9 Drawing from Densities

9.1 Introduction

Simulation consists of drawing from a density, calculating a statistic for each draw, and averaging the results. In all cases, the researcher wants to calculate an average of the form $\bar{t} = \int t(\varepsilon) f(\varepsilon)\, d\varepsilon$, where $t(\cdot)$ is a statistic of interest and $f(\cdot)$ is a density. To approximate this average through simulation, the researcher must be able to take draws from the density $f(\cdot)$. For some densities, this task is simple. However, in many situations, it is not immediately clear how to draw from the relevant density. Furthermore, even with simple densities, there may be ways of taking draws that provide a better approximation to the integral than a sequence of purely random draws.

We explore these issues in this chapter. In the first sections, we describe the most prominent methods that have been developed for taking purely random draws from various kinds of densities. These methods are presented in a progressive sequence, starting with simple procedures that work with a few convenient densities and moving to ever more complex methods that work with less convenient densities. The discussion culminates with the Metropolis–Hastings algorithm, which can be used with (practically) any density. The chapter then turns to the question of whether and how a sequence of draws can be taken that provides a better approximation to the relevant integral than a purely random sequence. We discuss antithetics, systematic sampling, and Halton sequences and show the value that these types of draws provide in estimation of model parameters.

9.2 Random Draws

9.2.1. Standard Normal and Uniform

If the researcher wants to take a draw from a standard normal density (that is, a normal with zero mean and unit variance) or a standard

uniform density (uniform between 0 and 1), the process from a programming perspective is very easy. Most statistical packages contain random number generators for these densities. The researcher simply calls these routines to obtain a sequence of random draws. In the sections below, we refer to a draw of a standard normal as η and a draw of a standard uniform as μ.

The draws from these routines are actually *pseudo-random* numbers, because nothing that a computer does is truly random. There are many issues involved in the design of these routines. The intent in their design is to produce numbers that exhibit the properties of random draws. The extent to which this intent is realized depends, of course, on how one defines the properties of "random" draws. These properties are difficult to define precisely, since randomness is a theoretical concept that has no operational counterpart in the real world. From a practical perspective, my advice is the following: unless one is willing to spend considerable time investigating and resolving (literally, re-solving) these issues, it is probably better to use the available routines rather than write a new one.

9.2.2. Transformations of Standard Normal

Some random variables are transformations of a standard normal. For example, a draw from a normal density with mean b and variance s^2 is obtained as $\varepsilon = b + s\eta$. A draw from a lognormal density is obtained by exponentiating a draw from a normal density: $\varepsilon = e^{b+s\eta}$. The moments of the lognormal are functions of the mean and variance of the normal that is exponentiated. In particular, the mean of ε is $\exp(b + (s^2/2))$, and its variance is $\exp(2b + s^2) \cdot (\exp(s^2) - 1)$. Given values for the mean and variance of the lognormal, the appropriate values of b and s to use in the transformation can be calculated. It is more common, however, to treat b and s as the parameters of the lognormal and calculate its mean and variance from these parameters.

9.2.3. Inverse Cumulative for Univariate Densities

Consider a random variable with density $f(\varepsilon)$ and corresponding cumulative distribution $F(\varepsilon)$. If F is invertible (that is, if F^{-1} can be calculated), then draws of ε can be obtained from draws of a standard uniform. By definition, $F(\varepsilon) = k$ means that the probability of obtaining a draw equal to or below ε is k, where k is between zero and one. A draw μ from the standard uniform provides a number between zero and one. We can set $F(\varepsilon) = \mu$ and solve for the corresponding ε: $\varepsilon = F^{-1}(\mu)$.

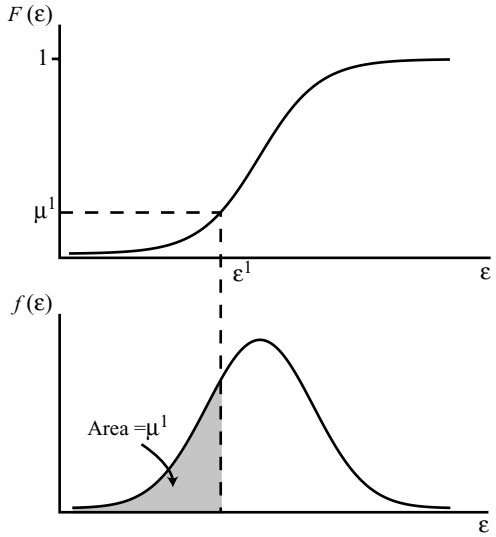

Figure 9.1. Draw of μ^1 from uniform and create $\varepsilon^1 = F^{-1}(\mu)$.

When ε is drawn in this way, the cumulative distribution of the draws is equal to F, such that the draws are equivalent to draws directly from F. An illustration is provided in Figure 9.1. A draw μ^1 from a standard uniform translates into the value of ε labeled ε^1, at which $F(\varepsilon^1) = \mu^1$.

The extreme value distribution, which is the basis for multinomial logit models, provides an example. The density is $f(\varepsilon) = \exp(-\varepsilon) \cdot \exp(-\exp(-\varepsilon))$ with cumulative distribution $F(\varepsilon) = \exp(-\exp(-\varepsilon))$. A draw from this density is obtained as $\varepsilon = -\ln(-\ln \mu)$.

Note that this procedure works only for univariate distributions. If there are two or more elements of ε, then $F^{-1}(\mu)$ is not unique, since various combinations of the elements of ε have the same cumulative probability.

9.2.4. Truncated Univariate Densities

Consider a random variable that ranges from a to b with density proportional to $f(\varepsilon)$ within this range. That is, the density is $(1/k)f(\varepsilon)$ for $a \leq \varepsilon \leq b$, and 0 otherwise, where k is the normalizing constant that insures that the density integrates to 1: $k = \int_a^b f(\varepsilon)\, d\varepsilon = F(b) - F(a)$. A draw from this density can be obtained by applying the procedure in Section 9.2.3 while assuring that the draw is within the appropriate range.

Draw μ from a standard uniform density. Calculate the weighted average of $F(a)$ and $F(b)$ as $\bar{\mu} = (1 - \mu)F(a) + \mu F(b)$. Then calculate

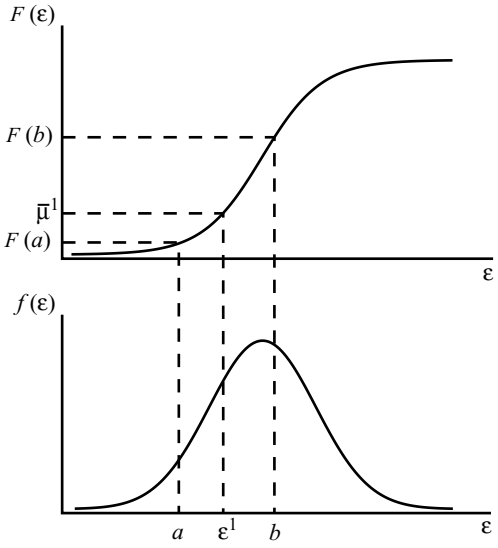

Figure 9.2. Draw of $\bar{\mu}^1$ between $F(a)$ and $F(b)$ gives draw ε^1 from $f(\varepsilon)$ between a and b.

$\varepsilon = F^{-1}(\bar{\mu})$. Since $\bar{\mu}$ is between $F(a)$ and $F(b)$, ε is necessarily between a and b. Essentially, the draw of μ determines how far to go between a and b. Note that the normalizing constant k is not used in the calculations and therefore need not be calculated. Figure 9.2 illustrates the process.

9.2.5. Choleski Transformation for Multivariate Normals

As described in Section 9.2.2, a univariate normal with mean b and variance s^2 is obtained as $\varepsilon = b + s\mu$, where μ is standard normal. An analogous procedure can be used to draw from a multivariate normal. Let ε be a vector with K elements distributed $N(b, \Omega)$. A Choleski factor of Ω is defined as a lower-triangular matrix L such that $\mathrm{LL}' = \Omega$. It is often called the generalized square root of Ω or generalized standard deviation of ε. With $K = 1$ and variance s^2, the Choleski factor is s, which is just the standard deviation of ε. Most statistical and matrix manipulation packages have routines to calculate a Choleski factor for any positive definite, symmetric matrix.

A draw of ε from $N(b, \Omega)$ is obtained as follows. Take K draws from a standard normal, and label the vector of these draws $\eta = \langle \eta_1, \ldots, \eta_K \rangle'$. Calculate $\varepsilon = b + L\eta$. We can verify the properties of ε. It is normally distributed, since the sum of normals is normal. Its

mean is b: $E(\varepsilon) = b + LE(\eta) = b$. And its covariance is Ω: $\text{Var}(\varepsilon) = E(L\eta(\eta L)') = LE(\eta\eta')L' = L\text{Var}(\eta)L' = LIL' = LL' = \Omega$.

