More Importance Sampling better title?

Valérie Ventura

Department of Statistics

Carnegie Mellon University

Pittsburgh PA 15213

Abstract

Assume that we want to estimate several quantities $\gamma_i(\theta) = \int c_i(x) f_\theta(x)\, dx$ via importance sampling, that is, using a single sample from a sampling density g. For convenience and efficiency, sampling from g should be easy, g and f_θ must overlap, and g must be chosen to minimize the variance of the estimates of $\gamma_i(\theta)$. But if several characteristics $\gamma_i(\theta)$ must be estimated, the method is unavoidably limited as a variance reduction technique, because g can be optimal only for one particular $\gamma_i(\theta)$. Worse, it can potentially be very non-optimal for other $\gamma_i(\theta)$'s. This paper presents an improvement upon importance sampling via double importance sampling (Ventura, 1997), which consists of applying the importance sampling identity twice to $\gamma_i(\theta)$ to obtain $\gamma_i(\theta) = \int c_i(x)\ [f_\theta(x)/h_{i\theta}(x)]\ [h_{i\theta}(x)/g(x)]\ g(x)\, dx$, where g is the same sampling density and h some function whose integral is known, or known up to the same normalizing constant as g. Contrary to importance sampling where only g can be chosen, double importance sampling allows the choices of g and $h_{i\theta}$: the recommended strategy is to choose g for ease of simulation and good coverage to prevent variance explosions of the estimates, while a different $h_{i\theta}$ can be designed for each pair (i, θ) to induce substantial variance reduction; $h_{i\theta}$ is not necessarilly a density, nor does it need to be sampled from. Moreover, all estimates can theoretically achieve zero variance if $h_{i\theta} \propto c_i f_\theta$, even if c_i does not have constant sign. The derivation of this result via double importance sampling (a) points to several strategies for designing $h_{i\theta}$, which, in particular, yields a regression estimator that has good properties, and (b) gives a reliable way to check if double importance sampling estimates are superior to the usual estimates, so that one can only do better but never worse than importance sampling alone. This method proves particularly effective if the sampling density g is much more spread than f_θ, or if f_θ is more spread than $c_i f_\theta$, as in small probability estimation.

Key Words: control variate, importance sampling, likelihood estimation, multiple regression estimate, ratio estimate, regression estimate, surface estimation, variance reduction.

1 Introduction

Suppose that we want to estimate

$$\gamma_i(\theta) = \mathrm{E}_{f_\theta}\{c_i(X)\} = \int c_i(x) f_\theta(x)\, dx \tag{1}$$

via simulation, for possibly varying values of a parameter θ, and possibly several functions c_i, where X is a random variable with density f_θ, and c_i is some function of x. Here and elsewhere, capital letters like X denote random variables with realizations x, and a subscript f_θ to an expectation indicates with respect to which density integration must be made. Applications include response surface estimation (Section 3.2), sensitivity analysis, and other problems where characteristics $\gamma_i(\theta)$ must be estimated for varying parameters of a density f_θ (Beckman and McKay, 1987, Hesterberg, 1988, 1995, Tukey, 1987), efficient bootstrap likelihood calculations (Davison, Hinkley, and Worton, 1992, 1995) and nested bootstrap simulations, where estimation must be made with respect to many first-level bootstrap distributions f_θ.

The importance sampling identity (Hammersley and Handscomb, 1964; Kahn, 1949) can be used to rewrite (1) as

$$\gamma_i(\theta) = \int c_i(x) \tfrac{f_\theta(x)}{g(x)}\, g(x)\, dx = \int c_i(x)\, w_\theta(x)\, g(x)\, dx \equiv \int t_{i\theta}(x)\, g(x)\, dx, \tag{2}$$

