Stat 362 Supplementary notes

Art B. Owen

November 2002

Abstract

These are supplementary notes for some of the Stat 362 topics that are not adequately covered in either of the course text books. They're intended to give some explanations to help students understand their course notes. They're not in polished textbook form, they're just here to fill some gaps.

Some parts are incomplete:

1. There are citations but no bibliography listing the cited references.
2. A citation like Patterson (195x) refers to a reference that may be filled in later, if these notes evolve further.
3. Similarly a reference like Figure XXX usually refers to a figure that was sketched on the blackboard in class, or handed out.
4. You don't have to do the exercises, unless they also turn up on one of the problem sets.
5. Chapter xxx might refer to something that belongs in these notes some day. It might appear later, or it might have been covered in class.
6. Some Propositions may turn into Theorems or Lemmas or other such items.
7. Notation isn't perfectly consistent. Later I'll decide on rules for when to use X instead of x and so forth.

   Similarly, notation for the variance of estimator A for a sampling method B can be written $\mathrm{Var}(\hat{I}_A)$ or $\mathrm{Var}_B(\hat{I})$ or $\mathrm{Var}_B(\hat{I}_A)$. Eventually one form will prove most useful.

   Much of the literature works with observations indexed from 0 to $n-1$, the rest use 1 through n. Almost nobody indexes the variables from 0 to $d-1$ though.
8. In the literature the dimension is either d or s. I like to use dimension d for the problem, and dimension s for a method of solution. Later we'll look at tricks for bridging the gap when $s < d$.

Contents

1 Latin hypercube sampling and more **2**
- 1.1 Introduction to LHS . . . 2
- 1.2 Effectiveness of LHS . . . 4
- 1.3 Anova of $[0,1]^d$ . . . 5
  - 1.3.1 Anova: more notes . . . 9
- 1.4 Randomized orthogonal arrays . . . 9

2 Quasi-Monte Carlo **11**
- 2.1 Discrepancy . . . 11
- 2.2 Dimension $d = 1$ . . . 14
- 2.3 Halton and Hammersley sequences . . . 15
- 2.4 (t, m, s)-nets and (t, s)-sequences . . . 18

Chapter 1

Latin hypercube sampling and more

1.1 Introduction to LHS

Latin hypercube sampling is a form of simultaneous stratification on all d variables of the unit cube $[0,1]^d$. There are two versions. In the centered version (called lattice sampling by Patterson (195x))

$$X_i^j = \frac{\pi_j(i) - 1/2}{n}, \quad i = 1, \ldots, n,\ j = 1, \ldots, d \tag{1.1}$$

where π_j are independent uniform random permutations of 1 through n. In the unbiased version, due to Beckman, Conover and McKay (197x),

$$X_i^j = \frac{\pi_j(i) - U_i^j}{n}, \quad i = 1, \ldots, n,\ j = 1, \ldots, d \tag{1.2}$$

where U_i^j are independent $U(0,1)$ random variables, independent of the permutations π_j.

Figure xxx shows a Latin hypercube sample with $n = 25$ and $d = 2$, and Figure xxx shows a centered version. Each of 25 rows has exactly one of the 25 points as does each of 25 columns in each figure. Thus there are 50 strata of area $1/25$ getting 1 of the 25 points. In d dimensions there are nd strata of volume $1/n$ each getting a proportional allocation of 1 of the n points.

The name "Latin hypercube" stems from a relationship between Latin hypercubes and Latin squares. A Latin square is an n by n array of cells. Each cell has one of n different symbols (usually letters) written in it. Every row of the array has each of the n symbols exactly once. So does every column. Suppose the symbols are A, B, and so forth. Then the cells occupied by the letter A constitute a Latin hypercube sample of n points in the $d = 2$ dimensional coordinates of the array. There is a rich combinatorial theory underlying the

construction of Latin squares. That theory does not play a role in the construction or analysis of Latin hypercubes, but it does become relevant in the study of randomize orthogonal arrays.

A Latin hypercube sample tends to be more uniformly distributed through the unit cube than an IID sample. A histogram of X_1^j through X_n^j with n equal width cells would be perfectly flat for each $j = 1, \dots, d$ while the corresponding histograms for IID samples would typically be uneven. Perhaps the worst case Latin hypercube sample has all n points arranged on the diagonal in the d dimensional cube. The probability of such a sample is $(n!)^{1-d}$. Of course a worst case sample for IID sampling could have all of the points crowded in a small box, instead of along a line.

Exercise Explain why the diagonal probability is not $(n!)^{-d}$.

Exercise Let $B \subset [0,1]^d$ be a box of volume V. Suppose that B has probability $(n!)^{-d}$ of containing all n points, when the points are IID. What is V?

Harder exercise Consider the smallest box that contains all n of the IID points. What is the probability that this box has volume $(n!)^{1-d}$ or less?

For either form of LHS, the estimate of $I = \int f(x)dx$ is still

$$\hat{I}_{LHS} = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

In the randomly centered version of a Latin hypercube sample each $X_i \sim U(0,1)^d$, generating unbiased sample means, as shown below.

Proposition 1.1 *For the randomly centered version of Latin hypercube sampling* (1.2) *each* $X_i \sim U(0,1)^d$.