To be concrete, consider a three-dimensional ε with zero mean. A draw of ε is calculated as

$$\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{pmatrix} = \begin{pmatrix}s_{11} & 0 & 0\\ s_{21} & s_{22} & 0\\ s_{31} & s_{32} & s_{33}\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix},$$

or

$$\begin{aligned}\varepsilon_1 &= s_{11}\eta_1,\\ \varepsilon_2 &= s_{21}\eta_1 + s_{22}\eta_2,\\ \varepsilon_3 &= s_{31}\eta_1 + s_{32}\eta_2 + s_{33}\eta_3\,.\end{aligned}$$

From this we see that $\text{Var}(\varepsilon_1) = s_{11}^2$, $\text{Var}(\varepsilon_2) = s_{21}^2 + s_{22}^2$, and $\text{Var}(\varepsilon_3) = s_{31}^2 + s_{32}^2 + s_{33}^2$. Also, $\text{Cov}(\varepsilon_1, \varepsilon_2) = s_{11}s_{21}$, and so on. The elements ε_1 and ε_2 are correlated because of the common influence of η_1 on both of them. They are not perfectly correlated because η_2 enters ε_2 without affecting ε_1. Similar analysis applies to ε_1 and ε_3, and ε_2 and ε_3. Essentially, the Choleski factor expresses K correlated terms as arising from K independent components, with each component *loading* differently onto each term. For any pattern of covariance, there is some set of loadings from independent components that reproduces that covariance.

9.2.6. Accept–Reject for Truncated Multivariate Densities

The procedure in Section 9.2.4 for drawing from truncated densities applies only to univariate distributions. With multivariate densities, drawing from a truncated support is more difficult. We describe an accept–reject procedure that can always be applied. However, as we will see, there are disadvantages of the approach that might cause a researcher to choose another approach when possible.

Suppose we want to draw from multivariate density $g(\varepsilon)$ within the range $a \le \varepsilon \le b$ where a and b are vectors with the same length as ε. That is, we want to draw from $f(\varepsilon) = \frac{1}{k}g(\varepsilon)$ if $a \le \varepsilon \le b$, and equal zero otherwise, where k is the normalizing constant. We can obtain draws from f by simply drawing from g and retaining ("accepting") the draws that are within the relevant range and discarding ("rejecting") the draws that are outside the range. The advantage of this procedure is that it can be applied whenever it is possible to draw from the untruncated density. Importantly, the normalizing constant, k, does not need to be known for the truncated density. This fact is useful because the normalizing constant is usually difficult to calculate.

The disadvantage of the procedure is that the number of draws that are accepted (that is, the number of draws from f that are obtained) is not fixed but rather is itself random. If R draws are taken from g, then the expected number of accepts is kR. This expected number is not known without knowing k, which, as stated, is usually difficult to calculate. It is therefore hard to determine an appropriate number of draws to take from g. More importantly, the actual number of accepted draws will generally differ from the expected number. In fact, there is a positive probability of obtaining no accepts from a fixed number of draws. When the truncation space is small (or, more precisely, when k is small), obtaining no accepts, and hence no draws from the truncated density, is a likely event.

This difficulty can be circumvented by drawing from g until a certain number of accepted draws is obtained. That is, instead of setting in advance the number of draws from g that will be taken, the researcher can set the number of draws from f that are obtained. Of course, the researcher will not know how long it will take to attain the set number.

In most situations, other procedures can be applied more easily to draw from a multivariate truncated density. Nevertheless, it is important to remember that, when nothing else seems possible with a truncated distribution, the accept–reject procedure can be applied.

9.2.7. Importance Sampling

Suppose ε has a density $f(\varepsilon)$ that cannot be easily drawn from by the other procedures. Suppose further that there is another density, $g(\varepsilon)$, that can easily be drawn from. Draws from $f(\varepsilon)$ can be obtained as follows. Take a draw from $g(\varepsilon)$ and label it ε^1. Weight the draw by $f(\varepsilon^1)/g(\varepsilon^1)$. Repeat this process many times. The set of weighted draws is equivalent to a set of draws from f.

To verify this fact, we show that the cumulative distribution of the weighted draws from g is the same as the cumulative distribution of draws from f. Consider the share of draws from g that are below some value m, with each draw weighted by f/g. This share is

$$\begin{aligned}\int \frac{f(\varepsilon)}{g(\varepsilon)} I(\varepsilon < m) g(\varepsilon)\, d\varepsilon &= \int_{-\infty}^{m} \frac{f(\varepsilon)}{g(\varepsilon)} g(\varepsilon)\, d\varepsilon \\ &= \int_{-\infty}^{m} f(\varepsilon)\, d\varepsilon = F(m)\,.\end{aligned}$$

In simulation, draws from a density are used to calculate the average of a statistic over that density. Importance sampling can be seen as a change in the statistic and a corresponding change in the density that

makes the density easy to draw from. Suppose we want to calculate $\int t(\varepsilon)f(\varepsilon)\,d\varepsilon$, but find it hard to draw from f. We can multiply the integrand by $g \div g$ without changing its value, so that the integral is $\int t(\varepsilon)[f(\varepsilon)/g(\varepsilon)]g(\varepsilon)\,d\varepsilon$. To simulate the integral, we take draws from g, calculate $t(\varepsilon)[f(\varepsilon)/g(\varepsilon)]$ for each draw, and average the results. We have simply transformed the integral so that it is easier to simulate.

The density f is called the target density, and g is called the proposal density. The requirements for importance sampling are that (1) the support of $g(\varepsilon)$ needs to cover the support of f, so that any ε that could arise with f can also arise with g, and (2) the ratio $f(\varepsilon)/g(\varepsilon)$ must be finite for all values of ε, so that this ratio can always be calculated.

A useful illustration of importance sampling arises with multivariate truncated normals. Suppose we want to draw from $N(0, \Omega)$ but with each element being positive (i.e., truncated below at zero). The density is

$$f(\varepsilon) = \frac{1}{k(2\pi)^{\frac{1}{2}K}|\Omega|^{1/2}} e^{-\frac{1}{2}\varepsilon'\Omega^{-1}\varepsilon}$$

for $\varepsilon \geq 0$, and 0 otherwise, where K is the dimension of ε and k is the normalizing constant. (We assume for the purposes of this example that k is known. In reality, calculating k might itself take simulation.) Drawing from this density is difficult, because the elements of ε are correlated as well as truncated. However, we can use the procedure in Section 9.2.4 to draw independent truncated normals and then apply importance sampling to create the correlation. Draw K univariate normals truncated below at zero, using the procedure in Section 9.2.4. These draws collectively constitute a draw of a K-dimensional vector ε from the positive quadrant support with density

$$g(\varepsilon) = \frac{1}{m(2\pi)^{\frac{1}{2}K}} e^{-\frac{1}{2}\varepsilon'\varepsilon},$$

where $m = 1/2^K$. For each draw, assign the weight

$$\frac{f(\varepsilon)}{g(\varepsilon)} = \frac{m}{k}|\Omega|^{-1/2} e^{\varepsilon'(\Omega^{-1}-I)\varepsilon}.$$

The weighted draws are equivalent to draws from $N(0, \Omega)$ truncated below at zero.

As a sidelight, note that the accept–reject procedure in Section 9.2.6 is a type of importance sampling. The truncated distribution is the target, and the untruncated distribution is the proposal density. Each draw from the untruncated density is weighted by a constant if the draw is

within the truncation space and weighted by zero if the draw is outside the truncation space. Weighting by a constant or zero is equivalent to weighting by one (accept) or zero (reject).

9.2.8. Gibbs Sampling

For multinomial distributions, it is sometimes difficult to draw directly from the joint density and yet easy to draw from the conditional density of each element given the values of the other elements. Gibbs sampling (the term was apparently introduced by Geman and Geman, 1984) can be used in these situations. A general explanation is provided by Casella and George, (1992), which the reader can use to supplement the more concise description that I give in the following.

Consider two random variables ε_1 and ε_2. Generalization to higher dimension is obvious. The joint density is $f(\varepsilon_1, \varepsilon_2)$, and the conditional densities are $f(\varepsilon_1|\varepsilon_2)$ and $f(\varepsilon_2|\varepsilon_1)$. Gibbs sampling proceeds by drawing iteratively from the conditional densities: drawing ε_1 conditional on a value of ε_2, drawing ε_2 conditional on this draw of ε_1, drawing a new ε_1 conditional on the new value of ε_2, and so on. This process converges to draws from the joint density.

To be more precise: (1) Choose an initial value for ε_1, called ε_1^0. Any value with nonzero density can be chosen. (2) Draw a value of ε_2, called ε_2^0, from $f(\varepsilon_2|\varepsilon_1^0)$. (3) Draw a value of ε_1, called ε_1^1, from $f(\varepsilon_1|\varepsilon_2^0)$. (4) Draw ε_2^1 from $f(\varepsilon_2|\varepsilon_1^1)$, and so on. The values of ε_1^t from $f(\varepsilon_1|\varepsilon_2^{t-1})$ and the values of ε_2^t from $f(\varepsilon_2|\varepsilon_1^{t-1})$ constitute a sequence in t. For sufficiently large t (that is, for sufficiently many iterations), the sequence converges to draws from the joint density $f(\varepsilon_1, \varepsilon_2)$.