and we estimate $\gamma_i(\theta)$ by its empirical analogue,

$$\hat{\gamma}_i(\theta) = Q^{-1} \sum_{q=1}^{Q} t_{i\theta}(x_q), \tag{3}$$

where $x_1, \ldots, x_Q$ is a random sample of size Q from the sampling density g, used to estimate all $\gamma_i(\theta)$'s; this is the basic estimate, also reported in Table 1. For importance sampling to be practical and efficient, sampling from g should be easy, g must provide good coverage of the sample spaces of many densities f_θ, that is, g must overlap with the f_θ, and ideally g must be chosen to minimize the variance of (1). These goals might be conflicting; for example, the variance of $\hat{\gamma}_i(\theta)$ is minimized if $g \propto |c_i|\, f_\theta$ (Therneau 1983), a density which might be difficult to sample from. Also, this choice for g depends on the pair (i, θ), so can only be optimal for the associated $\gamma_i(\theta)$; worse, it might yield other estimates with huge, possibly infinite, variances. To guard against possible variance explosions, Hesterberg (1988, 1995) recommends use of a defensive mixture, that is of a sampling density g that includes f_θ as a mixture component; if estimation must be made with respect to many distributions simultaneously, as in (1), one can use a defensive mixture of several or all f_θ; see Section 2.1. But the drawback to safety is possible loss of efficiency, since, by construction, a defensive mixture is not optimal for estimation. Double importance sampling allows safety and efficiency simultaneously.

The rest of this section summarizes results of importance sampling, while double importance sampling and associated issues are described in Section 2. Section 3 contains examples, and Section 4 conclusions.

A systematic component of importance sampling is the random variable $w_\theta(X) = f_\theta(X)/g(X)$, whose expectation with respect to g is one. This extra information can be used to make adjustments to the basic estimate (3); see Cochran (1977, pp. 151, 193, 196). To simplify the notation here and in the rest of the paper, we will often drop the subscripts i and θ when there is no ambiguity. Let us thus denote by T and W the random variables $t(X) = c(X)f(X)/g(X)$ and $w(X) = f(X)/g(X)$ respectively, with observed values $t(x_q) = t_q$ and $w(x_q) = w_q$, and associated sample averages $\bar{t}$ and $\bar{w}$. The regression estimate is obtained by making the standard linear adjustment to (3) as

$$\hat{\gamma}_{reg} = \hat{\gamma} - \hat{\beta}(\bar{w} - 1), \tag{4}$$

which is the ordinary least squares estimate of γ for the linear model $T = \gamma + \beta(W - 1) + \epsilon$, where $\mathrm{E}_g(\epsilon) = 0$, the quantity of interest γ is the intercept, and β is the unknown regression slope of T against W with least squares estimate $\hat{\beta}$. A multiplicative, rather than a linear, adjustment yields the ratio estimate $\hat{\gamma}_{rat} = \hat{\gamma}/\bar{w}$. Table 1 gives formulae for all estimates discussed, as well as their first order asymptotic variances.

We prefer the regression estimator to the others. Indeed, the first order asymptotic variance of $\hat{\gamma}_{reg}$ is the smallest since $\rho^2 > 0$ and $(\beta - \gamma)^2 > 0$ (Cochran, 1977, eq. (7.34), p. 196), with equality with the

method		estimate	equation reference	Q × variance
integration	$\hat{\gamma}$	$\bar{t}$	(3)	V_T
ratio	$\hat{\gamma}_{rat}$	$\bar{t}/\bar{w}$		$V_T + \gamma^2 V_W - 2\gamma\rho(V_T V_W)^{1/2}$
regression	$\hat{\gamma}_{reg}$	$\bar{t} - \hat{\beta}(\bar{w} - 1)$	(4)	$V_T\,(1 - \rho^2)$
double ratio	$\hat{\gamma}'_{rat}$	$\bar{t}/\bar{w}'$		$V_T + \gamma^2 V_{W'} - 2\gamma\rho'(V_T V_{W'})^{1/2}$
double regression	$\hat{\gamma}'_{reg}$	$\bar{t} - \hat{\beta}'(\bar{w}' - 1)$		$V_T\,(1 - \rho'^2)$

