Importance Sampling

Used for a number of purposes:

- Variance reduction
- Allows for difficult distributions to be sampled from.
- Sensitivity analysis
- Reusing samples to reduce computational burden.

Idea is to sample from a different distribution that picks points in "important" regions of the sample space.

Want

$$E\left[f(X)\right] = \int f(x)\,g(x)\,dx$$

Instead of sampling from density (or probability mass function) $g(x)$, sample from a distribution with density (or pmf) $h(x)$.

Since we are sampling from the "wrong" distribution we have to make adjustments in our estimator.

$$\begin{aligned} E_g\left[f(X)\right] &= \int f(x)g(x)dx \\ &= \int f(x)\frac{g(x)}{h(x)}h(x)dx \\ &= E_h\left[f(X)\frac{g(X)}{h(X)}\right] \end{aligned}$$

This suggests the following estimation scheme

1) Sample $x_1,\ldots,x_n$ from $h(x)$.

2) Calculate weights

$$w_i = \frac{g(x_i)}{h(x_i)}$$

3) Use estimator

$$\hat{\mu}_{f,IS} = \frac{1}{n}\sum_{i=1}^{n} w_i f(x_i)$$

So instead of a regular average, this estimator is a weighted average.

So points that occur more often under $h(x)$ than $g(x)$ get downweighted and those that occur less often get upweighted.

Notice that $\hat{\mu}_{f,IS}$ is an unbiased estimate of $E_g\left[f\left(X\right)\right]$ regardless of which proposal distribution $h\left(x\right)$ as long as $h\left(x\right)$ has the same support as $g\left(x\right)$, i.e.

$$g\left(x\right)>0 \text{ implies that } h\left(x\right)>0$$

Note that $h\left(x\right)>0$ can be allowed to occur when $g\left(x\right)=0$, though doing this tends to be inefficient (but there are times you want to do this).

Since $\hat{\mu}_{f,IS}$ is unbiased, the main idea is to pick a distribution $h(x)$ that reduces the variance.

$$\mathrm{Var}_h\left(\frac{f(X)g(X)}{h(X)}\right) = E_h\left[\left(\frac{f(X)g(X)}{h(X)}\right)^2\right] - \mu_f^2$$

To do this, we want $h(x)$ to look like $f(x)g(x)$, i.e. make

$$\frac{f(x)g(x)}{h(x)}$$

look like a constant.

The optimal $h(x)$ satisfies

$$h(x) = \frac{|f(x)|g(x)}{\int |f(x)|g(x)dx}$$

Note that this usually can't be determined, due to the normalizing constant.

However this does give us a motivation for picking $h(x)$.

Example: Monte Carlo Evaluation of a Likelihood Ratio (Genetics Example)

Assume that you have a missing data model where $X = \left(X_{obs}, X_{mis}\right)$. Then the observed data likelihood ratio satistifies

$$l\left(\theta_1, \theta_0\right) = \frac{L\left(\theta_1\right)}{L\left(\theta_0\right)} = \frac{p_{\theta_1}\left(X_{obs}\right)}{p_{\theta_0}\left(X_{obs}\right)}$$

$$= E_{\theta_0}\left[\frac{p_{\theta_1}\left(X_{obs}, X_{mis}\right)}{p_{\theta_0}\left(X_{obs}, X_{mis}\right)} \middle| X_{obs}\right]$$

This can be estimated by sampling $z_1, \ldots, z_n$ from $p\left(X_{mis} \middle| X_{obs}\right)$ calculating

1) $f\left(z_i\right) = \dfrac{p_{\theta_1}\left(X_{obs}, z_i\right)}{p_{\theta_0}\left(X_{obs}, z_i\right)}$

2) $\hat{l}\left(\theta_1, \theta_0\right) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} f\left(z_i\right)$

Suppose that you are interested in getting $l\left(\theta_2, \theta_0\right)$, based on this Monte Carlo estimate.

This can be done with the importance sampling estimate

$$\hat{l}\left(\theta_2,\theta_0\right)=\frac{1}{n}\sum_{i=1}^{n} f\left(z_i\right)\frac{p_{\theta_2}\left(X_{obs},z_i\right)}{p_{\theta_1}\left(X_{obs},z_i\right)}$$

This can be shown to be an unbiased estimator of $l\left(\theta_2,\theta_0\right)$.