As an example, consider two standard normal deviates that are independent except that they are truncated on the basis of their sum: $\varepsilon_1 + \varepsilon_2 \leq m$. Figure 9.3 depicts the truncated density. The circles are contours of the untruncated density, and the shaded area represents the truncated density. To derive the conditional densities, consider first the untruncated normals. Since the two deviates are independent, the conditional density of each is the same as its unconditional density. That is, ignoring truncation, $\varepsilon_1|\varepsilon_2 \sim N(0, 1)$. The truncation rule is $\varepsilon_1 + \varepsilon_2 \leq m$ which can be re-expressed as $\varepsilon_1 \leq m - \varepsilon_2$. Therefore, $\varepsilon_1|\varepsilon_2$ is distributed as a univariate standard normal truncated from above at $m - \varepsilon_2$. Given ε_2, a draw of ε_1 is obtained with the procedure in Section 9.2.4: $\varepsilon_1 = \Phi^{-1}(\mu\Phi(m - \varepsilon_2))$, where μ is a standard uniform draw and $\Phi(\cdot)$ is the cumulative standard normal distribution. Draws from ε_2 conditional on ε_1 are obtained analogously. Drawing sequentially from these conditional densities eventually provides draws from the joint truncated density.

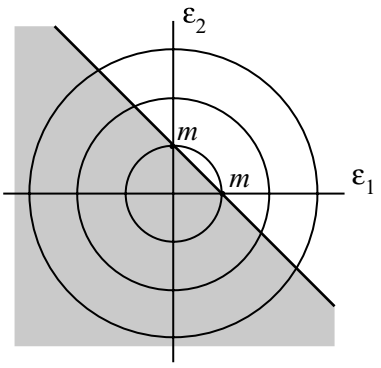

Figure 9.3. Truncated normal density.

9.2.9. Metropolis–Hastings Algorithm

If all else fails, the Metropolis–Hastings (MH) algorithm can be used to obtain draws from a density. Initially developed by Metropolis *et al.* (1953) and generalized by Hastings (1970), the MH algorithm operates as follows. The goal is to obtain draws from $f(\varepsilon)$.

1. Start with a value of the vector ε, labeled ε^0.
2. Choose a trial value of ε^1 as $\tilde{\varepsilon}^1 = \varepsilon^0 + \eta$, where η is drawn from a distribution $g(\eta)$ that has zero mean. Usually a normal distribution is specified for $g(\eta)$.
3. Calculate the density at the trial value $\tilde{\varepsilon}^1$, and compare it with the density at the original value ε^0. That is, compare $f(\tilde{\varepsilon}^1)$ with $f(\varepsilon^0)$. If $f(\tilde{\varepsilon}^1) > f(\varepsilon^0)$, then accept $\tilde{\varepsilon}^1$, label it ε^1, and move to step 4. If $f(\tilde{\varepsilon}^1) \leq f(\varepsilon^0)$, then accept $\tilde{\varepsilon}^1$ with probability $f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$, and reject it with probability $1 - f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$. To determine whether to accept or reject $\tilde{\varepsilon}^1$ in this case, draw a standard uniform μ. If $\mu \leq f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$, then keep $\tilde{\varepsilon}^1$. Otherwise, reject $\tilde{\varepsilon}^1$. If $\tilde{\varepsilon}^1$ is accepted, then label it ε^1. If $\tilde{\varepsilon}^1$ is rejected, then use ε^0 as ε^1.
4. Choose a trial value of ε^2 as $\tilde{\varepsilon}^2 = \varepsilon^1 + \eta$, where η is a new draw from $g(\eta)$.
5. Apply the rule in step 3 to either accept $\tilde{\varepsilon}^2$ as ε^2 or reject $\tilde{\varepsilon}^2$ and use ε^1 as ε^2.
6. Continue this process for many iterations. The sequence ε^t becomes equivalent to draws from $f(\varepsilon)$ for sufficiently large t.

The draws are serially correlated, since each draw depends on the previous draw. In fact, when a trial value is rejected, the current draw is the

same as the previous draw. This serial correlation needs to be considered when using these draws.

The MH algorithm can be applied with any density that can be calculated. The algorithm is particularly useful when the normalizing constant for a density is not known or cannot be easily calculated. Suppose that we know that ε is distributed proportional to $f^*(\varepsilon)$. This means that the density of ε is $f(\varepsilon) = \frac{1}{k} f^*(\varepsilon)$, where the normalizing constant $k = \int f^*(\varepsilon)\, d\varepsilon$ assures that f integrates to 1. Usually k cannot be calculated analytically, for the same reason that we need to simulate integrals in other settings. Luckily, the MH algorithm does not utilize k. A trial value of ε^t is tested by first determining whether $f(\tilde{\varepsilon}^t) > f(\varepsilon^{t-1})$. This comparison is unaffected by the normalizing constant, since the constant enters the denominator on both sides. Then, if $f(\tilde{\varepsilon}^t) \leq f(\varepsilon^{t-1})$, we accept the trial value with probability $f(\tilde{\varepsilon}^t)/f(\varepsilon^{t-1})$. The normalizing constant drops out of this ratio.

The MH algorithm is actually more general than I describe here, though in practice it is usually applied as I describe. Chib and Greenberg, (1995) provide an excellent description of the more general algorithm as well as an explanation of why it works. Under the more general definition, Gibbs sampling is a special case of the MH algorithm, as Gelman, (1992) pointed out. The MH algorithm and Gibbs sampling are often called Markov chain Monte Carlo (MCMC, or MC-squared) methods; a description of their use in econometrics is provided by Chib and Greenberg (1996). The draws are Markov chains because each value depends only on the immediately preceding one, and the methods are Monte Carlo because random draws are taken. We explore further issues about the MH algorithm, such as how to choose $g(\varepsilon)$, in the context of its use with hierarchical Bayes procedures (in Chapter 12).

9.3 Variance Reduction

The use of independent random draws in simulation is appealing because it is conceptually straightforward and the statistical properties of the resulting simulator are easy to derive. However, there are other ways to take draws that can provide greater accuracy for a given number of draws. We examine these alternative methods in the following sections.

Recall that the objective is to approximate an integral of the form $\int t(\varepsilon) f(\varepsilon)\, d\varepsilon$. In taking a sequence of draws from the density $f(\cdot)$, two issues are at stake: coverage and covariance. Consider coverage first. The integral is over the entire density f. It seems reasonable that a more accurate approximation would be obtained by evaluating $t(\varepsilon)$ at values of

ε that are spread throughout the domain of f. With independent random draws, it is possible that the draws will be clumped together, with no draws from large areas of the domain. Procedures that guarantee better coverage can be expected to provide a better approximation.

Covariance is another issue. With independent draws, the covariance over draws is zero. The variance of a simulator based on R independent draws is therefore the variance based on one draw divided by R. If the draws are negatively correlated instead of independent, then the variance of the simulator is lower. Consider $R = 2$. The variance of $\check{t} = [t(\varepsilon_1) + t(\varepsilon_2)]/2$ is $[V(t(\varepsilon_1) + V(t(\varepsilon_2) + 2\,\mathrm{Cov}(t(\varepsilon_1), t(\varepsilon_2))]/4$. If the draws are independent, then the variance is $V(t(\varepsilon_r))/2$. If the two draws are negatively correlated with each other, the covariance term is negative and the variance becomes less than $V(t(\varepsilon_r))/2$. Essentially, when the draws are negatively correlated within an unbiased simulator, a value above $\bar{t} = E_r(t(\varepsilon))$ for one draw will tend to be associated with a value for the next draw that is below $E_r(t(\varepsilon))$, such that their average is closer to the true value $\bar{t}$.

The same concept arises when simulators are summed over observations. For example, the simulated log-likelihood function is a sum over observations of the log of simulated probabilities. If the draws for each observation's simulation are independent of the draws for the other observations, then the variance of the sum is simply the sum of the variances. If the draws are taken in a way that creates negative correlation over observations, then the variance of the sum is lower.

For a given observation, the issue of covariance is related to coverage. By inducing a negative correlation between draws, better coverage is usually assured. With $R = 2$, if the two draws are taken independently, then both could end up being at the low side of the distribution. If negative correlation is induced, then the second draw will tend to be high if the first draw is low, which provides better coverage.

