

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

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

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, a term that 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, if interested, to supplement the more concise description that I give below.

Consider two random variables  $\varepsilon_1$  and  $\varepsilon_2$ . 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  $\varepsilon_1$  conditional on a value of  $\varepsilon_2$ , drawing  $\varepsilon_2$  conditional on this draw of  $\varepsilon_1$ , drawing a new  $\varepsilon_1$  conditional on the new value of  $\varepsilon_2$ , and so on. This process converges to draws from the joint density.

To be more precise: (1) Choose an initial value for  $\varepsilon_1$ , called  $\varepsilon_1^0$ . Any value with non-zero density can be chosen. (2) Draw a value of  $\varepsilon_2$ , called  $\varepsilon_2^0$ , from  $f(\varepsilon_2|\varepsilon_1^0)$ . (3) Draw a value of  $\varepsilon_1$ , called  $\varepsilon_1^1$ , from  $f(\varepsilon_1|\varepsilon_2^0)$ . (4) Draw  $\varepsilon_2^1$ , from  $f(\varepsilon_2|\varepsilon_1^1)$ . And so on. The values of  $\varepsilon_1^t$ , from  $f(\varepsilon_1|\varepsilon_2^{t-1})$  and the values of  $\varepsilon_2^t$ , from  $f(\varepsilon_2|\varepsilon_1^{t-1})$  constitute a sequence in  $t$ . For sufficiently high  $t$  (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:

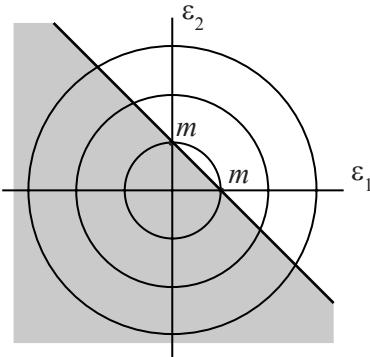


Figure 9.3: Truncated normal density.

$\varepsilon_1 + \varepsilon_2 \leq m$ . Figure 9.3 depicts the truncated density. The circles denote 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  $\varepsilon_2$ , a draw of  $\varepsilon_1$  is obtained with the procedure in section 9.2.4:  $\varepsilon_1 = \Phi^{-1}(\mu\Phi(m - \varepsilon_2))$ , where  $\mu$  is a standard uniform draw and  $\Phi(\cdot)$  is the cumulative standard normal distribution. Draws from  $\varepsilon_2$  conditional on  $\varepsilon_1$  are obtained analogously. Drawing sequentially from these conditional densities eventually provides draws from the joint truncated density.

### 9.2.9 Metropolis-Hastings algorithm

If all else fails, the Metropolis-Hastings (M-H) algorithm can be used to obtain draws from a density. Initially developed by Metropolis et al. (1953) and generalized by Hastings (1970), the M-H algorithm operates as follows. The goal is to obtain draws from  $f(\varepsilon)$ . (1) Start with a value of the vector  $\varepsilon$ , labeled  $\varepsilon^0$ . (2) Choose a “trial” value of  $\varepsilon^1$  as  $\tilde{\varepsilon}^1 = \varepsilon^0 + \eta$  where  $\eta$  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 to the density at the