Genetics example:

Observed Data Model

$$\left(Y_1,Y_2,Y_3,Y_4\right)\sim \text{Multi}\left(197,\left(\frac{\lambda}{4},\frac{1-\lambda}{4},\frac{1-\lambda}{4},\frac{2+\lambda}{4}\right)\right)$$

$$g\left(Y\middle|\lambda\right)=\left(\frac{\lambda}{4}\right)^{Y_1}\left(\frac{1-\lambda}{4}\right)^{Y_2+Y_3}\left(\frac{2+\lambda}{4}\right)^{Y_4}$$

Complete Data Model

$$\left(X_1,X_2,X_3,X_4,X_5\right)$$
$$\sim \text{Multi}\left(197,\left(\frac{\lambda}{4},\frac{1-\lambda}{4},\frac{1-\lambda}{4},\frac{\lambda}{4},\frac{1}{2}\right)\right)$$

$$g\left(X\left|\lambda\right.\right)=\left(\frac{\lambda}{4}\right)^{X_1+X_4}\left(\frac{1-\lambda}{4}\right)^{X_2+X_3}\left(\frac{1}{2}\right)^{X_5}$$

As seen before $X_4\left|Y_4\right. \sim \mathrm{Bin}\left(Y_4,\frac{\lambda}{2+\lambda}\right)$

The complete data likelihood ratio satisfies

$$\frac{g\left(Y,X\left|\lambda_1\right.\right)}{g\left(Y,X\left|\lambda_0\right.\right)}=\left(\frac{\lambda_1}{\lambda_0}\right)^{Y_1+X_4}\left(\frac{1-\lambda_1}{1-\lambda_0}\right)^{Y_2+Y_3}$$

Note that this implies the importance sample weight satisfies

$$w_i=c\left(Y_1,Y_2,Y_3,\theta_2,\theta_1\right)\left(\frac{\theta_2}{\theta_1}\right)^{z_i}$$

In this case $\hat{l}\left(\lambda_2,\lambda_0\right)$ has the form

$$\hat{l}\left(\lambda_2,\lambda_0\right)=\frac{c\left(Y_1,Y_2,Y_3,\lambda_2,\lambda_1\right)}{n}\sum_{i=1}^{n}f\left(z_i\right)\left(\frac{\lambda_2}{\lambda_1}\right)^{z_i}$$

As in many problems, the desired sampling distribution doesn't need to be known exactly, but only up to the normalizing constant (i.e. $l(x) = cg(x)$).

Importance sampling still works fine in this case.

1) Sample $x_1, \dots, x_n$ from $h(x)$.

2) Calculate weights

$$w_i = \frac{l(x_i)}{h(x_i)}$$

3) Use estimator

$$\begin{aligned}\hat{\mu}_{f,IS} &= \frac{1}{W} \sum_{i=1}^{n} w_i f(x_i) \\ &= \frac{1}{n\bar{w}} \sum_{i=1}^{n} w_i f(x_i)\end{aligned}$$

where $W = \sum w_i$.

Properties of this estimator

$$
\begin{aligned}
E_h\left[w(X)f(X)\right] &= \int \frac{l(x)}{h(x)} f(x)h(x)\,dx \\
&= \int f(x)cg(x)\,dx \\
&= cE_g\left[f(X)\right]
\end{aligned}
$$

$$
\begin{aligned}
E_h\left[w\right] &= \int \frac{l(x)}{h(x)} h(x)\,dx \\
&= \int cg(x)\,dx = c
\end{aligned}
$$

As this is a ratio estimator, it is no longer unbiased, but it is consistent.

In addition, when c is known, this estimator is often preferred to the unbiased one discussed last time as it often has a smaller mean square error (to be discussed later).

Efficiency of importance sampling

Effective sample size

$$
\mathrm{ESS}(n) = \frac{n}{1+\mathrm{var}_h\left(w(X)\right)}
$$

Since the weights are usually only known up to the normalizing constant, $\text{var}_h\left(w(X)\right)$ needs to be estimated by the coefficient of variation of the unnormalized weights

$$\text{cv}^2\left(w\right) = \frac{\sum_{i=1}^{n}\left(w_i - \bar{w}\right)^2}{\left(n-1\right)\bar{w}^2}$$

Assume for what follows that $c = 1$ (and this is known).