We describe below methods to attain better coverage for each observation's integral and to induce negative correlation over the draws for each observation as well as over observations. We assume for the sake of discussion that the integral is a choice probability and that the sum over observations is the simulated log-likelihood function. However, the concepts apply to other integrals, such as scores, and to other sums, such as moment conditions and market shares. Also, unless otherwise noted, we illustrate the methods with only two random terms so that the draws can be depicted graphically. The random terms are labeled ε^a and ε^b, and collectively as $\varepsilon = \langle \varepsilon^a, \varepsilon^b \rangle'$. A draw of ε from its density $f(\varepsilon)$ is denoted $\varepsilon_r = \langle \varepsilon_r^a, \varepsilon_r^b \rangle'$ for $r = 1, \ldots, R$. Thus, ε_3^a, for example, is the third draw of the first random term.

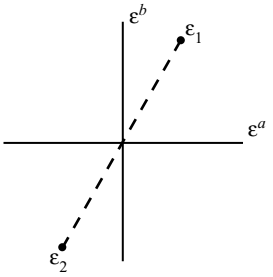

Figure 9.4. Reverse sign of both elements.

9.3.1. Antithetics

Antithetic draws, suggested by Hammersley and Morton (1956), are obtained by creating various types of mirror images of a random draw. For a symmetric density that is centered on zero, the simplest antithetic variate is created by reversing the sign of all elements of a draw. Figure 9.4 illustrates. Suppose a random draw is taken from $f(\varepsilon)$ and the value $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$ is obtained. The second "draw," which is called the antithetic of the first draw, is created as $\varepsilon_2 = \langle -\varepsilon_1^a, -\varepsilon_1^b \rangle'$. Each draw from f creates a pair of "draws," the original draw and its mirror image (mirrored through the origin). To obtain a total of R draws, $R/2$ draws are taken independently from f and the other $R/2$ are created as the negative of the original draws.

When the density is not centered on zero, the same concept is applied but through a different process. For example, the standard uniform density is between 0 and 1, centered on 0.5. A draw is taken, labeled μ_1, and its antithetic variate is created as $\mu_2 = 1 - \mu_1$. The variate is the same distance from 0.5 as the original draw, but on the other side of 0.5. In general, for any univariate density with cumulative function $F(\varepsilon)$, the antithetic of a draw ε is created as $F^{-1}(1 - F(\varepsilon))$. In the case of a symmetric density centered on zero, this general formula is equivalent to simply reversing the sign. In the remaining discussion we assume that the density is symmetric and centered on zero, which makes the concepts easier to express and visualize.

The correlation between a draw and its antithetic variate is exactly -1, so that the variance of their sum is zero: $V(\varepsilon_1 + \varepsilon_2) = V(\varepsilon_1) + V(\varepsilon_2) + 2\,\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0$. This fact does not mean that there is no variance in the simulated probability that is based on these draws. The simulated probability is a nonlinear function of the random terms, and so the correlation between $P(\varepsilon_1)$ and $P(\varepsilon_2)$ is less than one. The variance of the simulated probability $\breve{P} = \frac{1}{2}[P(\varepsilon_1) + P(\varepsilon_2)]$ is greater than zero. However, the

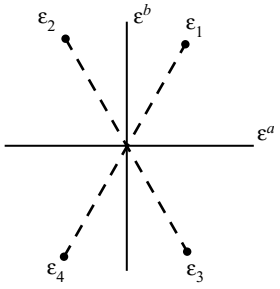

Figure 9.5. Reverse sign of each element, then of both.

variance of the simulated probabilities is less than $\frac{1}{2}V_r(P(\varepsilon_r))$, which is the variance with two independent draws.

As shown in Figure 9.4, reversing the sign of a draw gives evaluation points in opposite quadrants. The concept can be extended to obtain draws in each quadrant. A draw is taken, and then antithetic draws are created by reversing the sign of each element alone (leaving the sign of the other elements unchanged), reversing the sign of each pair of elements, each triplet of elements, and so on. For ε with two elements, this process creates three antithetic draws for each independent draw. For $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$, the antithetic draws are

$$\varepsilon_2 = \left\langle -\varepsilon_1^a, \varepsilon_1^b \right\rangle',$$
$$\varepsilon_3 = \left\langle \varepsilon_1^a, -\varepsilon_1^b \right\rangle',$$
$$\varepsilon_4 = \left\langle -\varepsilon_1^a, -\varepsilon_1^b \right\rangle'.$$

These draws are shown in Figure 9.5. Each quadrant contains a draw.

Better coverage and higher negative correlation can be obtained by shifting the position of each element as well as reversing their signs. In Figure 9.5, ε_1 and ε_2 are fairly close together, as are ε_3 and ε_4. This placement leaves large uncovered areas between ε_1 and ε_3 and between ε_2 and ε_4. Orthogonal draws with even placement can be obtained by switching element ε_1^a with ε_1^b while also reversing the signs. The antithetic draws are

$$\varepsilon_2 = \left\langle -\varepsilon_1^b, \varepsilon_1^a \right\rangle',$$
$$\varepsilon_3 = \left\langle \varepsilon_1^b, -\varepsilon_1^a \right\rangle',$$
$$\varepsilon_4 = \left\langle -\varepsilon_1^a, -\varepsilon_1^b \right\rangle',$$

which are illustrated in Figure 9.6. These concepts can, of course, be extended to any number of dimensions. For M-dimensional ε, each random draw creates 2^M antithetic draws (including the original one), with one in each quadrant.

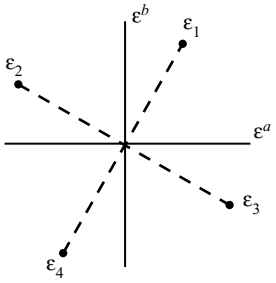

Figure 9.6. Switch positions and reverse signs.

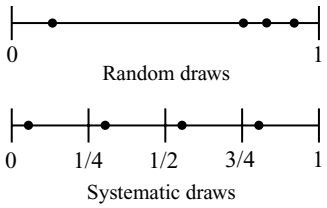

Figure 9.7. Draws from standard uniform.

Comparisons performed by Vijverberg (1997) and Sándor and András (2001) show that antithetics substantially improve the estimation of probit models. Similarly, Geweke (1988) has shown their value when calculating statistics based on Bayesian posteriors.

9.3.2. Systematic Sampling

Coverage can also be improved through systematic sampling (McGrath, 1970), which creates a grid of points over the support of the density and randomly shifts the entire grid. Consider draws from a uniform distribution between 0 and 1. If four draws are taken independently, the points may look like those in the top part of Figure 9.7, which provide fairly poor coverage. Instead, the unit interval is divided into four segments and draws taken in a way that assures one draw in each segment with equal distance between the draws. Take a draw from a uniform between 0 and 0.25 (by drawing from a standard uniform and dividing the result by 4). Label the draw ε_1. Three other draws are created as

$$\varepsilon_2 = 0.25 + \varepsilon_1,$$
$$\varepsilon_3 = 0.50 + \varepsilon_1,$$
$$\varepsilon_4 = 0.75 + \varepsilon_1.$$

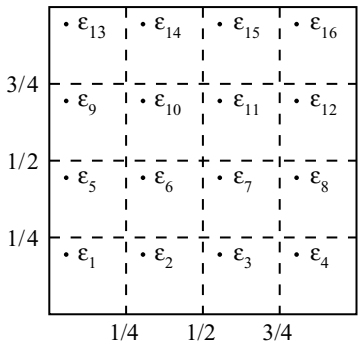

Figure 9.8. Systematic draws in two dimensions.

These draws look like those in the bottom part of Figure 9.7, which provide better coverage than independent draws.

The issue arises of how finely to segment the interval. For example, to obtain a total of 100 draws, the unit interval can be divided into 100 segments. A draw between 0 and 0.01 is taken, and then the other 99 draws are created from this one draw. Instead, the unit interval can be divided into fewer than 100 draws and more independent draws taken. If the interval is divided into four segments, then 25 independent draws are taken between 0 and 0.25, and three draws in the other segments are created for each of the independent draws. There is a tradeoff that the researcher must consider in deciding how fine a grid to use in systematic sampling. More segments provide more even coverage for a given total number of draws. However, fewer segments provide more randomness to the process. In our example with $R = 100$, there is only one random draw when 100 segments are used, whereas there are 25 random draws when four segments are used.

The randomness of simulation draws is a necessary component in the derivation of the asymptotic properties of the simulation-based estimators, as described in Chapter 10. Many of the asymptotic properties rely on the concept that the number of random draws increases without bound with sample size. The asymptotic distributions become relatively accurate only when enough random draws have been taken. Therefore, for a given total number of draws, the goal of better coverage, which is attained with a more finely defined segmentation, needs to be traded off against the goal of having enough randomness for the asymptotic formulas to apply, which is attained with a more coarsely defined segmentation. The same issue applies to the antithetics discussed earlier.