Table 1: *Comparison of importance sampling methods.* $T = c(X)f(X)/g(X)$ *and* $W = f(X)/g(X)$, *with observed values* $t(x_q) = t_q$ *and* $w(x_q) = w_q$, *and sample averages* $\bar{t}$ *and* $\bar{w}$; $W' = h(X)/g(X)$ *with observed values* w'_q *and sample average* $\bar{w}'$; $\hat{\beta}$ *is the slope of the regression of* t_q *on* w_q, *and* $\hat{\beta}'$ *that of the regression of* t_q *on* w'_q. V_T, V_W, *and* $V_{W'}$ *are the population variances of* T, W, *and* W', $\rho = \mathrm{corr}(T, W)$ *and* $\rho' = \mathrm{corr}(T, W')$ *their population correlations.*

basic estimate only if $\rho = 0$, and with the ratio estimate if $\beta = \gamma$. Typically $\rho = \mathrm{corr}(T, W) \neq 0$ because $T = c(X)W$ depends on W, while $\beta = \gamma$ occurs only when $T = \gamma W + \epsilon$, $\mathrm{E}_g(\epsilon) = 0$, that is when $c(x)$ in (1) is the very constant γ we want to estimate. However, while the basic estimate (3) is unbiased and consistent for all (i, θ), provided that g is non-zero on the set $\mathcal{X} = \{x : \exists\theta, \exists i, c_i(x) f_\theta(x) \neq 0\}$, non-linear components of the relationship between $T = c(X)f(X)/g(X)$ and $W = w(X)$ will induce bias of order Q^{-1} in $\hat{\gamma}_{rat}$ and $\hat{\gamma}_{reg}$. Although asymptotically negligible to their standard errors, of order $Q^{-1/2}$, the biases will be substantial in finite simulations if g does not cover the sample space of f adequately, hence the need to design and monitor g; see Section 2.1.

2 Double importance sampling

The discussion above suggests that importance sampling is unavoidably limited as a variance reduction technique. Now consider using the importance sampling identity once more to rewrite (1) as

$$\gamma_i(\theta) = \int c_i(x)\, \tfrac{f_\theta(x)}{h_{i\theta}(x)}\, \tfrac{h_{i\theta}(x)}{g(x)}\, g(x)\, dx \equiv \int c'_{i\theta}(x)\, w'_{i\theta}(x)\, g(x)\, dx, \tag{5}$$

where $c'_{i\theta}(x) = c_i(x)\, f_\theta(x)/h_{i\theta}(x)$ and $w'_{i\theta}(x) = h_{i\theta}(x)/g(x)$ (Ventura, 1997). The functions $h_{i\theta}$ need not be densities. All we need is for their integrals $\int h_{i\theta}(x)\, dx = H_{i\theta}$ to be known at least up to the same normalizing constant as g; without loss of generality, we can $H_{i\theta} = 1$. Also, for consistency, it is required that $h_{i\theta}$ be non-zero whenever $c_i\, f_\theta$ is non-zero.

Equation (5) has the same structure as (2) (only with primes added to c and w), and g is the same sampling density; it yields the same basic estimate (3) because the integrand with respect to g, $c'\, w' = c\, w = t$, remains unchanged, but different ratio and regression estimates $\hat{\gamma}'_{rat}$ and $\hat{\gamma}'_{reg}$, since the weight function is now $w' = h/g$ instead of $w = f/g$. Formulae to obtain the new estimates are similar to $\hat{\gamma}_{rat}$ and $\hat{\gamma}_{reg}$ in Table 1, but with w and $\hat{\beta}$ replaced by w' and $\hat{\beta}'$, where $\hat{\beta}'$ is the estimated slope of regression of $T = cf/g(X)$ on $W' = h/g(X)$ ($\hat{\beta}$ was the estimated slope of the same T on $W = f/g(X)$); note that if $\int h(x)\, dx = 0$, $\hat{\gamma}'_{rat}$ is not defined.

Double importance sampling simply consists of choosing the control variate $W' = h(X)/g(X)$, or equivalently h in (5), rather than letting the sampling density impose the use of $W = f(X)/g(X)$. Our method leads to interesting constructions of h; see Section 2.2, and the examples of Section 3. Since regression estimators are asympotically superior to ratio estimators, we choose h to minimize the variance of $\hat{\gamma}'_{reg}$. From Table 1, the asymptotic variance of $\hat{\gamma}'_{reg}$ is $Q^{-1} V_T\, (1 - {\rho'}^2)$, where $V_T = \mathrm{var}(\hat{\gamma})$, which can be made zero if h is so that $|\rho'| = |\mathrm{corr}(T, W')| = 1$. For example $|\rho'| = 1$ if $t \propto w'$, equivalent to