We have two estimators

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} w_i f\left(x_i\right) \qquad \text{(unbiased)}$$

$$\hat{\mu} = \frac{1}{n\bar{w}}\sum_{i=1}^{n} w_i f\left(x_i\right) \equiv \frac{\tilde{\mu}}{\bar{w}} \qquad \text{(ratio)}$$

Let $Z = w(X)\,f\left(X\right)$. Then by the delta method

$$E_h\left[\hat{\mu}\right] \approx E_h\left[\bar{Z}\left(1-\left(\bar{w}-1\right)+\left(\bar{w}-1\right)^2\right)\right]$$

$$\approx \mu - \frac{\text{cov}_h\left(w, Z\right)}{n} + \frac{\mu\,\text{Var}_h\left(W\right)}{n}$$

In addition,

$$\operatorname{Var}_h(\hat{\mu})$$
$$\approx \frac{1}{n}\left(\mu^2 \operatorname{Var}_h(w) + \operatorname{Var}_h(Z) - 2\mu \operatorname{Cov}_h(w, Z)\right)$$

For the unbiased estimator,

$$E_h[\tilde{\mu}] = \mu \text{ and } \operatorname{Var}_h(\tilde{\mu}) = \operatorname{Var}_h(Z)/n$$

Thus

$$\operatorname{MSE}(\tilde{\mu}) = \frac{\operatorname{Var}_h(Z)}{n}$$

and

$$\operatorname{MSE}(\hat{\mu}) = \left(E_h[\hat{\mu}] - \mu\right)^2 + \operatorname{Var}_h(\hat{\mu})$$
$$= \frac{\operatorname{MSE}_h(\tilde{\mu})}{n}$$
$$+ \frac{1}{n}\left(\mu^2 \operatorname{Var}_h(w) - 2\mu \operatorname{Cov}_h(w, Z)\right) + O\left(n^{-2}\right)$$

Thus the ratio estimate is to be preferred when

$$2\mu \operatorname{Cov}_h(w, Z) > \mu^2 \operatorname{Var}_h(w)$$

(assuming $\mu > 0$). That is when $w(X)$ and $w(X)f(X)$ are highly correlated.

In addition, the formula for $\mathrm{Var}_h(\hat{\mu})$ implies (whole bunch of steps omitted)

$$\mathrm{Var}_h(\tilde{\mu}) \approx \mathrm{Var}_g(f(X))\left(1+\mathrm{Var}_h(w(X))\right)/n$$

Rearranging this gives

$$\frac{\mathrm{Var}_g(f(X))}{\mathrm{Var}_h(w(X)f(X))} \approx \frac{1}{1+\mathrm{Var}_h(w(X))}$$

One way of thinking of this statement is that n samples from the proposal distribution $h(x)$ is worth $n/\left(1+\mathrm{Var}_h(w(X))\right)$ samples drawn from the target distribution $g(x)$.

The nice thing with this rule of thumb is that it doesn’t depend on the function being integrated.

Thus the effective sample size is a useful measure of the efficiency of the method when different functions f are investigated with a single sample.

One consequence of this is that we want to keep the coefficient of variation of the weights well behaved.

One way of doing this is to have the proposal distribution $h(x)$ be heavier tailed than the target distribution $g(x)$. This will help minimize

$$E_h\left[\left(\frac{g(X)}{h(X)}\right)^2\right] = \int\left(\frac{g(x)}{h(x)}\right)^2 h(x)\,dx$$

$$= E_g\left[\frac{g(X)}{h(X)}\right]$$

which will keep $\mathrm{Var}_h\left(g(X)/h(X)\right)$ small.

For example, use a t distribution with moderate degrees of freedom instead of a normal.

Marginalization in importance sampling

Let $g\left(X_1, X_2\right)$ and $h\left(X_1, X_2\right)$ be two probability densities where the support of g is a subset of the support of h (e.g. $g\left(X_1, X_2\right) > 0$ implies $h\left(X_1, X_2\right) > 0$).