Systematic sampling can be performed in multiple dimensions. Consider a two-dimensional uniform on the unit square. A grid is created by dividing each dimension into segments. As shown in Figure 9.8,

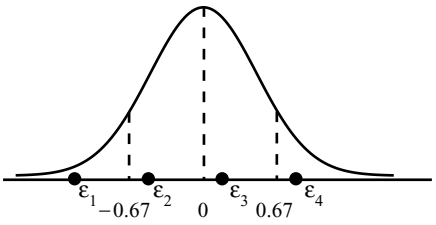

Figure 9.9. Systematic draws for univariate normal.

when each dimension is divided into four segments, the unit square is partitioned into 16 areas. A draw between 0 and 0.25 is taken for each element, giving $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$, where $0 < \varepsilon_1^a < 0.25$ and $0 < \varepsilon_1^b < 0.25$. This draw falls somewhere in the bottom-left area in Figure 9.8. Fifteen other draws are then created as the "origin" of each area, plus $\langle \varepsilon_1^a, \varepsilon_1^b \rangle'$. For example, the point that is created for the bottom-right area is $\varepsilon_4 = \langle (0.75 + \varepsilon_1^a), (0 + \varepsilon_1^b) \rangle'$.

These draws are defined for a uniform distribution. When f represents another density, the points are transformed using the method described in Section 9.2.3. In particular, let F be the cumulative distribution associated with univariate density f. Systematic draws from f are created by transforming each systematic draw from a uniform by F^{-1}. For example, for a standard normal, four equal-sized segments of the density are created with breakpoints: $\Phi^{-1}(0.25) = -0.67$, $\Phi^{-1}(0.5) = 0$, and $\Phi^{-1}(0.75) = 0.67$. As shown in Figure 9.9, these segments are equal-sized in the sense that each contains the same mass. The draws for the standard normal are created by taking a draw from a uniform between 0 and 0.25, labeled μ_1. The corresponding point on the normal is $\varepsilon_1 = \Phi^{-1}(\mu_1)$, which falls in the first segment. The points for the other three segments are created as $\varepsilon_2 = \Phi^{-1}(0.25 + \mu_1)$, $\varepsilon_3 = \Phi^{-1}(0.5 + \mu_1)$, and $\varepsilon_4 = \Phi^{-1}(0.75 + \mu_1)$.

Draws of multidimensional random terms are obtained similarly, provided that the elements are independent. For example, if ε consists of two elements each of which is standard normal, then draws analogous to those in Figure 9.8 are obtained as follows: Draw μ_1^a and μ_1^b from a uniform between 0 and 0.25. Calculate ε_1 as $\langle \Phi^{-1}(\mu_1^a), \Phi^{-1}(\mu_1^b) \rangle'$. Calculate the other 15 points as ε_r as $\langle \Phi^{-1}(x_r + \mu_1^a), \Phi^{-1}(y_r + \mu_1^b) \rangle'$, where $\langle x_r, y_r \rangle'$ is the origin of area r in the unit square.

The requirement that the elements of ε be independent is not restrictive. Correlated random elements are created through transformations of independent elements, such as the Choleski transformation. The independent elements are drawn from their density, and then the correlation is created inside the model.

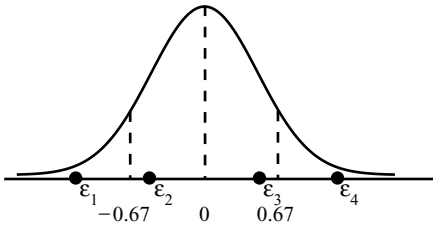

Figure 9.10. Symmetric systematic draws.

Obviously, numerous sets of systemtically sampled draws can be obtained to gain more randomization. In two dimensions with four segments in each dimension, 64 draws are obtained by taking 4 independent draws in the 0-to-$\frac{1}{4}$ square and creating 15 other draws from each. This procedure provides greater randomization but less fine coverage than defining the draws in terms of eight segments in each dimension such that each random draw in the 0 to $\frac{1}{8}$ square translates into 64 systematic draws.

The draws for the normal distribution that are created as just described are not symmetric around zero. An alternative approach can be used to assure such symmetry. For a unidimensional normal, 4 draws that are symmetric around zero are obtained as follows. Draw a uniform between 0 and 0.25, labeled μ_1. Create the draw from the normal as $\varepsilon_1 = \Phi^{-1}(\mu_1)$. Create the draw for the second segment as $\varepsilon_2 = \Phi^{-1}(0.25 + \mu_1)$. Then create the draws for the third and fourth segments as the negative of these draws: $\varepsilon_3 = -\varepsilon_2$ and $\varepsilon_4 = -\varepsilon_1$. Figure 9.10 illustrates the draws using the same μ_1 as for Figure 9.9. This procedure combines systematic sampling with antithetics. It can be extended to multiple dimensions by creating systematic draws for the positive quadrant and then creating antithetic variates for the other quadrants.

9.3.3. Halton Sequences

Halton sequences (Halton, 1960) provide coverage and, unlike the other methods we have discussed, induce a negative correlation over observations. A Halton sequence is defined in terms of a given number, usually a prime. The sequence is most easily understood though an example. Consider the prime 3. The Halton sequence for 3 is created by dividing the unit interval into three parts with breaks at $\frac{1}{3}$ and $\frac{2}{3}$, as shown in the top panel of Figure 9.11. The first terms in the sequence are these breakpoints: $\frac{1}{3}$, $\frac{2}{3}$. Then each of the three segments is divided into thirds, and the breakpoints for these segments are added to the sequences in a

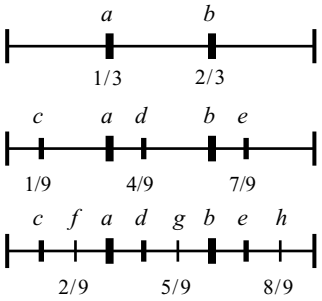

Figure 9.11. Halton sequence for prime 3.

particular way. The sequence becomes $\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}$. Note that the lower breakpoints in all three segments ($\frac{1}{9}, \frac{4}{9}, \frac{7}{9}$) are entered in the sequence before the higher breakpoints ($\frac{2}{9}, \frac{5}{9}, \frac{8}{9}$.) Then each of the nine segments is divided into thirds, with the breakpoints added to the sequences. The sequence becomes $\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \frac{1}{27}, \frac{10}{27}, \frac{19}{27}, \frac{4}{27}, \frac{13}{27}$, and so on. This process is continued for as many points as the researcher wants to obtain.

From a programming perspective, it is easy to create a Halton sequence. The sequence is created iteratively. At each iteration t, the sequence is denoted s^t, which is a series of numbers. The sequence is extended in each iteration with the new sequence being $s_{t+1} = \{s_t, s_t + 1/3^t, s_t + 2/3^t\}$. Start with 0 as the initial sequence: $s_0 = \{0\}$. The number zero is not actually part of a Halton sequence, but considering it to be the first element facilitates creation of the sequence, as we will see. It can be dropped after the entire sequence is created. In the first iteration, add $1/3^1$ ($= \frac{1}{3}$) and then $2/3^1$ ($= \frac{2}{3}$) to this element and append the results, to get $\{0, \frac{1}{3}, \frac{2}{3}\}$. The sequence has three elements. In the second iteration, add $1/3^2$ ($= \frac{1}{9}$) and then $2/3^2$ ($= \frac{2}{9}$) to each element of the sequence and append the results:

$$
\begin{aligned}
0 &= 0,\\
1/3 &= 1/3,\\
2/3 &= 2/3,\\
0 + 1/9 &= 1/9,\\
1/3 + 1/9 &= 4/9,\\
2/3 + 1/9 &= 7/9,\\
0 + 2/9 &= 2/9,\\
1/3 + 2/9 &= 5/9,\\
2/3 + 2/9 &= 8/9.
\end{aligned}
$$

The new sequence consists of nine elements.

In the third iteration, add $1/3^3$ ($= \frac{1}{27}$) and then $2/3^3$ ($= \frac{1}{27}$) to each element of this sequence and append the results:

$$
\begin{aligned}
0 &= 0,\\
1/3 &= 1/3,\\
2/3 &= 2/3,\\
1/9 &= 1/9,\\
4/9 &= 4/9,\\
7/9 &= 7/9,\\
2/9 &= 2/9,\\
5/9 &= 5/9,\\
8/9 &= 8/9,\\
0 + 1/27 &= 1/27,\\
1/3 + 1/27 &= 10/27,\\
2/3 + 1/27 &= 19/27,\\
1/9 + 1/27 &= 4/27,\\
4/9 + 1/27 &= 13/27,\\
7/9 + 1/27 &= 22/27,\\
2/9 + 1/27 &= 7/27,\\
5/9 + 1/27 &= 16/27,\\
8/9 + 1/27 &= 25/27,\\
0 + 2/27 &= 2/27,\\
1/3 + 2/27 &= 11/27,\\
2/3 + 2/27 &= 20/27,\\
1/9 + 2/27 &= 5/27,\\
4/9 + 2/27 &= 14/27,\\
7/9 + 2/27 &= 23/27,\\
2/9 + 2/27 &= 8/27,\\
5/9 + 2/27 &= 17/27,\\
8/9 + 2/27 &= 26/27.
\end{aligned}
$$

The sequence now consists of 27 elements. In the fourth iteration, add $1/3^4$ ($= \frac{1}{81}$) and then $2/3^4$ ($= \frac{2}{81}$) to each element of the sequence and append the results, and so on.