$$h_{opt}(x) = \gamma^{-1} c(x) f(x), \tag{6}$$

where γ is the characteristic we want to estimate in (1). Additionally, if ${\rho'}^2 = 1$, the quadratic component of the regression of T on W' is zero (Cochran (1977) p.198), so that the bias of $\hat{\gamma}'_{reg}$ is of order Q^{-2}, one order smaller than the biases of $\hat{\gamma}_{reg}$ and $\hat{\gamma}_{rat}$. Note that (6) is the optimal sampling density for the basic estimate (3) when c has constant sign, but, contrary to importance sampling, we need not sample from h_{opt}, a potentially difficult task. Additionally, $\hat{\gamma}'_{reg}$ theoretically achieves zero variance whenever $h = h_{opt}$, whereas zero variance is only possible in importance sampling if c in (1) has constant sign, and $g = h_{opt}$; this is because g is a density, hence a positive function, whereas h can be any function we please. Further, independently of the sampling density g, we can use a different $h_{i\theta}$ for each $\gamma_i(\theta)$ in (1) to increase efficiency. A good strategy is to choose g for ease of simulation, and good coverage of the combined sample space of the f_θ in (1) to prevent variance explosion. Then a different $h_{i\theta}$ can be designed for each pair (i, θ) to induce substantial variance reduction in $\gamma_i(\theta)$; $h_{i\theta}$ need not be sampled from, and all regression estimates $\hat{\gamma}'_{reg}$ can theoretically achieve zero variance with only one simulation from g, if $h \propto cf$.

We now address practical issues: briefly the design and monitoring of the sampling density g, then possible constructions of h. Finally, a criterion is given to chose between importance sampling estimates versus their double importance sampling counterparts.

2.1 Sampling density

If g and f_θ do not overlap sufficiently, some estimates might have huge variances, and ratio or regression estimates, obtained either by simple or double importance sampling, might have non-negligible bias. In extreme case, ratio and regression estimates, and even the unbiased basic estimates (3), will be pseudo-biased, as described in Ventura (1998); see also Natarajan and McCulloch (1998) for a similar phenomenon in Gibbs sampling simulations, and Scharfstein, Rotnisky and Robins (1999) in semi-parametric non-ignorable missing data problems. Basically, a pseudo-biased estimate is asymptotically unbiased with large variance, but very likely has a large empirical bias and a low estimated variance even when calculated from a very large number of samples; hence it does not estimate what it should with very high probability, yet appears to be perfect since its empirical variance is low.

To illustrate this, consider estimating $\gamma = \Phi(\Phi^{-1}(0.01))$ using the sampling density $g(x) = \phi(x - \mu)$, where ϕ and Φ are the standard normal PDF and CDF. Figure 1 shows the average of 200 repeat basic and regression estimates of γ, as functions of μ, along with simulated standard deviation bounds. Values of μ close to $\Phi^{-1}(0.01) = -2.326$ yield good estimates. For $\mu \in [-3, -5]$, the regression estimate (4) becomes increasingly biased while the basic estimate (3) remains unbiased; estimated simulation variances of both estimates increase, so we can tell that (3) and (4) degrade. But as μ decreases below -5, (4) tends to a 'perfect' 1.0 and (3) tends to a 'perfect' zero, none of which the true value, where