Then

$$\mathrm{Var}_h\left(\frac{g\left(X_1, X_2\right)}{h\left(X_1, X_2\right)}\right) \geq \mathrm{Var}_h\left(\frac{g\left(X_1\right)}{h\left(X_1\right)}\right)$$

where $g\left(X_1\right)$ and $h\left(X_1\right)$ are the respective marginal distributions.

$$\begin{aligned} E_h\left[\frac{g\left(X_1, X_2\right)}{h\left(X_1, X_2\right)}\middle| X_1\right] & \\ &= \int \frac{g\left(X_1, X_2\right)}{h\left(X_1, X_2\right)} h\left(X_2 \middle| X_1\right) dX_2 \\ &= \int \frac{g\left(X_1, X_2\right)}{h\left(X_1\right) h\left(X_2 \middle| X_1\right)} h\left(X_2 \middle| X_1\right) dX_2 \\ &= \frac{g\left(X_1\right)}{h\left(X_1\right)} \int g\left(X_2 \middle| X_1\right) dX_2 = \frac{g\left(X_1\right)}{h\left(X_1\right)} \end{aligned}$$

Thus

$$\mathrm{Var}_h\left(\frac{g(X_1,X_2)}{h(X_1,X_2)}\right) \geq \mathrm{Var}_h\left(E\left[\frac{g(X_1,X_2)}{h(X_1,X_2)}\middle|X_1\right]\right)$$

$$= \mathrm{Var}_h\left(\frac{g(X_1)}{h(X_1)}\right)$$

In addition

$$\mathrm{Var}_h\left(\frac{g(X_1,X_2)}{h(X_1,X_2)}\right) - \mathrm{Var}_h\left(\frac{g(X_1)}{h(X_1)}\right)$$

$$= E\left[\mathrm{Var}_h\left(\frac{g(X_1,X_2)}{h(X_1,X_2)}\middle|X_1\right)\right]$$

The implication of this result is that where possible, minimize the number of variables you sample as this will increase $\mathrm{ESS}(n)$.

However the computational burden must also be considered.

For example, if

$$\mathrm{Var}_h\left(\frac{g(X_1,X_2)}{h(X_1,X_2)}\right)=2\,\mathrm{Var}_h\left(\frac{g(X_1)}{h(X_1)}\right)$$

but the computational time involved in sampling $h(X_1)$ is 4 times the time involved sampling $h(X_1,X_2)$, sampling over the second space is to be preferred.

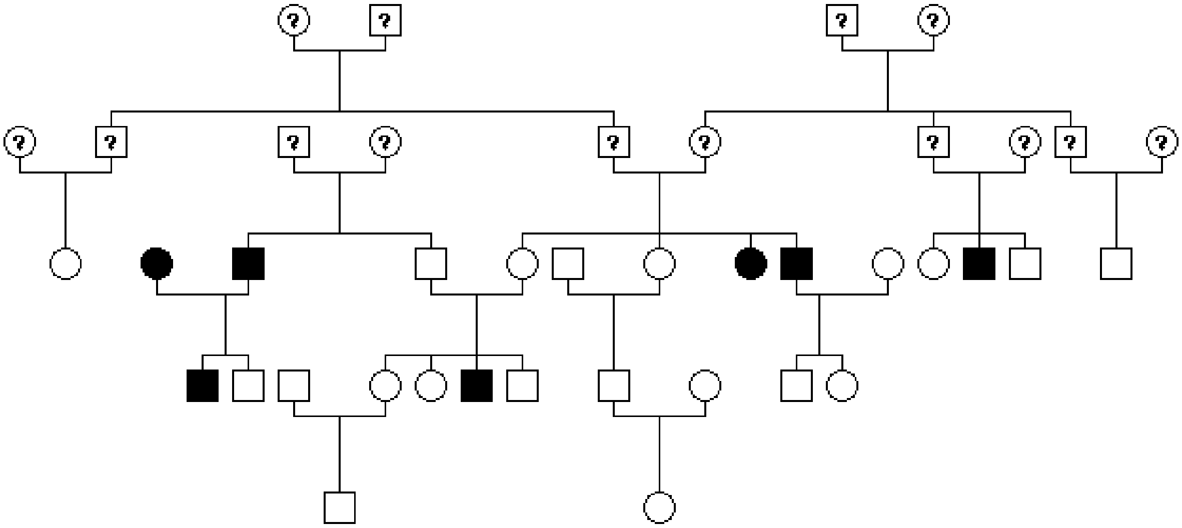
Similarly, Rao-Blackwellization can be with importance sampling, giving an estimate of the form

$$\breve{\mu}=\frac{1}{W}\sum_{i=1}^{n} w_i E_g\left[f\left(X_1,X_2\right)\middle|x_{1,i}\right]$$