Note that the sequence cycles over the unit interval every three numbers:

$$
\begin{array}{ccc}
0 & 1/3 & 2/3\\
1/9 & 4/9 & 7/9\\
2/9 & 5/9 & 8/9\\
1/27 & 10/27 & 19/27\\
4/27 & 13/27 & 22/27\\
7/27 & 16/27 & 25/27
\end{array}
$$

2/27	11/27	20/27
5/27	14/27	23/27
8/27	17/27	26/27

Within each cycle the numbers are ascending.

Halton sequences for other prime numbers are created similarly. The sequence for 2 is $\{\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \ldots\}$. In general, the sequence for prime k is created iteratively, with the sequence at iteration $t+1$ being $s_{t+1} = \{s_t, s_t + 1/k^t, s_t + 2/k^t, \ldots, s_t + (k-1)/k^t\}$. The sequence contains cycles of length k, where each cycle consists of k ascending points on the unit interval equidistant from each other.

Since a Halton sequence is defined on the unit interval, its elements can be considered as well-placed "draws" from a standard uniform density. The Halton draws provide better coverage than random draws, on average, because they are created to progressively fill in the unit interval evenly and ever more densely. The elements in each cycle are equidistant apart, and each cycle covers the unit interval in the areas not covered by previous cycles.

When using Halton draws for a sample of observations, one long Halton sequence is usually created and then part of the sequence is used for each observation. The initial elements of the sequence are discarded for reasons we will discuss. The remaining elements are then used in groups, with each group of elements constituting the "draws" for one observation. For example, suppose there are two observations, and the researcher wants $R = 5$ draws for each. If the prime 3 is used, and the researcher decides to discard the first 10 elements, then a sequence of length 20 is created. This sequence is

0	1/3	2/3
1/9	4/9	7/9
2/9	5/9	8/9
1/27	10/27	19/27
4/27	13/27	22/27
7/27	16/27	25/27
2/27	11/27.	

After eliminating the first 10 elements, the Halton draws for the first observation are $\{\frac{10}{27}, \frac{19}{27}, \frac{4}{27}, \frac{13}{27}, \frac{22}{27}\}$ and the Halton draws for the second observation are $\{\frac{7}{27}, \frac{16}{27}, \frac{25}{27}, \frac{2}{27}, \frac{11}{27}\}$. These draws are illustrated in Figure 9.12. Note that the gaps in coverage for the first observation are filled by the draws for the second observation. For example, the large gap between $\frac{4}{27}$ and $\frac{10}{27}$ for the first observation is filled in by the midpoint of this gap, $\frac{7}{27}$, for the second observation. The gap between $\frac{13}{27}$ and $\frac{19}{27}$ is

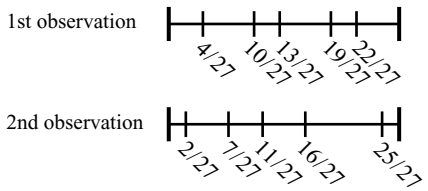

Figure 9.12. Halton draws for two observations.

filled in by its midpoint, $\frac{16}{27}$, for the second observation, and so on. The pattern by which Halton sequences are created makes them such that each subsequence fills in the gaps of the previous subsequences.

Because of this filling-in property, simulated probabilities based on Halton draws tend to be self-correcting over observations. The draws for one observation tend to be negatively correlated with those for the previous observation. In our example, the average of the first observation's draws is above 0.5, while the average of the draws for the second observation is below 0.5. This negative correlation reduces error in the simulated log-likelihood function.

When the number of draws used for each observation rises, the coverage for each observation improves. The negative covariance across observations diminishes, since there are fewer gaps in each observation's coverage to be filled in by the next observation. The self-correcting aspect of Halton draws over observations is greatest when few draws are used for each observation so that the correction is most needed. However, accuracy improves with more Halton draws, since coverage is better for each observation.

As described so far, the Halton draws are for a uniform density. To obtain a sequence of points for other univariate densities, the inverse cumulative distribution is evaluated at each element of the Halton sequence. For example, suppose the researcher wants draws from a standard normal density. A Halton sequence is created for, say, prime 3, and the inverse cumulative normal is taken for each element. The resulting sequence is

$$
\begin{aligned}
&\Phi^{-1}\tfrac{1}{3} = -0.43,\\
&\Phi^{-1}\tfrac{2}{3} = 0.43,\\
&\Phi^{-1}\tfrac{1}{9} = -1.2,\\
&\Phi^{-1}\tfrac{4}{9} = -0.14,\\
&\Phi^{-1}\tfrac{7}{9} = 0.76,\\
&\qquad\vdots
\end{aligned}
$$

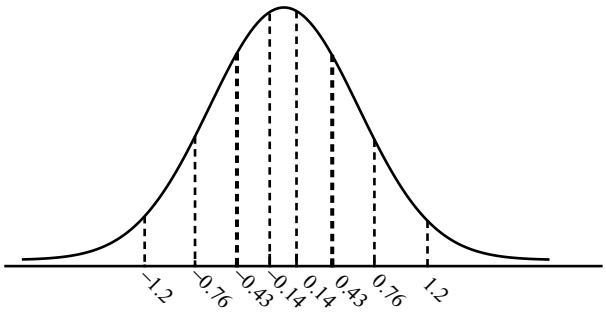

Figure 9.13. Halton draws for a standard normal.

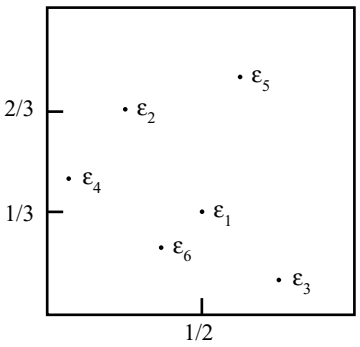

Figure 9.14. Halton sequence in two dimensions for primes 2 and 3.

This sequence is depicted in Figure 9.13. It can be considered the same as for the unit interval, as dividing the density into three segments of equal mass, with breakpoints at -0.43 and $+0.43$, and then dividing each segment into three subsegments of equal mass, and so on.

Halton sequences in multiple dimensions are obtained by creating a Halton sequence for each dimension with a different prime for each dimension. For example, a sequence in two dimensions is obtained by creating pairs from the Halton sequence for primes 2 and 3. The points are

$$\begin{aligned}
\varepsilon_1 &= \left\langle \tfrac{1}{2}, \tfrac{1}{3} \right\rangle, \\
\varepsilon_2 &= \left\langle \tfrac{1}{4}, \tfrac{2}{3} \right\rangle, \\
\varepsilon_3 &= \left\langle \tfrac{3}{4}, \tfrac{1}{9} \right\rangle, \\
\varepsilon_4 &= \left\langle \tfrac{1}{8}, \tfrac{4}{9} \right\rangle, \\
\varepsilon_5 &= \left\langle \tfrac{5}{8}, \tfrac{7}{9} \right\rangle, \\
\varepsilon_6 &= \left\langle \tfrac{3}{8}, \tfrac{2}{9} \right\rangle, \\
&\vdots
\end{aligned}$$

This sequence is depicted in Figure 9.14. To obtain draws for a

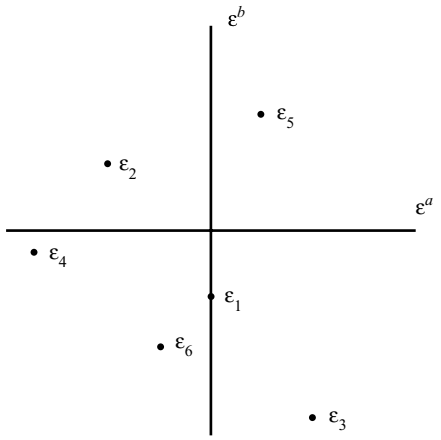

Figure 9.15. Halton sequence for two-dimensional standard normal.

two-dimensional independent standard normal, the inverse cumulative normal is taken of each element of these pairs. The draws are

$$\begin{aligned}
\varepsilon_1 &= \langle 0, -0.43 \rangle, \\
\varepsilon_2 &= \langle -0.67, 0.43 \rangle, \\
\varepsilon_3 &= \langle 0.67, -1.2 \rangle, \\
\varepsilon_4 &= \langle -1.15, -0.14 \rangle, \\
\varepsilon_5 &= \langle 0.32, 0.76 \rangle, \\
\varepsilon_6 &= \langle -.32, -0.76 \rangle, \\
&\vdots
\end{aligned}$$

which are shown in Figure 9.15.