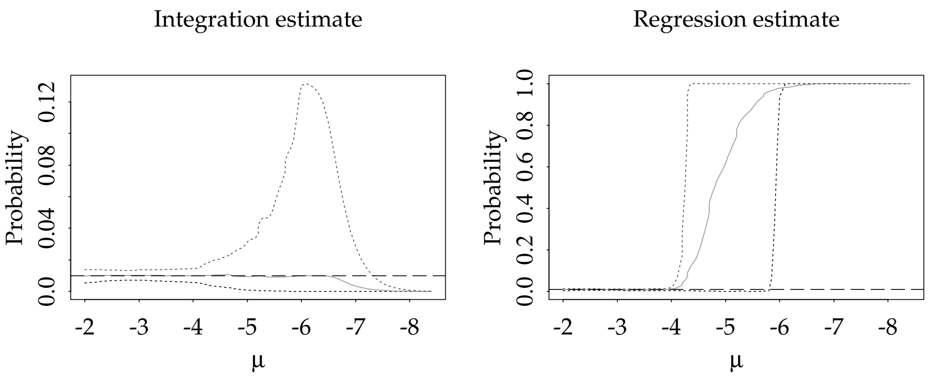


Figure 1: *Pseudo-bias of integration and regression estimates of the standard normal left tail* 1% *probability. The sampling density is* $N(\mu, 1)$*, with* μ *varying in* $[-8, -2]$*. The solid lines are averages of 200 repeat estimates (3) and (4) from simulations of size* $Q = 100$*, the dotted lines* 95% *simulated confidence bounds, and the dashed line the true value,* 0.01.

'perfect' means that the estimated simulation variances are almost zero; we then say that (3) and (4) are pseudo-biased. That is, one might think to have estimated a characteristic almost perfectly, judging from the small simulation variance, whereas the estimate does not estimate what it should with probability close to one.

Pseudo-bias happens because, as μ decreases, g provides increasingly poorer coverage of ϕ, which then makes the distribution of W increasingly skewed. This can be seen in Figure 2, where the PDF of W is plotted for $\mu = -5$ and -6. Extreme skewness of W can usually be avoided if g provides good coverage (exceptions include cases where f_θ is discontinuous at $x_0(\theta)$ in the middle of its range with, for example, $\lim_{x \to x_0(\theta)} f_\theta(x) = \infty$). Hesterberg's (1988, 1995) solution is to include f_θ as a mixture component in proportion $\alpha \in [0.1, 0.5]$ in g; in his terminology, g is a defensive mixture. This ensures coverage, hence negligible bias and no pseudo-bias. This also ensures safety, since in this case, the variance of $\hat{\gamma}_{rat}(\theta)$ is at most a factor α^{-1} larger than the variance of the estimate obtained by direct simulation from f_θ (Hesterberg, 1995). If estimation must be made with respect to many distributions simultaneously, as in (1), one can use a defensive mixture of several or all f_θ, with possibly other components, depending on the problem. In this paper, a mixture sampling density will be denoted as

$$g(x) = \sum_{k=1}^{K} \alpha_k \, g_k(x), \tag{7}$$

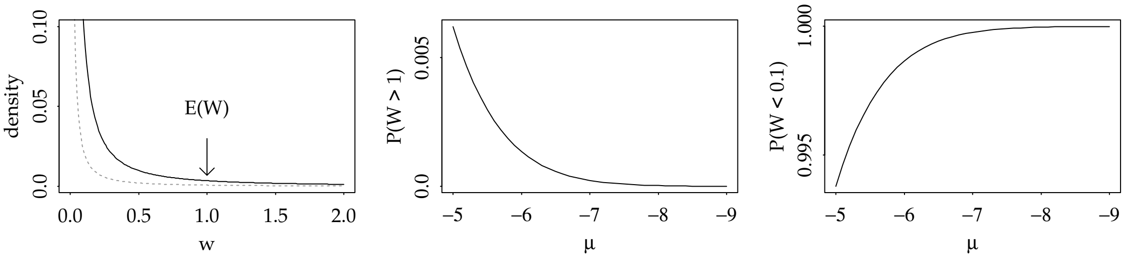


Figure 2: *Density function of W when $f(x) = \phi(x)$ and $g(x) = \phi(x - \mu)$. The left panel shows the pdf of W for $\mu = -5$ (solid line) and $\mu = -6$ (dotted line), the middle and right panels show the probabilities of observing a weight greater than* $\mathrm{E}(W) = 1$ *and smaller than* 0.1 *respectively, as functions of μ. The cdf of W is $F_W(w) = 1 - \Phi\left(\frac{\mu^2 - 2\log w}{2\mu} - \mu\right)$.*

with K mixture components g_k, in proportion α_k each. Hesterberg (1995) further showed that deterministic sampling from (7), that is drawing exactly $Q_k = Q\alpha_k$ observations from g_k rather than $Q_k \sim Bin(Q, \alpha_k)$ s.t. $\sum Q_k = 1$, reduces variability in the estimates.

Finally, even if we believe that g provides good coverage, it is important to monitor the importance sampling weights to detect skewness, since pseudo-bias cannot be detected by looking at empirical properties of estimators. Some diagnostics are described by Hesterberg (1995) and Ventura (1997), but none are entirely reliable.