Proof Sketch of proof to come:

1. each of n cells $[(k-1)/n, k/n)$ for $k = 1, \dots, n$ has equal probability $1/n$ of containing X_i^1 for each i. (Cell choice driven by $\pi_1(i)$.)
2. placement within each cell is uniform, based on U_i^1.
3. so $X_i^1 \sim U(0,1)$
4. same for $j = 2, \dots, d$.
5. variables used to generate $X_1^j, \dots, X_n^j$ indep of those for $X_1^k, \dots, X_n^k$ for $j \neq k$.

□

Proposition 1.2 *Centered Latin hypercube sampling is unbiased.*

Proof

$$E(\hat{I}_{LHS}) = E\biggl(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\biggr) = \frac{1}{n}\sum_{i=1}^{n} E\Bigl(f(X_i)\Bigr)$$

and $E(f(X_i)) = \int f(x)dx$ by Proposition 1.1. $\square$

For the centered version of LHS the sample mean has a bias. There each X_i^j is uniformly distributed on values $(k-1/2)/n$ for $k = 1, \ldots, n$. It follows that X_i is uniformly distributed over a grid of n^d points and so

$$E(\hat{I}_{CLHS}) = \frac{1}{n^d}\sum_{i_1=1}^{n}\cdots\sum_{i_d=1}^{n} f\left(\left(\frac{i_1-1/2}{n},\ldots,\frac{i_d-1/2}{n}\right)\right). \tag{1.3}$$

The right side of (1.3) is the product of d midpoint rules of n points. It thus can be expected to have the accuracy of a one dimensional midpoint rule on n points. If f is smooth, then such a midpoint rule has error $O(n^{-2})$. This is a small error compared to the Monte Carlo error rate, which we shall see is still $O_p(n^{-1/2})$ for both kinds of LHS.

1.2 Effectiveness of LHS

Stratification with proportional allocation never increases variance compared to IID sampling, and can reduce it. Therefore it is natural to expect that d separate kinds of proportional stratification applied simultaneously as in LHS should might reduce variance too. Stratification works best when f is nearly constant within strata and varies significantly between strata. For multiple stratification we should expect a good result if f is nearly constant within strata defined by any of the stratifications we're using. More generally, if f is nearly a sum of functions each of which is nearly constant within one of the stratifications in use, we should expect a good result.

LHS works well for functions that are additive or nearly additive in the d input variables. An additive function is a sum of d functions each of which depends on just one input variable. Later, in Chapter xxx, we'll construct the closest additive function to f. For now suppose that

$$f(X) = \sum_{j=1}^{d} g_j(X^j) + h(X)$$

where either $h(x) = 0$ for all x or $\int h(x)^2dx$ is small compared to σ^2. Then

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n} g_j(X_i^j) + \frac{1}{n}\sum_{i=1}^{n} h(X_i). \tag{1.4}$$

In the right side of (1.4) the averages of $g_j(X_i^j)$ are one dimensional integral estimates corresponding to either a midpoint rule (centered LHS) or a stratified sample (unbiased LHS). The former can have errors as small as $O(n^{-2})$ and the latter as small as $O_p(n^{-3/2})$ for general but smooth g_j. The average $h(X_i)$ in (1.4) is a standard Monte Carlo average, so we can expect an error of size $O_p(n^{-1/2})$, but with a small variance when $h(x)$ is small.

Let f_{add} be the additive function closest to f, that is $\int(f(x)-g(x))^2 \geq \int(f(x)-f_{\mathrm{add}}(x))^2dx$ if g is additive. Then the following facts are known about LHS

1. If $\int f(x)^2dx < \infty$ $\mathrm{Var}(\hat{I}_{\mathrm{lhs}}) = \frac{1}{n}\int f_{\mathrm{res}}(x)^2dx + o(n^{-1})$ (Michael Stein (1987) Technometrics)
2. $\mathcal{L}(\hat{I}_{\mathrm{lhs}} - I)$ under LHS is very close to $\mathcal{L}((1/n)\sum_{i=1}^n f_{\mathrm{res}}(X_i))$ under IID sampling. (Owen 1992 JRSS-B, Loh 199x Annals of Stat) In particular $n^{1/2}(\hat{I}_{\mathrm{lhs}} - I) \to N(0, \int f_{\mathrm{res}}(x)^2)dx$.
3. If $\int f(x)^2dx < \infty$ then $\mathrm{Var}_{\mathrm{lhs}}(\hat{I}) \leq [n/(n-1)]\mathrm{Var}_{\mathrm{iid}}(\hat{I})$

$\mathcal{L}$ means the "law" or distribution of

1.3 Anova of $[0,1]^d$

In LHS it was beneficial to decompose the integrand f into a sum of functions, d of which depended on just one input variable.

More generally it is useful to decompose an integrand into a sum of 2^d functions, one for each subset of $\{1,2,\ldots,d\}$. The function for subset u depends on X only through X^j for $j \in u$.

For example, consider the function $f(x) = x^1x^2x^3$ on $[0,1]^3$. It can be written

$$\begin{aligned} f(x) =& \frac{1}{8} + \frac{1}{4}\left(x^1 - \frac{1}{2}\right) + \frac{1}{4}\left(x^2 - \frac{1}{2}\right) + \frac{1}{4}\left(x^3 - \frac{1}{2}\right) \\ &+ \frac{1}{2}\left(x^1 - \frac{1}{2}\right)\left(x^2 - \frac{1}{2}\right) + \frac{1}{2}\left(x^1 - \frac{1}{2}\right)\left(x^3 - \frac{1}{2}\right) + \frac{1}{2}\left(x^2 - \frac{1}{2}\right)\left(x^3 - \frac{1}{2}\right) \\ &+ \left(x^1 - \frac{1}{2}\right)\left(x^2 - \frac{1}{2}\right)\left(x^3 - \frac{1}{2}\right). \end{aligned} \tag{1.5}$$

There are 8 functions in (1.5), one for each subset of $\{1,2,3\}$. Each of these functions integrates to zero over the range of any of the variables in it.