When creating sequences in several dimensions, it is customary to eliminate the initial part of the series. The initial terms of two Halton sequences are highly correlated, through at least the first cycle of each sequence. For example, the sequences for 7 and 11 begin with $\{\frac{1}{7}, \frac{2}{7}, \frac{3}{7}, \frac{4}{7}, \frac{5}{7}, \frac{6}{7}\}$ and $\{\frac{1}{11}, \frac{2}{11}, \frac{3}{11}, \frac{4}{11}, \frac{5}{11}, \frac{6}{11}\}$. These first elements fall on a line in two dimensions, as shown in Figure 9.16. The correlation dissipates after each sequence has cycled through the unit interval, since sequences with different primes cycle at different rates. Discarding the initial part of the sequence eliminates the correlation. The number of initial elements to discard needs to be at least as large as the largest prime that is used in creating the sequences.

The potential for correlation is the reason that prime numbers are used to create the Halton sequences instead of nonprimes. If a nonprime is used, then there is a possibility that the cycles will coincide throughout the entire sequence, rather than for just the initial elements. For example,

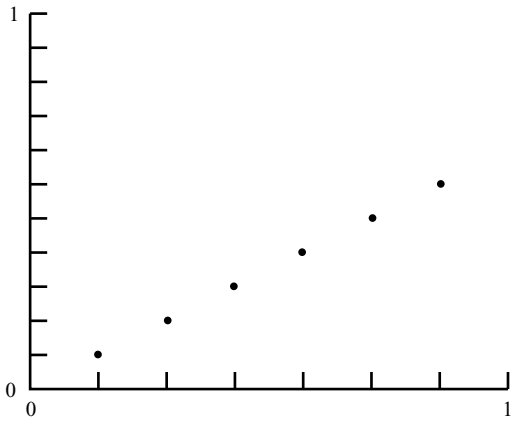

Figure 9.16. First six elements of Halton sequence for primes 7 and 11.

if Halton sequences are created for 3 and 6, the sequence for 3 cycles twice for every one cycle of the sequence for 6. Since the elements within a cycle are ascending, the elements in each cycle of the sequence for 3 are correlated with the elements in the cycle of the sequence for 6. Using only prime numbers avoids this overlapping of cycles.

The superior coverage and the negative correlation over observations that are obtained with Halton draws combine to make Halton draws far more effective than random draws for simulation. Spanier and Maize (1991) have shown that a small number of Halton draws provide relatively good integration. In the context of discrete choice models, Bhat (2001) found that 100 Halton draws provided more precise results for his mixed logit than 1000 random draws. In fact, the simulation error with 125 Halton draws was half as large as with 1000 random draws and somewhat smaller than with 2000 random draws. Train (2000), Munizaga and Alvarez-Daziano (2001), and Hensher (2001) confirm these results on other datasets.

As illustration, consider the mixed logit model that is described extensively in Chapter 11. Briefly, the model is for households' choice of electricity supplier. In a stated-preference survey, respondents were presented with a series of hypothetical choice situations. In each situation, four energy suppliers were described and the respondent was asked which company he would choose. The suppliers were differentiated on the basis of their price, whether the company required the customer to sign a long-term contract, whether the supplier was the local energy utility, whether the supplier was a well-known company, and whether the supplier offered time-of-day (TOD) or seasonal rates. A mixed logit model was estimated with these six characteristics as explanatory

Table 9.1. *Means of parameter estimates*

	1000 Random Draws	100 Halton Draws
Price	−0.8607	−0.8588
Contract length:		
Mean	−0.1955	−0.1965
Std. dev.	0.3092	0.3158
Local utility:		
Mean	2.0967	2.1142
Std. dev.	1.0535	1.0236
Known company:		
Mean	1.4310	1.4419
Std. dev.	0.8208	0.6894
TOD rates:		
Mean	−8.3760	−8.4149
Std. dev.	2.4647	2.5466
Seasonal rates:		
Mean	−8.6286	−8.6381
Std. dev.	1.8492	1.8977

variables. The coefficient of each variable was assumed to be normally distributed, except for the price coefficient, which was assumed to be fixed. The model therefore contained five random terms for simulation. A complete description of the data, the estimated model, and its implications are given in Chapter 11, where the content of the model is relevant to the topic of the chapter. For now, we are concerned only with the issue of Halton draws compared to random draws.

To investigate this issue, the model was estimated with 1000 random draws and then with 100 Halton draws. More specifically, the model was estimated five times using five different sets of 1000 random draws. The mean and standard deviation of the estimated parameters from these five runs were calculated. The model was then estimated five times with Halton sequences. The first model used the primes $2, 3, 5, 7, 11$ for the five dimensions of simulation. The order of the primes was switched for the other models, so that the dimension for which each prime was used changed in the five runs. The average and standard deviation of the five sets of estimates were then calculated.

The means of the parameter estimates over the five runs are given in Table 9.1. The mean for the runs based on random draws are given in the first column, and the means for the runs based on Halton draws are given in the second column. The two sets of means are very similar. This result indicates that the Halton draws provide the same estimates, *on average*, as random draws.

Table 9.2. *Standard deviations of parameter estimates*

	1000 Random Draws	100 Halton Draws
Price	0.0310	0.0169
Contract length:		
Mean	0.0093	0.0045
Std. dev.	0.0222	0.0108
Local utility:		
Mean	0.0844	0.0361
Std. dev.	0.1584	0.1180
Known company:		
Mean	0.0580	0.0242
Std. dev.	0.0738	0.1753
TOD rates:		
Mean	0.3372	0.1650
Std. dev.	0.1578	0.0696
Seasonal rates:		
Mean	0.4134	0.1789
Std. dev.	0.2418	0.0679

The standard deviations of the parameter estimates are given in Table 9.2. For all but one of the 11 parameters, the standard deviations are lower with 100 Halton draws than with 1000 random draws. For eight of the parameters, the standard deviations are half as large. Given that both sets of draws give essentially the same means, the lower standard deviations with the Halton draws indicate that a researcher can expect to be closer to the expected values of the estimates using 100 Halton draws than 1000 random draws.

These results show the value of Halton draws. Computer time can be reduced by a factor of ten by using Halton draws instead of random draws, without reducing, and in fact increasing, accuracy.

These results need to be viewed with caution, however. The use of Halton draws and other quasi-random numbers in simulation-based estimation is fairly new and not completely understood. For example, an anomaly arose in the analysis that serves as a warning. The model was reestimated with 125 Halton draws instead of 100. It was estimated five times under each of the five orderings of the prime numbers as described earlier. Four of the five runs provided very similar estimates. However, the fifth run gave estimates that were noticeably different from the others. For example, the estimated price coefficient for the first four runs was -0.862, -0.865, -0.863, and -0.864, respectively, while the fifth gave -0.911. The standard deviations over the five sets of estimates were

lower than with 1000 random draws, confirming the value of the Halton draws. However, the standard deviations were greater with 125 Halton draws than with 100 Halton draws, due to the last run with 125 draws providing such different results. The reason for this anomaly has not been determined. Its occurrence indicates the need for further investigation of the properties of Halton sequences in simulation-based estimation.

9.3.4. Randomized Halton Draws

Halton sequences are systematic rather than random. However, the asymptotic properties of simulation-based estimators are derived under the assumption that the draws are random. There are two ways that this issue can be addressed. First, one can realize that draws from a random number generator are not actually random either. They are systematic, like anything done by a computer. A random number generator creates draws that have many of the characteristics of truly random draws, but in fact they are only pseudorandom. In this regard, therefore, Halton draws can be seen as a systematic way of approximating integration that is more precise than using pseudorandom draws, which are also systematic. Neither matches the theoretical concept of randomness, and, in fact, it is not clear that the theoretical concept actually has a real-world counterpart. Both meet the basic underlying goal of approximating an integral over a density.

Second, Halton sequences can be transformed in a way that makes them random, at least in the same way that pseudorandom numbers are random. Bhat (forthcoming) describes the process, based on procedures introduced by Tuffin (1996):

1. Take a draw from a standard uniform density. Label this random draw μ.
2. Add μ to each element of the Halton sequence. If the resulting element exceeds 1, subtract 1 from it. Otherwise, keep the resulting element as is (without subtracting 1).

The formula for this transformation is $s_n = \text{mod}(s_o + \mu)$, where s_o is the original element of the Halton sequence, s_n is the transformed element, and mod takes the fractional part of the argument in parentheses.

The transformation is depicted in Figure 9.17. Suppose the draw of μ from the uniform density is 0.40. The number 0.33 is the first element of the Halton sequence for prime 3. This element is transformed, as shown in the top panel, to $0.33 + 0.40 = 0.73$, which is just a 0.40 move up

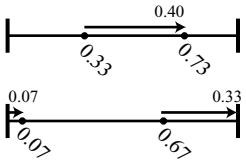

Figure 9.17. Random transformation of Halton draws with $\mu = 0.40$.