original value  $\varepsilon^0$ . 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  $\varepsilon^1$ , and move to step 4. If  $f(\tilde{\varepsilon}^1) \leq f(\varepsilon^0)$ , then accept  $\tilde{\varepsilon}^1$  with probability  $\frac{f(\tilde{\varepsilon}^1)}{f(\varepsilon^0)}$ , and reject it with probability  $1 - \frac{f(\tilde{\varepsilon}^1)}{f(\varepsilon^0)}$ . To determine whether to accept or reject  $\tilde{\varepsilon}^1$  in this case, draw a standard uniform  $\mu$ . If  $\mu \leq \frac{f(\tilde{\varepsilon}^1)}{f(\varepsilon^0)}$ , then keep  $\tilde{\varepsilon}^1$ . Otherwise, reject  $\tilde{\varepsilon}^1$ . If  $\tilde{\varepsilon}^1$  is accepted, then label it  $\varepsilon^1$ . If  $\tilde{\varepsilon}^1$  is rejected, then use  $\varepsilon^0$  as  $\varepsilon^1$ . (4) Choose a trial value of  $\varepsilon^2$  as  $\tilde{\varepsilon}^2 = \varepsilon^1 + \eta$ , where  $\eta$  is a new draw from  $g(\eta)$ . (5) Apply the rule in step 3 to either accept  $\tilde{\varepsilon}^2$  as  $\varepsilon^2$  or reject  $\tilde{\varepsilon}^2$  and use  $\varepsilon^1$  as  $\varepsilon^2$ . (6) Continue this process for many iterations. The sequence  $\varepsilon^t$  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 M-H 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  $\varepsilon$  is distributed proportional to  $f^*(\varepsilon)$ . This means that the density of  $\varepsilon$  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 M-H algorithm does not utilize  $k$ . A trial value of  $\varepsilon^t$  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  $\frac{f(\tilde{\varepsilon}^t)}{f(\varepsilon^{t-1})}$ . The normalizing constant drops out of this ratio.

The M-H 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 M-H algorithm, as Gelman (1992) pointed out. The M-H 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 ex-

plore further issues about the M-H 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  $f(\cdot\cdot\cdot)$ , 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  $\varepsilon$  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 1 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  $\bar{t} = [t(\varepsilon_1) + t(\varepsilon_2)]/2$  is  $[V(t(\varepsilon_1)) + V(t(\varepsilon_2)) + 2Cov(t(\varepsilon_1), t(\varepsilon_2))]/4$ . 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  $\varepsilon^a$  and  $\varepsilon^b$ , and collectively as  $\varepsilon = \langle \varepsilon^a, \varepsilon^b \rangle'$ . A draw of  $\varepsilon$  from its density  $f(\varepsilon)$  is denoted  $\varepsilon_r = \langle \varepsilon_r^a, \varepsilon_r^b \rangle'$  for  $r = 1, \dots, R$ . Thus,  $\varepsilon_3^a$ , e.g., is the third draw of the first random term.

### 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  $\mu_1$ , and its antithetic variate is created as  $\mu_2 = 1 - \mu_1$ . The

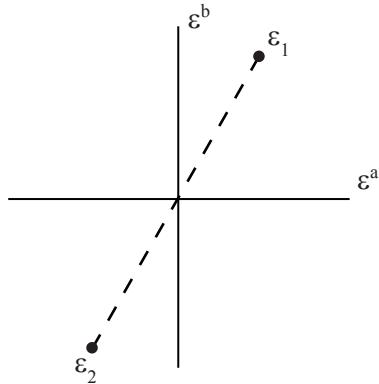


Figure 9.4: Reverse sign of both elements.

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  $\varepsilon$  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, such that the variance of their sum is zero:  $V(\varepsilon_1 + \varepsilon_2) = V(\varepsilon_1) + V(\varepsilon_2) + 2Cov(\varepsilon_1, \varepsilon_2) = 0$ . 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 non-linear 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  $\check{P} = (1/2)[P(\varepsilon_1) + P(\varepsilon_2)]$  is greater than zero. However, the variance of the simulated probabilities is less than  $(1/2)V_r(P(\varepsilon_r))$ , 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  $\varepsilon$  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

$$\begin{aligned}\varepsilon_2 &= \langle -\varepsilon_1^a, \varepsilon_1^b \rangle' \\ \varepsilon_3 &= \langle \varepsilon_1^a, -\varepsilon_1^b \rangle' \\ \varepsilon_4 &= \langle -\varepsilon_1^a, -\varepsilon_1^b \rangle'.\end{aligned}$$

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

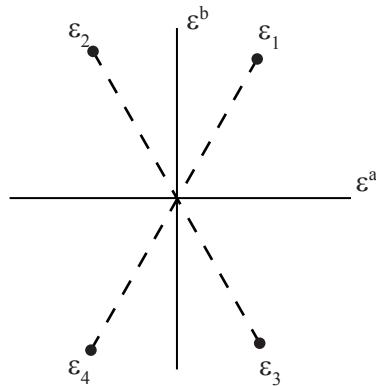


Figure 9.5: Reverse sign of each element then both.