The ANOVA decomposition is a representation of the form

$$f(X) = \sum_u f_u(X) \tag{1.6}$$

where the sum is over subsets $u \subseteq \{1,2,\ldots,d\}$ and $f_u(X)$ is a function on $[0,1]^d$ that depends on X only through X^j where $j \in u$. In particular, for the emptyset, $f_\emptyset(X)$ is a constant. For a given function f, more than one representation of the form (1.6) is possible. We describe below how to specify the unique ANOVA decomposition of f.

Operationally the summation in (1.6) may be carried out in order of increasing subset size

$$f_\emptyset(X) + \sum_{r=1}^{d} \sum_{j_1<\cdots<j_r} f_{\{j_1,j_2,\ldots,j_r\}}(X) \tag{1.7}$$

where each $j_k \in \{1, 2, \ldots, d\}$ or it might be carried out in "odometer order" as

$$\sum_{j_1 \in \{0,1\}} \cdots \sum_{j_d \in \{0,1\}} f_{\{r|j_r=1\}}(X). \tag{1.8}$$

Most of the time we hide the details of how subsets are summed over, and use the short form (1.6).

To understand the non-uniqueness in (1.6) consider three sets $u, v, w \subseteq \{1, \ldots, d\}$ with $w \subset u$ and $w \subset v$, and $u \neq v$. Then if we replace f_u by $f_u + f_w$ and f_v by $f_v - f_w$ (1.6) still holds.

When working with subsets, we employ the following notation: $|u|$ is the cardinality of u, $-u$ is the set theoretic complement $\{1, 2, \ldots, d\} - u$, X^u is the vector of $|u|$ components made up of X^j for $j \in u$, and dX^u describes integration with respect to all the variables X^j for $j \in u$. Integrating a function $f(X)$ with respect to dX^u leaves a function of X that depends on X only through X^{-u}. (For instance, if $g(z, y) = \cos(z + y)$, then $\int_0^1 g(z, y)dy = \sin(z + 1) - \sin(z)$ depending on z, but not on y which was "integrated out".) The set $[0, 1]^u$ is a copy of $[0, 1]^{|u|}$ to which X^u belongs.

To select a unique representation of the form (1.6), it seems reasonable to attribute as much structure as possible to the smallest subsets of variables. The ANOVA components are defined recursively through

$$f_u(X) = \int_{[0,1]^{-u}} \Big(f(X) - \sum_{v \subset u, v \neq u} f_v(X) \Big) dX^{-u}. \tag{1.9}$$

That is we first subtract from f anything that can be attributed to X^v for a proper subset v of u, then to get the effect of X^u, we average this residual over values of X^j for $j \notin u$. For $u = \emptyset$, applying natural conventions to (1.9) leads to $f_\emptyset(X) = \int (f(X) - 0)dX = I$. Then

$$f_{\{j\}}(X) = \int \Big(f(X) - I \Big) \prod_{k \neq j} dX^k, \tag{1.10}$$

is the "main effect" for variable j.

Because f_v only depends on X^j for $j \in v \subset u$ it follows that f_v is unaffected by the value of X^{-u}, and so we may also write

$$f_u(X) = \int_{[0,1]^{-u}} f(X)X^{-u} - \sum_{v \subset u, v \neq u} f_v(X). \tag{1.11}$$

Equation (1.11) is simpler to work with in analytical expressions. There may be a numerical disadvantage to (1.11) relative to (1.9) if the former requires the subtraction of two large numbers of nearly equal size.

Lemma 1.1 *Suppose that* $\int f(x)^2 dx < \infty$ *and that* $j \in u \subset \{1, \ldots, d\}$. *Then* $\int_0^1 f_u(x)dx^j = 0$.

Proof The proof is by induction on $|u|$. For $|u| = 1$, let $u = \{j\}$ and then by (1.10)

$$\begin{aligned}\int_0^1 f_{\{j\}}(X)dX^j &= \int_0^1 \int_{[0,1]^{-\{j\}}} (f(X) - I) \prod_{k\neq j} dX^k dX^j \\ &= \int_{[0,1]^d} (f(X) - I)dX \\ &= 0.\end{aligned}$$

Now suppose that $\int_0^1 f_v(X)dX^j = 0$ for $j \in v$ whenever $1 \leq |v| \leq r < d$ and suppose that $|u| = r + 1$ and that $j \in u$, and let $-u + j$ be a shorthand for the union of j and $-u$. To complete the induction,

$$\begin{aligned}\int_0^1 f_u(X)dX^j &= \int_{[0,1]^{-u+j}} \Big(f(X) - \sum_{v\subset u, v\neq u} f_v(X)\Big)dX^{-u+j} \\ &= \int_{[0,1]^{-u+j}} \Big(f(X) - \sum_{v\subset u, v\neq u, j\neq v} f_v(X)\Big)dX^{-u+j} \\ &= \int_{[0,1]^{-u+j}} \Big(f(X) - \sum_{v\subseteq u-\{j\}} f_v(X)\Big)dX^{-u+j} \\ &= \int_{[0,1]^{-u+j}} \Big(f(X) - \sum_{v\subset u-\{j\} v\neq u-\{j\}} f_v(X)\Big)dX^{-u+j} + f_{u-\{j\}}(X) \\ &= f_{u-\{j\}}(X) - f_{u-\{j\}}(X) \\ &= 0. \ \square\end{aligned}$$

The ANOVA splits the square integrable functions on $[0, 1]^d$ into 2^d mutually orthogonal sets.