2.2 Choosing h

Assume that the defensive sampling mixture (7) was already chosen. We must now chose h for efficient estimation, which must be such that, as best we can, $h \propto cf$. In this section, we offer alternatives to construct h, some of which are automatic.

2.2.1 h as an importance sampler

The first option is to make h look like cf "by eye", which might be possible in some applications. Related to this approach is to use $h_{i\theta}$ to be the sampling density one could have used to reduce the variability of the estimate of the corresponding $\gamma_i(\theta)$, had use of a denfensive mixture not been necessary. Although an $h_{i\theta}$ could be one of the mixture components g_k in (7), it is not at all necessary, which is advantageous because (a) it might be difficult to sample from $h_{i\theta}$, and (b) one might simply have too many quantities $\gamma_i(\theta)$ to estimate to include, as mixture components, densities designed to improve each of these. This approach is particularly effective if the sample space of the sampling density g is much more spread that the sample space of f_θ, as often occurs if estimates must be obtained for a wide range of values of θ, or if $c_i f_\theta$ is more concentrated than f_θ or g, as would occur if $\gamma_i(\theta)$ is a tail probability of f_θ, because then h is chosen to overlap with the important region of the sample space; see the examples in Section 3.

A more systematic alternative of obtaining h as correlated as possible with cf is to use some scatterplot smoothing technique to regress the dependent variable $Y = c(X)\, f(X)$, with observed values $y_q = c(x_q)\, f(x_q)$, on functions of the predictor X, with observed values x_q. See, for example, Venables and Ripley (1997, p. 323) for smoothing techniques. That is we estimate H in $c(X)\, f(X) = H(X) + \epsilon$ where, H could be a spline basis, or a set of other predictors if they are available. Then we use $h(x_q) = \hat{y}_q = \hat{H}(x_q)$, $q = 1 \ldots Q$, where $\hat{y}_q$ are the fitted values of the model. If H is a spline basis, then the resulting h is piecewise polynomial, so its integral easily calculated. That a close to optimal h can be contructed automatically once the sample from g is taken is a useful feature of the method that is not available in importance sampling alone. Spline fitting is difficult in high dimensions, but we can always obtain h by regression if predictors are available. Predictors that are always available are the mixture components g_k in (7).

2.2.2 Regression construction of h

Assume here that we have predictors, at least the mixture components $g_k(X)$ in (7), and maybe others that we will also denote by g_k for notational convenience. We need to be able to evaluate g_k at all simulation points x_q and to know $\int g_k(x)dx$ at least up to the same normalizing constant as g. If g_k is a mixture component, then $\int g_k(x)dx = 1$, and not only can we evaluate $g_k(x_q)$, but these were evaluated already to obtain the denominators of the importance sampling weights $g(x_q) = \sum_{k=1}^{K} \alpha_k\, g_k(x_q)$. So using these regressors involves little more than recycling some quantities we already have.

The aim is to construct $h(X)$ that is as correlated as possible with $cf_\theta(X)$, which we do by regressing $cf_\theta(X)$ on all the covariates $g_k(X)$, without intercept, and use the fitted values for h. That is, we fit the model (note, no intercept)

$$c(x_q)f_\theta(x_q)/g(x_q) = \sum_{k=1}^{K} \beta_k \, g_k(x_q)/g(x_q) + \epsilon_q, \quad q = 1 \ldots Q \tag{8}$$

with least square estimates $\hat{\beta}_k$, and use

$$\hat{h}(x) = \sum_{k=1}^{K} \hat{\beta}_k \, g_k(x) \tag{9}$$

with integral $\sum_{k=1}^{K} \hat{\beta}_k$, immediately provided by (8).

Once $\hat{h}$ is fitted, we can calculate the double importance sampling estimator as described in Table 1. More explicitly here,

$$\hat{\gamma}'_{reg} = \bar{t} - \hat{\beta}' \left(Q^{-1} \sum_{q=1}^{Q} \hat{h}(x_q)/g(x_q) - \sum_{k=1}^{K} \hat{\beta}_k \right), \tag{10}$$

where $\hat{\beta}$ is the least square estimate of the slope of the linear regression, with intercept, of $c(x_q)f_\theta(x_q)/g(x_q)$ on $\hat{h}(x_q)/g(x_q)$.