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,  $\varepsilon_1$  and  $\varepsilon_2$  are fairly close together, as are  $\varepsilon_3$  and  $\varepsilon_4$ . This placement leaves large uncovered areas between  $\varepsilon_1$  and  $\varepsilon_3$  and between  $\varepsilon_2$  and  $\varepsilon_4$ . Orthogonal draws with even placement can be obtained by switching element  $\varepsilon_1^a$  with  $\varepsilon_1^b$  while also reversing the signs. The antithetic draws are

$$\begin{aligned}\varepsilon_2 &= \langle -\varepsilon_1^b, \varepsilon_1^a \rangle' \\ \varepsilon_3 &= \langle \varepsilon_1^b, -\varepsilon_1^a \rangle' \\ \varepsilon_4 &= \langle -\varepsilon_1^a, -\varepsilon_1^b \rangle',\end{aligned}$$

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

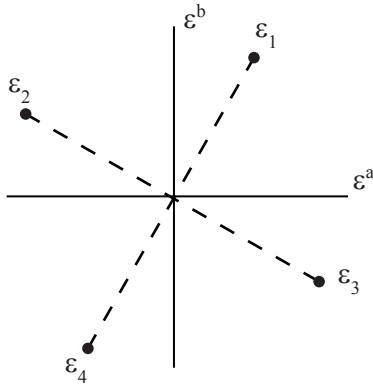


Figure 9.6: Switch positions and reverse signs.

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 could 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 .25 (by drawing from a standard uniform and dividing the result by 4.) Label the draw  $\varepsilon_1$ . Three other draws are created as

$$\begin{aligned}\varepsilon_2 &= .25 + \varepsilon_1 \\ \varepsilon_3 &= .50 + \varepsilon_1 \\ \varepsilon_4 &= .75 + \varepsilon_1\end{aligned}$$

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

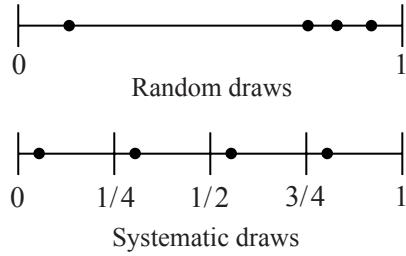


Figure 9.7: Draws from standard uniform.

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 above.

Systematic sampling can be performed in multiple dimensions. Consider a two-dimensional uniform on the unit square. A net is created

by dividing each dimension into segments. As shown in Figure 9.8, when each dimension is divided into four segments, the unit square is partitioned into 16 areas. A draw between 0 and .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  $\varepsilon_4 = \langle (0.75 + \varepsilon_1^a), (0 + \varepsilon_1^b) \rangle'$ .

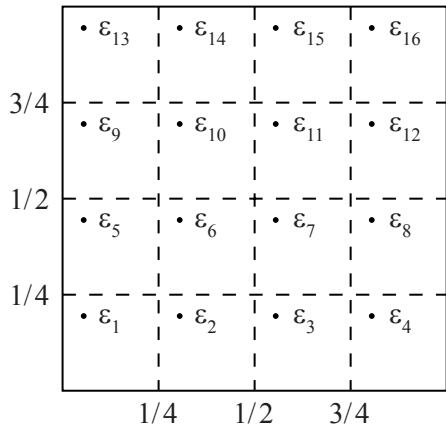


Figure 9.8: Systematic draws in two dimensions.

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:  $\Phi^{-1}(0.25) = -.67$ ,  $\Phi^{-1}(0.5) = 0$ , and  $\Phi^{-1}(0.75) = .67$ . 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  $\mu_1$ . 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 multi-dimensional random terms are obtained similarly,

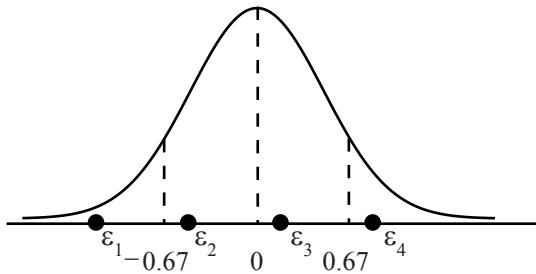


Figure 9.9: Systematic draws for univariate normal.