Proposition 1.3 *Suppose that* $\int f^2 dx < \infty$ *and* $\int g^2 dx < \infty$. *Let* $u, v \subseteq \{1, \ldots, d\}$ *and suppose that* $u \neq v$. *Then* $\int f_u(x)g_v(x)dx = 0$.

Proof Since $u \neq v$, there either exists $j \in u$ with $j \notin v$, or $j \in v$ with $j \notin u$. Without loss of generality suppose that $j \in u$ and $j \notin v$. Then

$$\begin{aligned}\int f_u(x)g_v(x)dx &= \int_{[0,1]^{-\{j\}}} \int_0^1 f_u(x)g_v(x)dx^j dx^{-\{j\}} \\ &= \int_{[0,1]^{-\{j\}}} \int_0^1 f_u(x)dx^j g_v(x)dx^{-\{j\}} \\ &= 0. \ \square\end{aligned}$$

Corollary 1.1 *Suppose that* $\int f^2 dx$ *and that* $u \neq v$ *are subsets of* $\{1, \ldots, d\}$. *Then* $\int f_u(x)f_v(x)dx = 0$.

Proof Take $f = g$ in Proposition 1.3. □

The value of the ANOVA decomposition is that it allows us to decompose the variance of f into parts that we can attribute to various subsets of input variables. Let

$$\sigma_u^2 = \begin{cases} \int f_u(x)^2 dx, & |u| > 0 \\ 0 & |u| = 0. \end{cases}$$

Proposition 1.4 *Suppose that* $\int f(x)^2 dx < \infty$. *and that* $\sigma^2 = \int (f(x) - I)^2 dx$. *Then*

$$\sigma^2 = \sum_{|u|>0} \sigma_u^2.$$

Proof

$$\begin{aligned} \int (f(x) - I)^2 dx &= \int \Big[\sum_{|u|>0} f_u(x) \Big]^2 dx \\ &= \int \Big[\sum_{|u|>0} f_u(x) \Big] \Big[\sum_{|v|>0} f_v(x) \Big] dx \\ &= \int \Big[\sum_{|u|>0} f_u(x)^2 \Big] dx \\ &= \sum_{|u|>0} \sigma_u^2. \ \square \end{aligned}$$

Now we're ready to exhibit the functions f_{add} and f_{res} used to describe LHS. $f_{\text{add}}(X) = \sum_{|u|\le 1} f_u(X)$ and $f_{\text{res}}(X) = \sum_{|u|>1} f_u(X)$. To a reasonable approximation $\text{Var}(\hat{I}_{\text{lhs}}) \doteq (1/n) \sum_{|u|>1} \sigma_u^2$.

Exercise: Let $f_{\text{add}}(x) = f_\emptyset(X) + \sum_{j=1}^d f_{\{j\}}(X) = \sum_{|u|\le 1} f_u(X)$, and let $g(x)$ be any other additive function of X. Then

$$\int (f(x) - f_{\text{add}}(x))^2 dx \le \int (f(x) - g(x))^2 dx.$$

Exercise: Let $f_{\text{two}}(x) = \sum_{|u|\le 2} f_u(X)$, and let $g(x)$ be any other function of X expressible as a sum of functions that depend on only 0, 1, or 2 of the components of X. Then

$$\int (f(x) - f_{\text{two}}(x))^2 dx \le \int (f(x) - g(x))^2 dx.$$

Exercise: Let u be a subset of $\{1, \dots, d\}$, let $f_{[u]}(x) = \sum_{v \subseteq u} f_v(X)$ let $g(x)$ be any other function of X that depends on X only through X^u. Then

$$\int (f(x) - f_{[u]})^2 dx \le \int (f(x) - g(x))^2 dx.$$

1.3.1 Anova: more notes

1. The anova is only given here for square integrable functions, though some of the definitions don't seem to need square integrability. The most useful properties relate to orthogonality, so restricting to square integrable functions is not a great loss.

2. The anova extends to product domains, not just $[0,1]^d$. It originated in agricultural field trials, on products of discrete domains (one of a fertilizers and one of b grains ploughed to one of c depths etc.) Take any d sets $\mathcal{X}_j$ for $j = 1, \dots, d$ with distributions p_j for $X^j \in \mathcal{X}_j$. Then an anova can be constructed for square integrable functions on $\mathcal{X} = \prod_{j=1}^d \mathcal{X}_j$ relative to the distribution $p(X) = \prod_{j=1}^d p_j(X^j)$ of $X \in \mathcal{X}$. The anova also extends to $d = \infty$.

3. The anova of $[0,1]^d$ has often been reinvented. Hoeffding seems to be the first. Sobol also invented it. Efron and Stein made good use of it studying the jackknife. There are some recent reproducing Hilbert space versions of it from Hickernell and from Wahba. Takemura (1983) gives a history of the anova decomposition.

