are considered for estimating θ from $X_1, \ldots, X_n$. Notation is simplified by using just f to stand for $f(x)$, and f_i for $f(X_i)$, etc.

Key words and phrases. Difference estimator, Horvitz–Thompson estimator, regression estimator, simulation, variance reduction.

Received May 2011; accepted May 2011.

2 Asymptotically optimum estimator

Suppose that g has the same support as f. Then for any fixed value of the constant c, the estimator

$$\hat{\theta}_c = \hat{\theta}_0 + c\Bigl(1 - n^{-1} \sum f_i / g_i\Bigr)$$

is unbiased, and has variance

$$\operatorname{var}(\hat{\theta}_c) = n^{-1} \int \Bigl\{ (\phi - c)\frac{f}{g} - (\theta - c) \Bigr\}^2 g \, \mathrm{d}x.$$

The variance here is easily shown to be minimized by the choice $c = \gamma$, say, where

$$\gamma = \frac{E_f(\phi f/g) - \theta}{E_f(f/g) - 1}; \tag{2.1}$$

and the minimum variance may be expressed as

$$\operatorname{var}(\hat{\theta}_\gamma) = \operatorname{var}(\hat{\theta}_0) - n^{-1} \frac{\{E_f(\phi f/g) - \theta\}^2}{E_f(f/g) - 1}.$$

The above optimality result follows directly from familiar ideas in the literature on survey sampling and on simulation. In the terminology of survey sampling, $\hat{\theta}_c$ is a difference estimator; see, for example, Särndal, Swensson and Wretman (1992), Section 6.3 for the general idea, and Van Deusen (1995) for application to the estimation of integrals. In the present context, if $h(x)$ is any function which is thought to approximate $\phi(x)$ to some extent, and whose mean $E_f\{h(X)\}$ is known, then $h(x)$ generates a difference estimator

$$\hat{\theta}_{h(x)} = E_f\{h(Y)\} + n^{-1} \sum_{i=1}^{n} (\phi_i - h_i) \frac{f_i}{g_i}$$

which is unbiased for θ. The estimator $\hat{\theta}_c$ above results from taking $h(x) \equiv c$, and optimality of the particular choice $c = \gamma$ follows either by simple calculus or as a special case of arguments given by Särndal, Swensson and Wretman (1992), Section 6.8. In terms of the literature on Monte Carlo methods, $\hat{\theta}_c$ is an example of the method of control variates. Since $E_g(f/g)$ is known to equal unity, f/g is available as a control variate in the estimation of $\theta = E_g(\phi f/g)$, and it is well known that the optimum choice of c is then $\operatorname{cov}_g(\phi f/g, f/g)/\operatorname{var}_g(f/g)$ (e.g., Ripley, 1987, p. 124), which is the same as γ above.

In practice, γ is usually unknown, and must also be estimated. If $\hat{\gamma}$ is an estimator such that $n^{1/2}(\hat{\gamma} - \gamma) = O_p(1)$, then

$$\begin{aligned} \hat{\theta}_{\hat{\gamma}} &= \hat{\theta}_\gamma + (\hat{\gamma} - \gamma)\Bigl(1 - n^{-1} \sum f_i / g_i\Bigr) \\ &= \hat{\theta}_\gamma + O_p(n^{-1}), \end{aligned}$$

since $n^{1/2}(1 - n^{-1}\sum f_i/g_i) = O_p(1)$ under sampling from g. Thus $\hat{\theta}_{\hat{\gamma}}$ has variance $\text{var}(\hat{\theta}_\gamma) + O(n^{-2})$ and asymptotically negligible bias of order $O(n^{-1})$: provided $\hat{\gamma}$ is $\sqrt{n}$-consistent, the optimum first-order efficiency of $\hat{\theta}_\gamma$ within the class $\hat{\theta}_c$ is attained by $\hat{\theta}_{\hat{\gamma}}$.

For estimation of γ from $X_1, \ldots, X_n$, various $\sqrt{n}$-consistent estimators immediately suggest themselves. One possibility is to estimate the unknowns $E_f(\phi f/g)$, $E_f(f/g)$ and θ in (2.1) by the corresponding unbiased sample quantities $n^{-1}\sum \phi_i(f_i/g_i)^2$, $n^{-1}\sum(f_i/g_i)^2$ and $\hat{\theta}_0$. Another is the ordinary least-squares estimate obtained by linear regression of $\phi_i f_i/g_i$ on f_i/g_i:

$$\hat{\gamma} = \frac{n^{-1}\sum \phi_i(f_i/g_i)^2 - (n^{-1}\sum f_i/g_i)\hat{\theta}_0}{n^{-1}\sum(f_i/g_i)^2 - (n^{-1}\sum f_i/g_i)^2}. \tag{2.2}$$