provided that the elements are independent. For example, if  $\varepsilon$  consists of two elements each of which is standard normal, then draws analogous to those in Figure 9.8 are obtained as follows: Draw  $\mu_1^a$  and  $\mu_1^b$  from a uniform between 0 and 0.25. Calculate  $\varepsilon_1$  as  $(\Phi^{-1}(\mu_1^a), \Phi^{-1}(\mu_1^b))'$ . Calculate the other 15 points as  $\varepsilon_r$  as  $(\Phi^{-1}(x_r + \mu_1^a), \Phi^{-1}(y_r + \mu_1^b))'$ , where  $(x_r, y_r)'$  is the origin of area  $r$  in the unit square.

The requirement that the elements of  $\varepsilon$  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.

Obviously, numerous sets of systematically 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 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 8 segments in each dimension such that each random draw in the 0 to 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 uni-dimensional normal, 4 draws that are symmetric around zero are obtained as follows. Draw a uniform between 0 and 0.25, labeled  $\mu_1$ . 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  $\varepsilon_4 = -\varepsilon_1$ . Fig-

ure 9.10 illustrates the draws using the same  $\mu_1$  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.

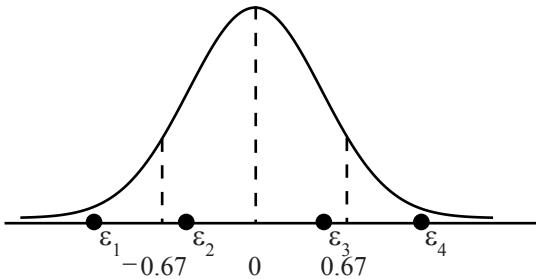


Figure 9.10: Symmetric systematic draws.

### 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 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 9 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

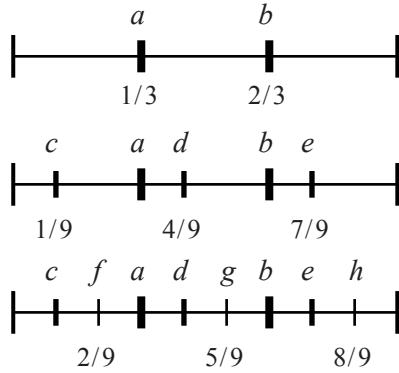


Figure 9.11: Halton sequence for prime 3.

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 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 sequence 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:

$$\begin{array}{ll}
 0 & 0 \\
 1/3 & 1/3 \\
 2/3 & 1/3 \\
 0 + 1/9 & 1/9 \\
 1/3 + 1/9 & = 4/9 \\
 2/3 + 1/9 & 7/9 \\
 0 + 2/9 & 2/9 \\
 1/3 + 2/9 & 5/9 \\
 2/3 + 2/9 & 7/9
 \end{array}$$

The new sequence consists of nine elements.

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

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

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 3 num-

bers:

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	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  $\{\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \dots\}$ . In general, the sequence for prime  $k$  is created iteratively, with the sequence at iteration  $t + 1$  being  $s_{t+1} = \{s_t, s_t + 1/k^t, s_t + 2/k^t, \dots, s_t + (k - 1)/k^t\}$ . 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 evermore 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

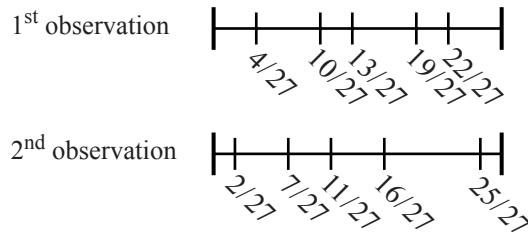


Figure 9.12: Halton draws for two observations.

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 filled in by its midpoint,  $\frac{16}{27}$ , for the second observation. And so on. The pattern by which Halton sequences are created make it 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 such 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:

$$\begin{aligned}\Phi^{-1}(1/3) &= -.43 \\ \Phi^{-1}(2/3) &= .43 \\ \Phi^{-1}(1/9) &= -1.2 \\ \Phi^{-1}(4/9) &= -.14 \\ \Phi^{-1}(7/9) &= .76 \\ &\dots\end{aligned}$$

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  $-.43$  and  $.43$ , and then dividing each segment into three subsegments of equal mass, and so on.