1.4 Randomized orthogonal arrays

Under IID sampling $\mathrm{Var}(\hat{I}) = (1/n)\sum_{|u|>0}\sigma_u^2$. Latin hypercube sampling takes out the additive part leaving $\mathrm{Var}(\hat{I}) = (1/n)\sum_{|u|>1}\sigma_u^2$. LHS achieve this by stratifying each input variable individually. The goal of randomized orthogonal array sampling is to reduce the variance to $(1/n)\sum_{|u|>2}\sigma_u^2$ by stratifying every pair of input variables. More generally it may aim to reduce the variance to $(1/n)\sum_{|u|>t}\sigma_u^2$ by stratifying every t-tuple of input variables for $2 \le t < d$. The case $t = d$ was covered in Chapter XXX on stratified sampling where we gave Haber's result that the variance is $O(n^{-1-2/d})$.

Definition 1.1 *Let A be an n by s matrix with elements $A_i^j \in \{0, 1, \dots, b-1\}$. A is an orthogonal array of strength $t \le s$, denoted $OA(n, s, b, t)$, if in each $n \times t$ submatrix of A, all b^t possible rows appear the same number λ of times.*

The number λ is called the index of the array. Clearly $\lambda = b^t/n$. For many important orthogonal array constructions $\lambda = 1$. Graphically suppose that A is a data matrix with one row per case and one column per observation. Then any scatter plot of t variables is a regular b^t grid (with λ overstrikes at each point).

If b is a prime number, then the Bose construction is available for $\mathrm{OA}(b^2, b+1, b, 2)$. This construction can handle up to $b+1$ variables of which any pair are sampled on a $b \times b$ grid. For $i = 1, \dots, n$, let $A_i^1 = \lfloor (i-1)/b \rfloor$ and $A_i^2 = (i-1) - bA_i^1$. As i goes from 1 to n, (A_i^1, A_i^2) goes like a 2 digit base b odometer. Now for $j = 3, \dots, b+1$, let $A_i^j = A_i^1 + (j-2) \times A_i^2$ using arithmetic modulo b.

The Bose construction is also available for $b = p^r$ where p is a prime number and $r \geq 1$ is an integer power, using the Galois field with p^r elements, denoted $\mathrm{GF}(p^r)$. See Chapter xxx. Write the elements of $\mathrm{GF}(p^r)$ in any order as $\phi(0), \ldots, \phi(b^r - 1)$. Now for $j = 3, \ldots, b+1$, let $A_i^j = \phi^{-1}(\phi(A_i^1) + \phi(j-2) \times \phi(A_i^2))$, doing the arithmetic in $\mathrm{GF}(b^r)$. Using integers modulo p^r will give incorrect results (when $r > 1$).

No understanding of GF is required or expected in Stat 362.

The points of the orthogonal array can be embedded into $[0,1]^d$ to give an integration rule. A moment's thought reveals that a naive embedding such as $X_i^j = (A_i^j + 1/2)/b$ will not work well. The points are too structured. For prime b we find that (A_i^1, A_i^2, A_i^3), $i = 1, \ldots, n$ lie on two planes. A random scrambling of the points can break up this pattern while preserving the uniformity of t dimensional coordinate projections of the points.

The centered version of a randomized orthogonal array has

$$X_i^j = \frac{\pi_j(A_i^j) + 1/2}{b}, \quad i = 1, \ldots, n,\ j = 1, \ldots, d \tag{1.12}$$

where π_j are independent uniform random permutations of 0 through $b-1$. There is an unbiased version

$$X_i^j = \frac{\pi_j(A_i^j) + U_i^j}{b}, \quad i = 1, \ldots, n,\ j = 1, \ldots, d \tag{1.13}$$

where U_i^j are independent $U(0,1)$ random variables.

Latin hypercube sampling corresponds to strength $t = 1$ apart from the minor change to permutations of 1 to n instead of 0 to $n-1$.

Orthogonal array sampling reduces the variance to $(1/n)\sum_{|u|>t} \sigma_u^2 + o(1/n)$. This is even so for $t = d$, which corresponds to stratification and a variance that is $O(n^{-1-2/d})$. In practice, smaller values of t are more useful. They require smaller n given d, or allow larger d given a bound on n.

Chapter 2

Quasi-Monte Carlo

In Monte Carlo sampling the points X_i tend to form clumps in some parts of $[0,1]^d$ and leave voids in others. The idea in quasi-Monte Carlo (QMC) is to find points that, to the extent possible, are spread out uniformly through $[0,1]^d$ with minimal clumps and voids. Though in most QMC schemes these points are deterministic, the goal in QMC is similar to that in stratified Monte Carlo. Because stratification schemes tend to reduce error, we anticipate QMC to reduce error too. We still estimate I by

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

but now $(X_1, \dots, X_n)$ are the points of a QMC integration rule.

[Fussy point: what if there are repeated values among the X_i? Then does the uniform distribution on X_i weight points proportionally to their number of occurences, or does it weight all occuring points once? The answer depends on whether $X_1, \dots, X_n$ are interpreted as an n-tuple or as a set. In class I took X_i with a uniform distribution on $i \in \{1, \dots, n\}$.]

2.1 Discrepancy

To make progress on this idea we need to define a measure of uniformity of points. In an integration problem $I = \int f(x)dx$ is the expectation of f with respect to the continuous uniform distribution on $[0,1]^d$ while the estimate $\hat{I} = (1/n)\sum_{i=1}^{n} f(X_i)$ is the expectation of f with respect to the discrete uniform distribution with probability $1/n$ on X_i. Therefore a reasonable way to measure uniformity of a list of n points is through a distance between the discrete uniform distribution on those points and the continuous uniform distribution on their domain.