By the argument given above, these and other possibilities are equivalent to first order, asymptotically as $n \to \infty$. The least-squares estimator $\hat{\gamma}$ in (2.2) has some particularly appealing properties. In the trivial case where $\phi(x) \equiv \theta$, $\gamma = \hat{\gamma} = \theta$, exactly: in contrast to the standard estimator $\hat{\theta}_0$, $\hat{\theta}_{\hat{\gamma}}$ estimates θ without error in this case. Similarly, in the ideal but impracticable setting where $\phi \propto g/f$, $\hat{\gamma} = \gamma = 0$, and again θ is estimated without error. The latter property, in particular, suggests that the least-squares choice $\hat{\gamma}$ of (2.2) should enjoy a certain advantage in terms of second-order efficiency, but this has not been investigated in detail.

The estimator $\hat{\theta}_{\hat{\gamma}}$, with $\hat{\gamma}$ as in (2.2), has been previously studied by Hesterberg (1988, 1995, 1996), a primary part of whose motivation was the equivariance of that estimator under the addition of a constant to $\phi(x)$. The asymptotic optimality above complements Hesterberg's empirical studies, from which it is concluded that $\hat{\theta}_{\hat{\gamma}}$ performs well, and asymptotically as well as any of the competitors considered, in a variety of importance-sampling problems. We note that one of the competitors studied by Hesterberg (1988, 1995, 1996) is the "ratio" estimator

$$\hat{\theta}_{\text{ratio}} = \frac{\hat{\theta}_0}{n^{-1}\sum f_i/g_i},$$

which may be viewed as a simple renormalization of $\hat{\theta}_0$ to achieve equivariance under $\phi(x) \mapsto \phi(x) + \text{constant}$. In the finite-population sampling literature this corresponds to the well-known form $\sum(\phi_i/\pi_i)/\sum(1/\pi_i)$, where $\{\pi_i\}$ are first-order sample inclusion probabilities, and in the survey-sampling context this estimator is usually found to perform better (e.g., Särndal, Swensson and Wretman, 1992, Section 5.7) than the unbiased Horvitz–Thompson estimator $N^{-1}\sum \phi_i/\pi_i$ which corresponds to $\hat{\theta}_0$ above. In the empirical studies of Hesterberg (1988, 1995, 1996) and in the example of Section 3 below, $\hat{\theta}_{\text{ratio}}$ is found to perform markedly worse in the importance-sampling context than $\hat{\theta}_0$. This may be explained heuristically in terms of the $\hat{\theta}_c$ family: while $\hat{\theta}_0$ is optimal for the "ideal" importance-sampling problem in which $\phi \propto g/f$, it is easily shown that $\hat{\theta}_{\text{ratio}} = \hat{\theta}_\theta + O_p(n^{-1})$, so that

$\hat{\theta}_{\text{ratio}}$ is asymptotically optimal for the trivial problem in which $\phi(x) \equiv \theta$. So $\hat{\theta}_{\text{ratio}}$ is efficient in some situations where importance sampling itself offers little or no variance reduction, but is typically out-performed by $\hat{\theta}_0$, which in turn is out-performed by $\hat{\theta}_\gamma$, when importance sampling is effective.

3 Example

As an illustration, we use the now-classical example of Siegmund (1976), considered also in Ripley (1987), Section 5.2. Suppose $Z_1, Z_2, \ldots$ are independently distributed as $N(\mu, 1)$, with partial sums $S_k = Z_1 + \cdots + Z_k$. For $a \leq 0 < b$ let

$$T = \min\{k : S_k \leq a \text{ or } S_k \geq b\},$$

and define $\theta = \text{pr}(S_T \geq b) = E_f\{I(S_T \geq b)\}$, where f is the density of S_T and $I(\cdot)$ is the indicator function. Siegmund (1976) suggests importance sampling with realizations $X_1, \ldots, X_n$ of S_T obtained by drawing the $\{Z_i\}$ not from $N(\mu, 1)$ but from $N(-\mu, 1)$; the ratio of densities f/g is then $\exp(2\mu S_T)$, and the standard importance-sampling estimator is

$$\hat{\theta}_0 = n^{-1} \sum I(X_i \geq b) \exp(2\mu X_i).$$

Table 1 shows the results of a simulation experiment with $a = -4$, $b = 7$ and $n = 10^4$, patterned after Ripley (1987), Table 5.2. The $\hat{\gamma}$ used here is the least-squares estimator from (2.2).