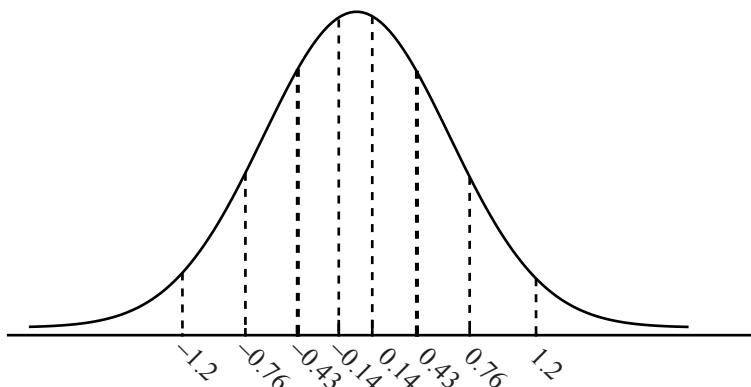


Figure 9.13: Halton draws for a standard normal.

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

$$\varepsilon_1 = \langle \frac{1}{2}, \frac{1}{3} \rangle$$

$$\varepsilon_2 = \langle \frac{1}{4}, \frac{2}{3} \rangle$$

$$\varepsilon_3 = \langle \frac{3}{4}, \frac{1}{9} \rangle$$

$$\varepsilon_4 = \langle \frac{1}{8}, \frac{4}{9} \rangle$$

$$\varepsilon_5 = \langle \frac{5}{8}, \frac{7}{9} \rangle$$

$$\varepsilon_6 = \langle \frac{3}{8}, \frac{2}{9} \rangle$$

...

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

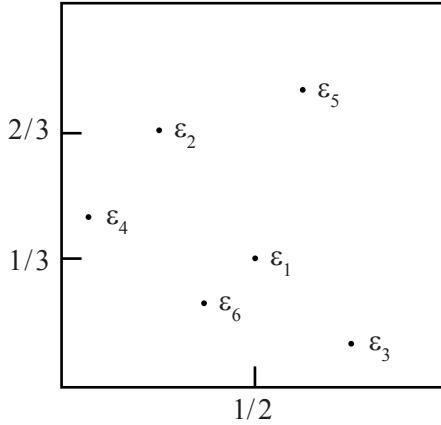


Figure 9.14: Halton sequence in two dimensions for primes 2 and 3.

dimensional independent standard normal, the inverse cumulative nor-

mal is taken of each element of these pairs. The draws are

$$\begin{aligned}\varepsilon_1 &= \langle 0, -.43 \rangle \\ \varepsilon_2 &= \langle -.67, .43 \rangle \\ \varepsilon_3 &= \langle .67, -1.2 \rangle \\ \varepsilon_4 &= \langle -1.15, -.14 \rangle \\ \varepsilon_5 &= \langle .32, .76 \rangle \\ \varepsilon_6 &= \langle -.32, -.76 \rangle \\ &\dots\end{aligned}$$

which are shown in Figure 9.15.

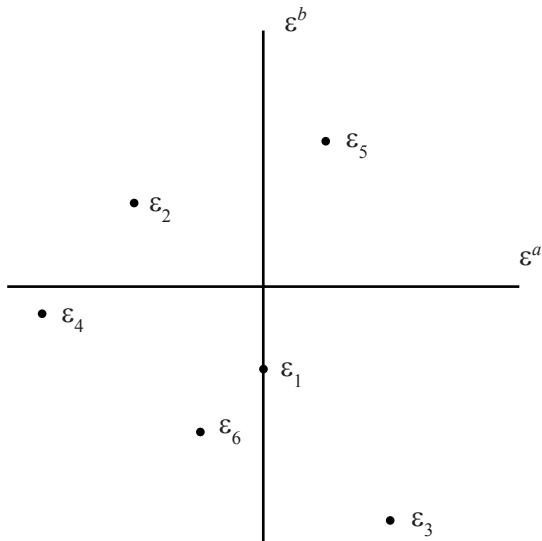


Figure 9.15: Halton sequence for 2-dimensional standard normal.