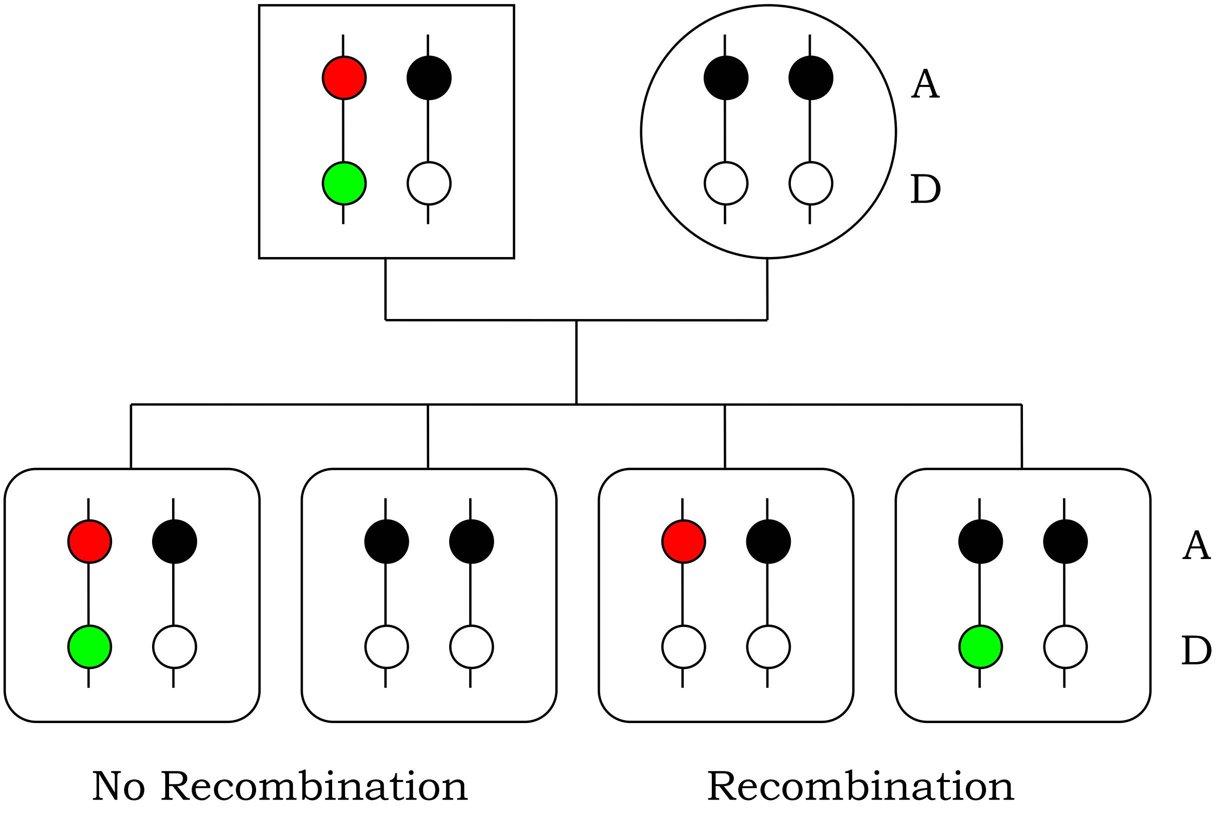
While this is a consistent estimate, it may not have a smaller variance than the non-Rao-Blackwellized estimate. However in many situations it should.