There are many ways to define a distance between distributions. One that has served well in the theory of quasi-Monte Carlo is the star discrepancy, developed next. First we generalize the notion of an interval to d dimensions.

Definition 2.1 *For $a, b \in \mathbb{R}^d$, the half-open interval $[0, a)$ is the set*

$$\prod_{j=1}^{d}[0, a^j) = \Big\{x \mid 0 \le x^j < a^j,\ j = 1, \ldots, d\Big\}$$

and the half-open interval $[a, b)$ is the set

$$\prod_{j=1}^{d}[a^j, b^j) = \Big\{x \mid a^j \le x^j < b^j,\ j = 1, \ldots, d\Big\}.$$

Definition 2.2 *The star discrepancy of $X_1, \ldots, X_n$ is*

$$D_n^*(X_1, \ldots, X_n) = \sup_{a \in [0,1)^d} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \in [0,a)} - \prod_{j=1}^{d} a^j \right|. \tag{2.1}$$

Open and closed intervals are defined analogously. An interval with lower endpoint 0 is also called an "anchored box" where the more general interval is an "un-anchored box".

The value $|(1/n)\sum_{i=1}^{n} 1_{X_i \in [0,a)} - \prod_{j=1}^{d} a^j|$ compares the fraction of points in the interval $[0, a)$ with its volume. The star discrepancy (2.3) is the supremum of such differences over all intervals $[0, a)$. For $d = 1$ it reduces to the Kolmogorov-Smirnov distance, except that the Kolmogorov-Smirnov distance is defined for distributions on $\mathbb{R}$, not just on the unit interval.

The origin plays a special role in the star discrepancy, as it belongs to every interval for which volume and the fraction of points are compared. Other discrepancies can be defined by altering the collection of sets over which the supremem in (2.3) is taken. Two such examples are:

Definition 2.3 *The discrepancy is*

$$D_n(X_1, \ldots, X_n) = \sup_{a<b,\ a,b \in [0,1)^d} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{a \le X_i < b} - \prod_{j=1}^{d} (b^j - a^j) \right|. \tag{2.2}$$

Definition 2.4 *The isotropic discrepancy is*

$$J_n(X_1, \ldots, X_n) = \sup_{convex\ C \subseteq [0,1)^d} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \in C} - \mathit{Vol}(C) \right|, \tag{2.3}$$

where $\mathit{Vol}(C)$ denotes the d dimensional volume of the convex set C.

Discrepancies over different classes, such as all hyper-rectangles not necessarily parallel to the sides of $[0, 1)^d$, or hyper-triangles, or spheres have also been investigated. [Beck (198x) did this. For these other classes one can do only a little better than Monte Carlo unless d is small.]

The practical use of discrepancies is that they can be used to bound integration error. A typical bound takes the form $|\hat{I} - I| \le \mathrm{DISC}(X_1, \ldots, X_n)\mathrm{NORM}(f)$

where DISC is a measure of the discrepancy of $X_1, \ldots, X_n$ and NORM is a measure of the size of the integrand f. Then an integration rule with a small discrepancy will give a small error over a whole class of integrands $\{f \mid \mathrm{NORM}(f) \le C\}$ for $C > 0$. The best known example is the Koksma-Hlawka inequality, described in Chapter xxx, defined in terms of the star discrepancy.

There are bounds among the discrepancies. For example Niedereitter and Wills (1976) show that $D_n \le J_n \le 4dD_n^{1/d}$. The first inequality is trivial. The second shows that although $D_n \to 0$ implies $J_n \to 0$, there may be a substantial dimension effect.

Proposition 2.1 $D_n^* \le D_n \le 2^d D_n^*$.

The key to the right hand inequality in Proposition 2.1 is the decomposition of the sub-interval $[a, b)^d$ into 2^d pieces. For $d = 2$ the decomposition may be written

$$\begin{aligned}&1_{[a^1,b^1)\times[a^2,b^2)}(x)\\ =&1_{[0,b^1)\times[0,b^2)}(x) - 1_{[0,b^1)\times[a^2,b^2)}(x) - 1_{[a^1,b^1)\times[0,b^2)}(x) + 1_{[0,b^1)\times[0,b^2)}(x), \quad (2.4)\end{aligned}$$

as illustrated in Figure xxx. Integrating the decomposition in (2.4) over $x \in [0, 1)^2$ gives a decomposition of the volume of $[a, b)$ into an alternating sum of volumes of intervals anchored at 0. Similarly summing (2.4) over a set of points $X_i \in [0, 1)^2$ gives a decomposition of counts. For general d there are 2^d parts in the decomposition, each having a discrepancy smaller than D_n^*.

Proof of Proposition 2.1
The left inequality in is trivial. Define the vector $ab_{[u]}$ through

$$ab_{[u]}^j = \begin{cases} b^j, & j \in u \\ a^j, & j \notin u. \end{cases}$$

Then by an inclusion-exclusion argument

$$1_{x\in[a,b)} = \sum_{u\subseteq\{1,\ldots,d\}} (-1)^{d-|u|} 1_{x\in[0,ab_{[u]})}$$

and so

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n 1_{X_i\in[0,ab_u)} - \prod_{j=1}^d (b^j - a^j)\right| &= \left|\sum_u \Big(\frac{1}{n}\sum_{i=1}^n 1_{X_i\in[a_{[u]},b)} - \prod_{j=1}^d ab^j\Big)\right| \\ &\le \sum_u \left|\Big(\frac{1}{n}\sum_{i=1}^n 1_{X_i\in[a_{[u]},b)} - \prod_{j=1}^d (b^j - a_{[u]}^j)\Big)\right| \\ &\le \sum_u D_n \\ &= 2^d D_n. \ \square\end{aligned}$$

Proposition 2.2 (Koksma) *For $d = 1$ and f differentiable,*

$$\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int_0^1 f(x)dx\Big| \le D_n^* \int_0^1 |f'(x)|dx.$$

Proposition 2.3 (Hlawka) *For $d \ge 1$*

$$\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int_{[0,1]^d} f(x)dx\Big| \le D_n^* V_{HK}(f),$$

where V_{HK} denotes the variation of f in the sense of Hardy and Krause.