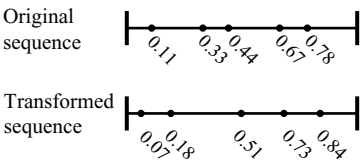

Figure 9.18. Randomization of Halton sequence in one dimension.

the line. The number 0.67 is the second element of the sequence. It is transformed by adding 0.4 and then, since the result exceeds 1, by subtracting 1 to get 0.07 ($0.67 + 0.40 - 1 = 0.07$). As shown in the bottom panel, this transformation is visualized as moving the original point up by a distance 0.40, but wrapping around when the end of the unit interval is reached. The point moves up 0.33 to where the line ends, and then wraps to the start of the line and continues to move up another 0.07, for a total movement of 0.40.

Figure 9.18 depicts the transformation for the first five elements of the sequence. The relation between the points and the degree of coverage are the same before and after the transformation. However, since the transformation is based on the random draw of μ, the numerical values of the transformed sequence are random. The resulting sequence is called a randomized Halton sequence. It has the same properties of coverage and negative correlation over observations as the original Halton draws, since the relative placement of the elements is the same; however, it is now random.

With multiple dimensions, the sequence used for each dimension is transformed separately based on its own draw from the standard uniform density. Figure 9.19 represents a transformation of a two-dimensional sequence of length 3 defined for primes 2 and 3. The sequence for prime 3 is given by the x-axis and obtains a random draw of 0.40. The sequence for prime 2 obtains a draw of 0.35. Each point in the original two-dimensional sequence is moved to the right by 0.40 and up by 0.35, wrapping as needed. The relation between the points in each dimension is maintained, and yet the sequence is now random.

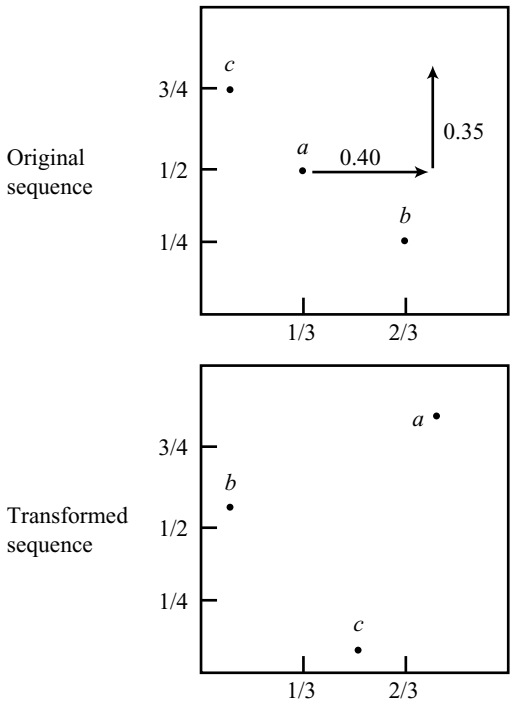

Figure 9.19. Randomization of Halton sequence in two dimensions.

9.3.5. Scrambled Halton Draws

Another issue with Halton draws arises when they are used in high dimensions. For simulation of high-dimensional integrals, Halton sequences based on large primes are necessary. For example, with 15 dimensions, the primes up to 47 are needed. However, Halton draws defined by large primes can be highly correlated with each other over large portions of the sequence. The correlation is not confined to the initial elements as described earlier, and so cannot be eliminated by discarding these elements. Two sequences defined by large and similar primes periodically become synchronized with each other and stay that way for many cycles.

Bhat (forthcoming) describes the problem and an effective solution. Figure 9.20 reproduces a graph from his paper that depicts the Halton sequence for primes 43 and 47. Clearly, these sequences are highly correlated.

This correlation can be removed while retaining the desirable coverage of Halton sequences by *scrambling* the digits of each element of the

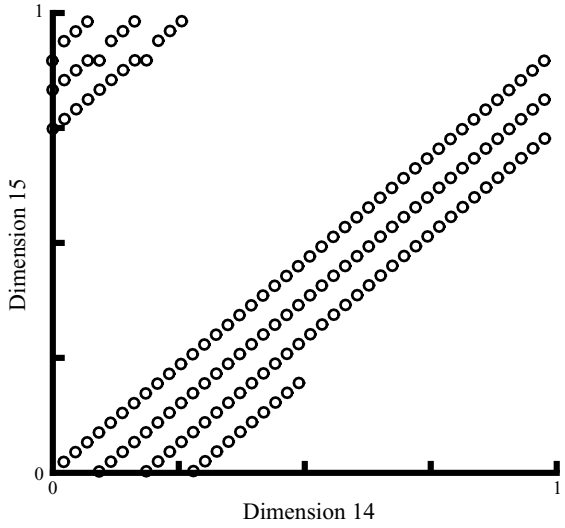

Figure 9.20. Standard Halton sequence.

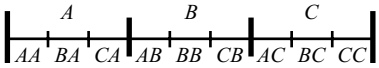

Figure 9.21. Segments for scrambling the Halton sequence.

sequences. The scrambling can be done in various ways. Braatan and Weller (1979) describe a procedure that is most easily explained through an example. Consider the Halton sequence for prime 3:

$$\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \ldots.$$

Recall that the sequence is created by dividing the unit interval into three segments, which we label A, B, and C in Figure 9.21. Each segment is divided into three subsegments, labeled AA (for subsegment A of segment A), BA (subsegment B of segment A), CA, AB, BB, CB, AC, BC, and CC. The Halton sequence is the starting point of each segment arranged alphabetically and ignoring A (i.e., ignore A, $\frac{1}{3}$ for B, $\frac{2}{3}$ for C), followed by the starting point of each subsegment arranged alphabetically and ignoring A (i.e., ignore AA, AB, and AC, $\frac{1}{9}$ for BA, $\frac{4}{9}$ for BB, $\frac{7}{9}$ for BC, $\frac{2}{9}$ for CA, $\frac{5}{9}$ for CB, and $\frac{8}{9}$ for CC.) Note that the segments and subsegments starting with A are ignored because their starting points either are 0 (for segment A) or are already included in the sequence (e.g., the starting point of subsegment AB is the same as the starting point of segment B).

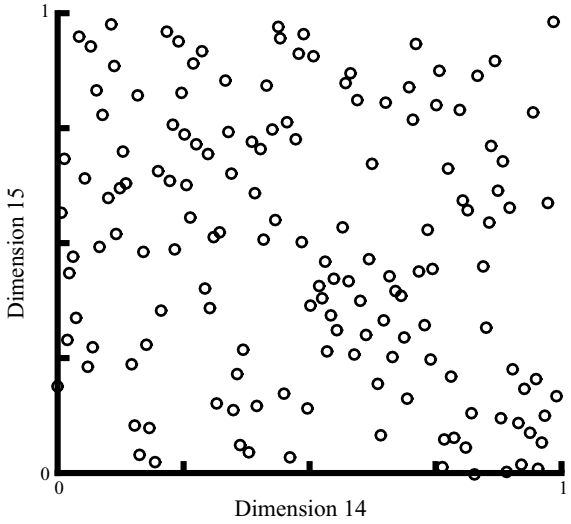

Figure 9.22. Scrambled Halton sequence.

The scrambled sequence is obtained by reversing B and C, that is, by considering C to be before B in the alphabet. The alphabetical listing is now: segments A C B, subsegments AA AC AB CA CC CB BA BC BB. The sequence is then created the same way as before but with this new alphabetical ordering: ignore A, $\frac{2}{3}$ for C, $\frac{1}{3}$ for B; ignore AA, AC, and AB, $\frac{2}{9}$ for CA, $\frac{8}{9}$ for CC, $\frac{5}{9}$ for CB, $\frac{1}{9}$ for BA, $\frac{7}{9}$ for BC, $\frac{4}{9}$ for BB. The orginal and scrambled sequences are:

Original	Scrambled
1/3	2/3
2/3	1/3
1/9	2/9
4/9	8/9
7/9	5/9
2/9	1/9
5/9	7/9
8/9	4/9

Different permutations of the letters are used for different primes. Figure 9.22, from Bhat (forthcoming), shows the scrambled sequence for primes 43 and 47. The points are not correlated as they are in the original sequence. Bhat demonstrates that scrambled sequences perform well for high-dimensional integrals in the same way that unscrambled ones do for low-dimensional integrals.

9.3.6. Other Procedures

We have described only a few of the most prominent and straightforward antithetic and quasi-random procedures. More complex procedures, with desirable theoretical properties, are described by Niederreiter (1978, 1988), Morokoff and Caflisch (1995), Joe and Sloan (1993), and Sloan and Wozniakowski (1998), to name only a few in this burgeoning area of research. As we have seen with Halton sequences, fairly simple procedures can provide large improvements over random draws. Comparisons performed by Sándor and András (2001) on probit and Sándor and Train (2002) on mixed logit indicate that the accuracy of simulation-based estimation of discrete choice models can be improved even further with the more complex procedures. It is important to remember, however, in the excitement of these methods, that accuracy can always be improved by simply using more draws. The researcher needs to decide whether learning and coding new methods of taking draws is more expedient, given her time constraints, than simply running her model with more draws.