The empirical efficiencies shown in Table 1 support the intuition developed in Section 2. When the variance reduction provided by importance sampling itself is very large—in the case $\mu = -0.5$ a factor of 9600 is reported by Ripley (1987)—$\hat{\theta}_0$ is close to optimal and no appreciable improvement is provided by $\hat{\theta}_{\hat{\gamma}}$. In other situations $\hat{\theta}_{\hat{\gamma}}$ is considerably more efficient than $\hat{\theta}_0$: when $\mu = -0.1$, for example, the variance reduction factor of 12 reported by Ripley (1987), for importance sampling with $\hat{\theta}_0$, is improved to a variance reduction factor of more than 200 by use of $\hat{\theta}_{\hat{\gamma}}$. In all cases the computational cost of $\hat{\theta}_{\hat{\gamma}}$ differs negligibly from that of $\hat{\theta}_0$.

Table 1 *Empirical performance of three estimators of* $\theta = \text{pr}(S_T \geq b)$. *Variances are all estimated from* 1000 *simulation runs, and are accurate to two significant digits*

		Estimated standard deviations			Est. rel. efficiency
μ	θ	$\hat{\theta}_0$	$\hat{\theta}_{\text{ratio}}$	$\hat{\theta}_{\hat{\gamma}}$	$\text{var}(\hat{\theta}_0)/\text{var}(\hat{\theta}_{\hat{\gamma}})$
−0.1	0.1444	0.0011	0.0026	0.00024	19.44
−0.2	0.0408	0.00020	0.0011	0.000092	4.52
−0.3	0.00991	0.000038	0.00041	0.000029	1.68
−0.5	0.000506	0.0000025	0.000063	0.0000024	1.02

Hesterberg (1988, 1995, 1996) provides further empirical evidence, finding that $\hat{\theta}_{\hat{\gamma}}$ is hugely more efficient than $\hat{\theta}_0$ in some problems, but that, as in the above example, there is little or no efficiency gain when the quantity θ being estimated is a very small probability.