Hlawka's theorem is known as the Koksma-Hlawka inequality as it generalizes Koksma's earlier theorem. The significance of this result is that a rate of convergence for D_n^* translates into a rate of convergence for an upper bound on $\hat{I} - I$.

2.2 Dimension $d = 1$

Given $d = 1$ and a sample size n, by almost any reasonable measure the most uniform set of points is obtained by spliting $[0, 1)$ into n congruent intervals $[(i-1)/n, i/n)$ for $i = 1, \ldots, n$, and using their centers $x_i = (i - 1/2)/n$. Niederreiter (1992) shows that

$$D_n^* = \frac{1}{2n} + \max_{1\le i\le n}\Big|X_{(i)} - \frac{i-1/2}{n}\Big|$$

where $X_{(i)}$ is the i'th largest of the X_i. Thus the midpoint rule minimizes D_n^*. It also minimizes D_n.

The problem becomes more difficult if we would like the $n + 1$ point rule $X_1, \ldots, X_{n+1}$ to include all the points of the n point rule. The $n + 1$ point midpoint rule does not extend the n point midpoint rule by the addition of a single point. We would like to find an infinite sequence X_i for $i \ge 1$ with a small discrepancy D_n or D_n^* for each initial subsequence.

If one were to construct a rule by hand, adding one point at a time, a natural rule to use would be something like $1/2, 1/4, 3/4, 1/8, 5/8$ and so forth. One takes the first point in the middle, leaving two equal width intervals on either side of it. The second point might reasonably be in the middle of one of these intervals, say the left one. Then the third point ought to be in the middle of the right interval. The first three points now leave four equal width intervals that one could split. Whichever one we pick to split, the next one should be on the other side of $1/2$. So we might take $1/8$ as the fourth point, then $5/8$ as the fifth.

By writing out i and Z_i in base 2 we see a pattern in this greedy selection method. Write $i = \sum_{k=0}^{\infty} a_k(i)2^k$ for bits $a_k(i) \in \{0, 1\}$ and then write $Z_i = \sum_{k=0}^{\infty} a_k(i)2^{-1-k}$. The pattern is the van Der Corput scheme. The van Der Corput scheme usually starts at $X_0 = 0$. It generalizes to arbitrary integer bases $b \ge 2$.

X_i	i base 2	X_i base 2
1/2	1	.1
1/4	10	.01
3/4	11	.11
1/8	100	.001
5/8	101	.101
3/8	110	.011
7/8	111	.111
1/16	1000	.0001
9/16	1001	.1001

Table 2.1: Van Der Corput sequence

Definition 2.5 *The radical inverse function ϕ_b in base $b \geq 2$ takes integers $i \geq 0$ to real values in $[0,1)$. Let $i = \sum_{k=0}^{\infty} a_k(i) b^k$ where $0 \leq a_k(i) < b$. Then*

$$\phi_b(i) = \sum_{k=0}^{\infty} a_k(i) b^{-k-1}.$$

A classical method for generating an equidistributed sequence of points in the unit interval is to take $X_n = n\theta \mod 1$ where θ is an irrational number. It is clear that a rational θ would not be effective because then X_n would have a finite period. Figure xxx-(a) shows the first 50 points generated by this method for $\theta = \sqrt{2}$, and Figure xxx-(b) shows the first 50 points generated for $\theta = \sqrt{101}$. The points for $\theta = \sqrt{2}$ have smaller discrepan Theoretical analysis of these rules show that the best values of θ are those for which continued fraction approximations

$$\theta = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}$$

for $a_j \geq 1$, converge the slowest.

A second classical method has $X_n = f(n) \mod 1$ for a function f satisfying $f(n+1) - f(n) \to 0$ and $n(f(n+1) - f(n)) \to \infty$ as $n \to \infty$. Figure XXX shows some examples.

2.3 Halton and Hammersley sequences

The greatest need for QMC methods is not for $d = 1$, but for d large enough that iterated one dimensional rules are ineffective. One of the simplest methods for higher d is the Halton sequence. The Halton sequence uses radical inverse generators in bases $b_j \geq 2$ for $j = 1, \ldots, d$. In order for these points to be equidistributed it is necessary for b_j to be relatively prime. That is for $j \neq k$ the bases b_j and b_k should not both be divisible by any positive integer other than 1.

Definition 2.6 (Halton (1960)) *The Halton sequence in* $[0,1)^d$ *has*

$$X_i^j = \phi_{b_j}(i), \quad i \geq 0, \quad 1 \leq j \leq d,$$

where $b_1 = 2$*,* $b_2 = 3$*, and* b_k *is the* k*'th largest prime number.*

Figure xxx shows the first 72 points of the Halton sequence for $d = 2$.