2.2.3 Importance sampling multiple regression estimator

We could make another use of the mixture components g_k to generalize the regression estimator (4) to a multiple regression estimator. Indeed, the random variables $W_k = g_k(X)/g(X)$ have mean one under sampling from g, so can be used as covariate to form the multiple regression estimator

$$\hat{\gamma}_{mreg} = \bar{t} - \sum_{k=1}^{K} \hat{\delta}_{\theta k} \left(Q^{-1} \sum_{q=1}^{Q} g_k(x_q)/g(x_k) - 1 \right) \tag{11}$$

where $\hat{\delta}_k$ are the least square estimate of the linear regression with intercept

$$c(x_q)f_\theta(x_q)/g(x_k) = \delta_0 + \sum_{k=1}^{K} \delta_k \, g_k(x_q)/g(x_k) + \epsilon_q, \quad q = 1 \ldots Q. \tag{12}$$

The multiple regression estimator $\hat{\gamma}_{mreg}$ is in fact (exactly?) the same as (10), which is obtained by finding $\hat{h}$ first, by means of the regression without intercept (8), then by regressing $c(x_q)f_\theta(x_q)$ on the fitted values (9).

2.3 Importance sampling versus double importance sampling

Should an efficient $\hat{\gamma}'_{reg}$ always be preferred to a safer $\hat{\gamma}_{reg}$? Indeed, it might not always be easy to design a good h. From Table 1, the asymptotic variances of the two estimators are $Q^{-1}V_T(1-\rho'^2)$ and $Q^{-1}V_T(1-\rho^2)$ respectively, where $V_T = \text{var}(\hat{\gamma})$, $\rho = \text{corr}(T, W)$ and $\rho' = \text{corr}(T, W')$. So $\hat{\gamma}'_{reg}$ is superior if $\rho'^2 > \rho^2$. The true correlations are unknown, but we verify in the examples of Section 3, that the sample correlations are reliable indicators of which estimate should be used. This should guard us from doing worse than importance sampling alone. Note that sample correlations are sensitive to outliers, and so they are reliable only if W and W' are not too skewed, that is if g provides adequate coverage of f_θ (which should be the case when we use a defensive mixture), and of $h_{i\theta}$.

The asymptotic variance of the multiple regression (11) is $Q^{-1}V_T(1 - R^2(T, [g_1(X), \ldots, g_k(X)])$, where $R(T, [g_1(X), \ldots, g_k(X)]$ is the multiple correlation coefficient of T and predictors $g_1(X), \ldots, g_k(X)$. We found, in the examples of Section 3, that the estimated aymptotic variance of (11) does not match its simulated variance well, but rather overestimates it.

3 Applications

3.1 An example

In this example, we illustrate several ways in which h can be constructed, show the performance of the method, and demonstrate that the sample correlations $\hat{\rho} = \text{corr}(T, W)$ and $\hat{\rho}' = \text{corr}(T, W')$ are reliable indicators of which regression estimator should be used.

Assume that X is univariate normal, with PDF $f_\theta(x) = \phi(x-\theta)$ and CDF $F_\theta(x) = \Phi(x-\theta)$, where ϕ and Φ are the standard normal PDF and CDF. For all $\theta \in [0, 10]$, we want to estimate the right tail probability

$$\gamma_1(\theta) = \mathrm{E}_{f_\theta}\left(I[X \geq F_\theta^{-1}(0.99)]\right), \tag{13}$$