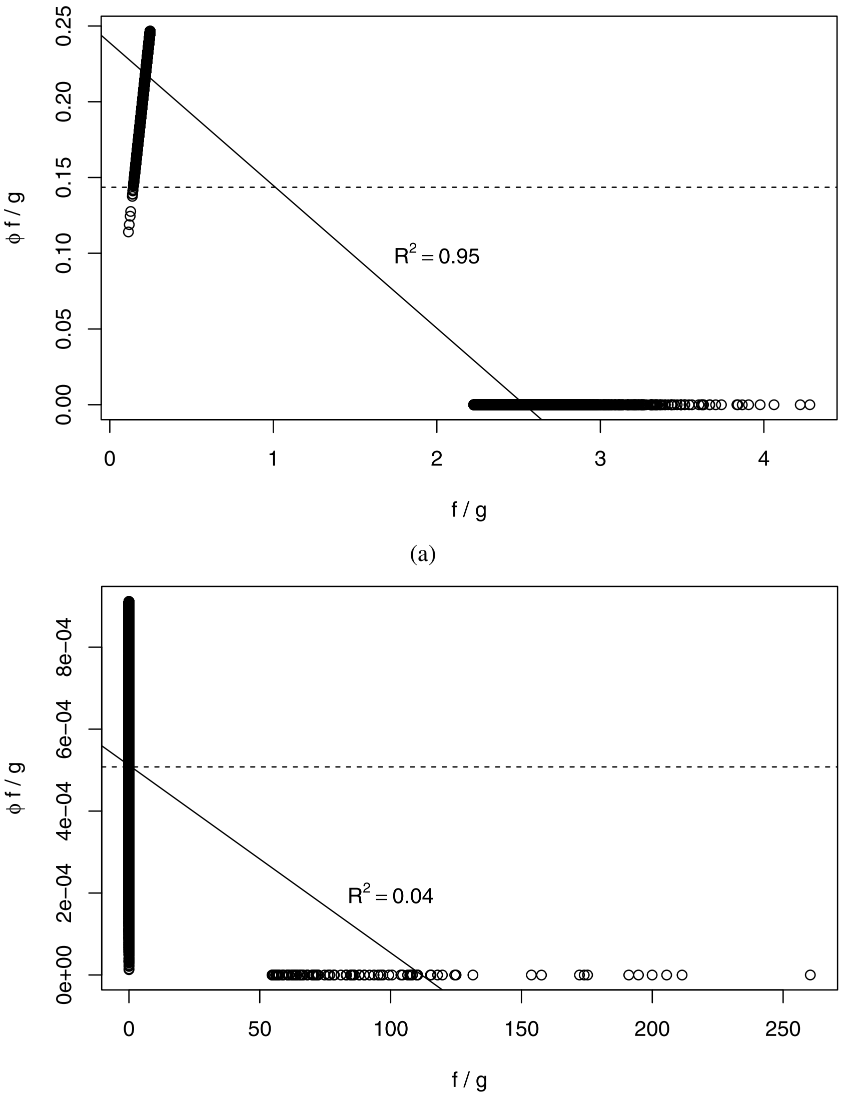
4 Discussion

We have considered here the "pure" importance-sampling problem in which nothing is known except that f and g are densities, so that the expectations $E_f(g/f) = E_g(f/g) = 1$ are known. With only this knowledge, $\hat{\theta}_c$ is the most general class possible of control-variate or difference estimators. If other functions are available whose expectation is known, such functions may be used as additional control variates, and may yield still further variance reduction; see Hesterberg (1996) for exploration of this.

The routine availability of f/g as a potentially helpful control variate seems to have been neglected in many, perhaps most, published applications of importance sampling. The resulting variance reduction is computationally inexpensive, and—as the above example shows—can be substantial. The gain in precision achieved by this device will be greatest when the regression of $\phi_i f_i/g_i$ on f_i/g_i explains a substantial fraction of the variance in the former. Figure 1 shows, for one typical sample drawn at each of $\mu = -0.1$ and $\mu = -0.5$ in the above example, the fitted regression line. Also shown in each panel of the figure is the corresponding statistical model implicit in the use of the standard estimator $\hat{\theta}_0$, which is a linear model with intercept only. The R^2 values are respectively 0.95 for $\mu = -0.1$ and 0.04 for $\mu = -0.5$; in the latter case, the low value of R^2 results in almost no reduction in variance relative to the standard estimator $\hat{\theta}_0$ (as was seen in Table 1). In Figure 1(b) there are fewer than 100 points (out of 10^4 in all) in the $\phi_i = 0$ group, whereas in Figure 1(a) there are more than 3000 such points.

An incidental point to note is that the plots displayed in Figure 1 show clearly that, from a statistical modeling perspective, a linear regression line is a poor fit to the data. It is evident that more variance could be explained by a suitable nonlinear fit. The use of nonlinear models in a probability sampling framework such as this is explored in some generality by Firth and Bennett (1998); unfortunately their results—on, for example, the use of generalized linear models—are rarely applicable in the present context, because expectations of nonlinear functions of f/g are required. The simple linear model's apparent lack of appeal as a statistical description of the sample does not, in any case, invalidate its use as shown above for the specific purpose of variance reduction.

The requirement that g has the same support as f is nontrivial: without it, the mean under importance sampling of $n^{-1}\sum f_i/g_i$ is not 1, and so $\hat{\theta}_c$ is biased. If a sampling density g is chosen whose support is a strict subset of that of f, as may



(a)

(b)

Figure 1 *Plot of* $\phi_i f_i/g_i$ *versus* f_i/g_i, *for two typical samples in the sequential testing example*: (a) *with* $\mu = -0.1$, (b) *with* $\mu = -0.5$. *The solid line is the fitted least squares regression which underlies* $\hat{\theta}_{\hat{\gamma}}$; *the dashed line, drawn at the sample mean of* $\phi_i f_i/g_i$, *represents the corresponding model underlying* $\hat{\theta}_0$.

be convenient for example if $\phi(x)$ is zero-valued in some interval, an appropriately adjusted definition of $\hat{\theta}_c$ would require knowledge of $E_f\{I(g > 0)\}$, which usually would be unavailable; in such a situation there appears to be no alternative to $\hat{\theta}_0$.

Acknowledgments

This work was supported by the United Kingdom Engineering and Physical Sciences Research Council.

References

Evans, M. and Swartz, T. B. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press. MR1859163

Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling (with discussion). *Journal of the Royal Statistical Society B* **60** 3–21. MR1625672

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. London: Chapman and Hall. MR0223065

Hesterberg, T. C. (1988). Advances in importance sampling. Ph.D. thesis, Stanford Univ. MR2637036

Hesterberg, T. C. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37** 185–194.

Hesterberg, T. C. (1996). Control variates and importance sampling for efficient bootstrap simulations. *Statistics and Computing* **6** 147–157.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley. MR0875224

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer. MR2080278

Särndal, C. E., Swensson, B. and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer. MR1140409

Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Annals of Statistics* **25** 673–684. MR0418369

Van Deusen, P. C. (1995). Difference sampling as an alternative to importance sampling. *Canadian Journal of Forest Research* **25** 487–490.

Department of Statistics
University of Warwick
Coventry CV4 7AL
United Kingdom
E-mail: d.firth@warwick.ac.uk