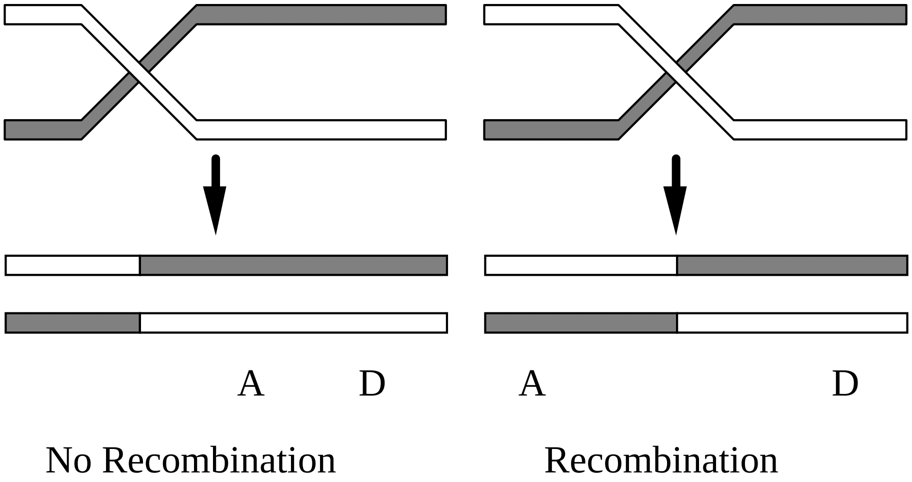
The following is an example where it did help.

Estimating recombination fractions with pedigree data



- 41 members
- n = 27 (nonfounders), f = 14 (founders)
- 8 markers from chromosome 19
- #alleles ranges from 6 to 8
- 14 members in top 2 generation have no marker data
- Want to use pedigree to estimate the distances between the 8 markers





Can use the recombination fraction (P[recombination]) as a measure of distance.

Data Structure

- $\mathbf{x}_l$ = data on locus l
  - o $\mathbf{x}_M = \left(\mathbf{x}_1,\ldots,\mathbf{x}_m\right)$ m markers
  - o $\mathbf{x}_D$ = disease / trait data
  - o $\mathbf{x} = \left(\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{x}_D\right)$
- $\mathbf{y}_l$ = haplotype for locus l
  - o $\mathbf{y} = \left(\mathbf{y}_1,\ldots,\mathbf{y}_m,\mathbf{y}_D\right)$
  - o Allele information with parental source
- $\mathbf{z}_l$ = inheritance vector for locus l
  - o $\mathbf{z} = \left(\mathbf{z}_1,\ldots,\mathbf{z}_m,\mathbf{z}_D\right)$
  - o Indicators of whether allele inherited came from the grandmother or grandfather

Assume that $\mathbf{z}_l$ is a vector of length $2n$. Then an estimate of θ_j, the recombination fraction between markers j and $j + 1$ is

$$\theta_j = \frac{1}{2n}\sum_{i=1}^{2n} I\left(z_{j,i} \neq z_{j+1,i}\right)$$

assuming that $\mathbf{z}$ is known.

Unfortunately, for most data sets, **z** can't be determined with certainty. However it is possible to do simulation and estimate the recombination fractions using Monte Carlo EM (MCEM)

As part of this procedure is it necessary to calculate $E\left[\mathbf{z}_l \middle| \mathbf{x}, \boldsymbol{\theta}\right]$ via Monte Carlo. There are two approaches to doing this.

1) Sample $\mathbf{y}^{(i)}$from $p_{\boldsymbol{\theta}}\left(\mathbf{y} \middle| \mathbf{x}\right)$ by importance sampling

   Sample $\mathbf{z}^{(i)}$ from $p_{\boldsymbol{\theta}}\left(\mathbf{z} \middle| \mathbf{x}, \mathbf{y}^{(i)}\right)$

   Estimate $E\left[\mathbf{z}_l \middle| \mathbf{x}, \boldsymbol{\theta}\right]$ by

$$\frac{1}{W} \sum_{i=1}^{m} w_i \mathbf{z}_l^{(i)}$$

2) Sample $\mathbf{y}^{(i)}$from $p_{\boldsymbol{\theta}}\left(\mathbf{y} \middle| \mathbf{x}\right)$ by importance sampling

   Estimate $E\left[\mathbf{z}_l \middle| \mathbf{x}, \boldsymbol{\theta}\right]$ by

$$\frac{1}{W} \sum_{i=1}^{m} w_i E\left[\mathbf{z}_l \middle| \mathbf{y}^{(i)}, \boldsymbol{\theta}\right]$$

Both approaches estimate exactly the same quantity, but the 2nd approach is much more efficient.

First the second approach has a smaller variance.

In addition, calculating the expected value is actually faster than sampling the ***z***'s. (usually doesn't work out this way).