where I is the indicator function, and the average

$$\gamma_2(\theta) = \mathrm{E}_{f_\theta}(X), \tag{14}$$

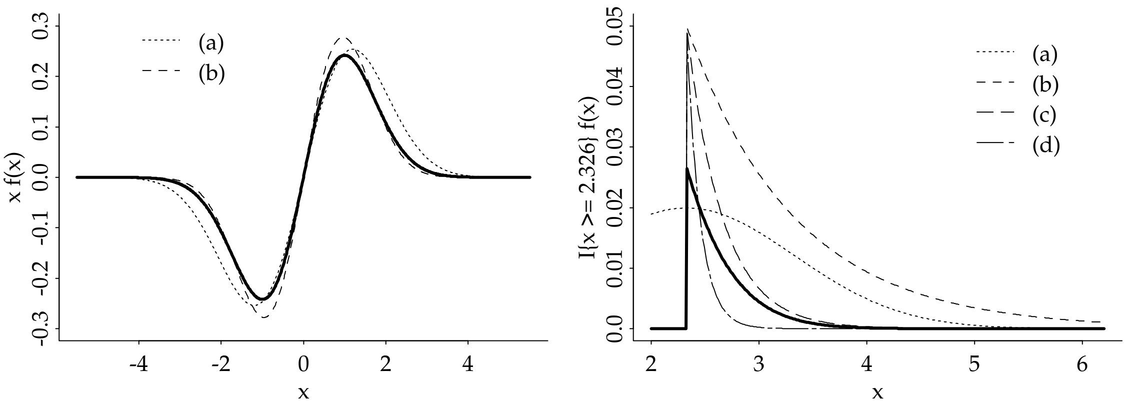
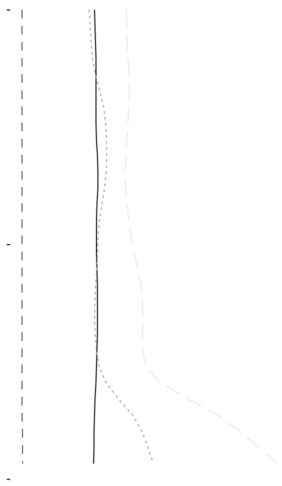


Figure 3: *Plots of cf (bold curves), and of the functions h given in Table 2. The left panel corresponds to the mean (14) with $c_2(x) = x$ and the right panel to the tail probability (13) with $c_1(x) = I\{x \geq 2.326\}$. Most curves were rescaled for easier visual comparison of cf and h.*

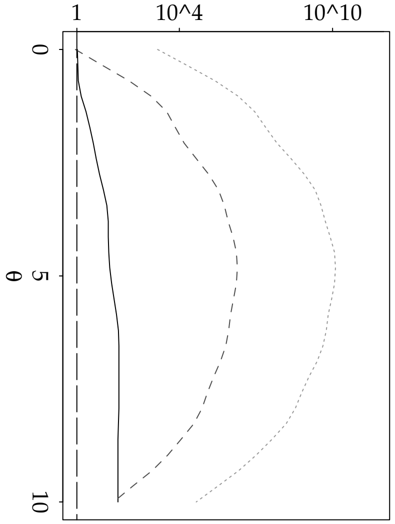
with true values 0.01 and θ respectively. An example of a mixture g that provides good coverage of the combined sample space of the f_θ includes, in equal mixing proportions, $K = 20$ components, all normal densities with variance 1, and means uniformly equidistant between 0 and 10; that is, $K = 20$, $\alpha_k = K^{-1}$ and $g_k(x) = \phi\,(x - (k-1)(10/19))$, $k = 1 \ldots K$ in (7).

First we choose h "by eye" to demonstrate the potential of the method in the special case when $\theta = 0$; we will later construct h by regression. Table 2 contains our choices of h, also diplayed in Figure 3. We produced 100 repeat estimates of $\gamma_1(0)$ and $\gamma_2(0)$ from which empirical efficiencies of regression versus basic estimates were calculated as ratios of estimated variances. Estimated asymptotic efficiencies were also obtained as the averages of the 100 repeats of $(1 - \hat\rho^2)^{-1}$ and $(1 - \hat\rho'^2)^{-1}$. Table 2 summarizes these efficiencies. Ratio estimates were inferior, hence not reported. All estimates have negligible bias. Efficiency gains can be substantial, and even more so the aggregate gains over all parameters of interest, depending on h; further, simulated and estimated asymptotic efficiencies agree well, so sample correlations appear to be accurate indicators of which regression estimator should be used: $\hat\gamma'_{reg}$ should be preferred to $\hat\gamma_{reg}$ only if a function h such that $\hat{\rho'}^2 > \hat\rho^2$ can be constructed. This guards us against doing worse than importance sampling alone.

Simulated efficiency for (12)



Simulated efficiency for (13)



Asymptotic efficiency for (13)