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  $\{\frac{1}{11}, \frac{2}{11}, \frac{3}{11}, \frac{4}{11}, \frac{5}{11}, \frac{6}{11} \dots\}$ . These first elements fall in 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 correlated portion. 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.

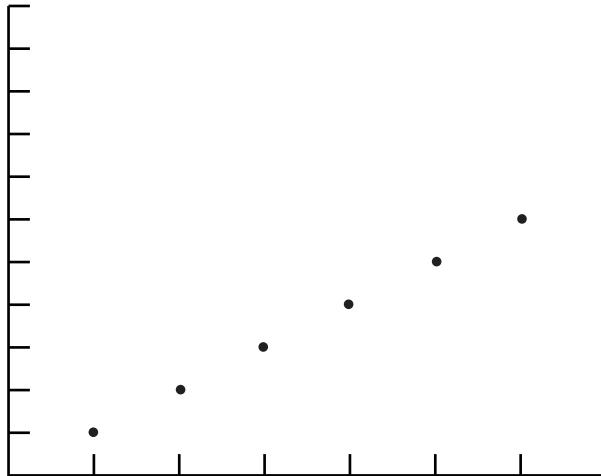


Figure 9.16: First six elements of Halton sequence for primes 7 and 11.

The potential for correlation is the reason that prime numbers are used to create the Halton sequences instead of non-primes. If a non-prime 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, 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 household's 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/she 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 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 that 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, such that the dimension for which each prime was used changed in the five runs. The average and standard deviation of the five sets of estimated 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.

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

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

8 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 indicates 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 in lieu 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 helpful warning. The model was re-estimated with 125 Halton draws instead of 100. It was estimated five times under each of the five orderings of the prime numbers as described above. 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

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

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 concept 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 the characteristics of truly random draws, but the fact that they are systematic is why they are called “pseudo-random draws.” In this regard, therefore, Halton draws can be seen as a systematic way of approximating integration that is more

precise than using pseudo-random 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 pseudo-random 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 as  $\mu$ .
2. Add  $\mu$  to each element of the Halton sequence. If the resulting element exceeds one, 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  $\text{mod}$  takes the fractional part of the term in parentheses.

The transformation is depicted in Figure 9.17. Suppose the draw of  $\mu$  from the uniform density is 0.40. The number .33 is the first element of the Halton sequence for prime 3. This element is transformed, as shown in the top panel, to  $.33 + .40 = .73$ , which is just a .40 move up the line. The number .67 is the second element of the sequence. It is transformed by adding .4 and then, since the result exceeds 1, by subtracting 1 to get 0.07.  $(.67 + .40 - 1 = .07)$ . As shown in the bottom panel, this transformation is visualized as moving the original point up by a distance of .40, but wrapping around when the end of the unit interval is reached. The point moves up .33 to where the line ends, and then wraps to the start of the line and continues to move up another .07, for a total movement of .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  $\mu$ , 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 are the same; however, it is now random.

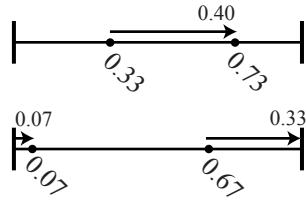
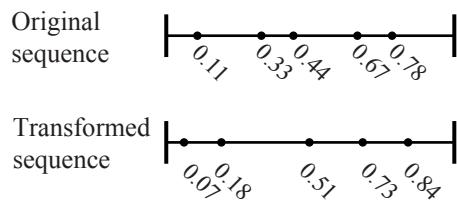
Figure 9.17: Random transformation of Halton draws with  $\mu = 0.40$ .

Figure 9.18: Randomization of Halton sequence in one dimension.

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 2-dimensional sequence of length 3 defined in primes 2 and 3. The sequence in 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 2-dimensional sequence is moved to the left by .40 and up by .35, wrapping as needed. The relation between the points in each dimension is maintained, and yet the sequence is now random.

### 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 above, and so cannot be eliminated by