The radical inverse points $\phi_b(0), \ldots, \phi_b(b^r - 1)$ contain 1 point in each of b^r equal length subintervals $[(k-1)/b^r, k/b^r)$ for $k = 1, \ldots, b^r$. The same holds for the points $\phi_b(\ell b^r), \ldots, \phi_b((\ell+1)b^r - 1)$ for integers $\ell \geq 0$. If we consider two relatively prime bases b_1 and b_2 then the first $b_1^{r_1} b_2^{r_2}$ points of the sequence $(\phi_{b_1}(i), \phi_{b_2}(i))$ have one point in each equal area subinterval $\prod_j [(k_j - 1)/b_j^{r_j}, k_j/b_j^{r_j})$ for $k_j = 1, \ldots, b_j^{r_j}$ and $j = 1, 2$. As in the one dimensional case, consecutive blocks of $b_1^{r_1} b_2^{r_2}$ points are also stratified. The Halton sequence obeys a full d dimensional stratification when $\prod_{j=1}^d b_j$ points (or a multiple thereof) are used.

The reason for using the first d primes is that smaller bases give better uniformity than larger ones. The smallest d relatively prime integers (greater than or equal to 2) are the first d primes. In practice the sequence need not be started at $X_0 = (0, \ldots, 0)$. Any n consecutive values $X_{m+1}, \ldots, X_{m+n}$ could be used.

The Halton sequence is surpassed by more recent constructions. But it remains useful. It can be used for any number of sample points n in any dimension d. One can "add a dimension" to an existing simulation. Having sampled $f(X_i)$ on $x \in [0,1)^d$ we might later modify f into a function g on $[0,1)^{d+1}$. It is straightforward to go back and calculate what X_i "would have been" had it been $d+1$ dimensional instead of d dimensional.

The Halton sequence allows us to simulate to whatever value n that we wish. If however we know the value of n in advance, then a scheme of Hammersley can be better.

Definition 2.7 (Hammersley (1960)) *The Hammersley sequence in* $[0,1)^d$ *has* $X_i^1 = i/n$ *for* $i = 0, \ldots, n-1$ *and* $X_i^j = \phi_{b_{j-1}}(i)$ *for* $j = 2, \ldots, d$ *where* $b_1 = 2$*,* $b_2 = 3$*, and* b_k *is the* k*'th largest prime number.*

The Hammersley scheme samples the first variable X_i^1 with equispaced points and then uses a $d-1$ dimensional Halton sequence for the rest of the variables. By taking smaller bases, better equidistribution is obtained. In practice the first dimension could as well be taken to be $X_i^1 = (i+1/2)/n$ and the others can be any n consecutive values from a $d-1$ dimensional Halton sequence.

Proposition 2.4 *For the Halton sequence with* $n \geq 1$*,*

$$D_n^*(X_0, \ldots, X_{n-1}) < \frac{d}{n} + \frac{1}{n} \prod_{j=1}^{d} \left(\frac{b_i - 1}{2 \log(b_j)} \log(n) + \frac{b_i + 1}{2} \right), \tag{2.5}$$

while for the Hammersley sequence

$$D_n^*(X_0,\dots,X_{n-1}) < \frac{d}{n} + \frac{1}{n}\prod_{j=1}^{d-1}\left(\frac{b_i-1}{2\log(b_j)}\log(n) + \frac{b_i+1}{2}\right). \qquad (2.6)$$

Proof These are Theorems 3.6 and 3.8 of Niederreiter (1992). □

From Proposition 2.4 we see that for a Hammersley sequence $D_n^* = O(\log(n)^{d-1}/n)$ while the Halton sequence satisfies $D_n^* = O(\log(n)^d/n)$. All n points of the n-point Halton sequence are also points of the $n+1$-point Halton sequence. The Hammersley sequence, by contrast, is not extendable this way but gets a smaller discrepancy by a factor of $\log(n)$.

The Halton sequence attains a discrepancy of $O(n^{-1}(\log n)^d)$ and so for a function of bounded variation in the sense of Hardy and Krause, $|\hat{I} - I| = O(n^{-1}(\log n)^d$. At first sight, this rate compares quite favorably with $O_p(n^{-1/2})$, the rate for Monte Carlo. Because $\log(n) = o(n)$ as $n \to \infty$, we find $n^{-1}(\log n)^d = O(n^{-1+\epsilon})$ for any $\epsilon > 0$. It appears that QMC can almost turn $n^{-1/2}$ errors into n^{-1}. If so it would be like using MC with $O(n^2)$ points.

There are examples where QMC can work that well. But for even moderately large d the asymptotic bound is too remote to ensure a good result from QMC. Powers of $\log(n)$ are not necessarily negligible compared to n until n is extremely large. For example consider the Hammersley rule for $d = 10$, which has error rate $n^{-1}(\log n)^9$. It is true that for some n_0 we have $n^{-1}(\log n)^9 \le n^{-1/2}$ whenever $n \ge n_0$. But the required value of n_0 is between 10^{34} and 10^{35}, past any reasonable sample size. This comparison ignores the constants in front of the rate, but the Hammersley constant can be much larger than the MC one, and it would have to be much smaller in order to bring n_0 down to a reasonable sample size.

Exercise: Consider the function $f(X) = 12^{d/2}\prod_{j=1}^d (X^j - 1/2)$. Show that it has variance $\sigma^2 = 1$ regardless of d. Show that $V_{HK}(f)$ increases exponentially in d. For $d = 10$, find n_0 such that the bound on D_n^* in (2.6) times $V_{HK}(f)$ is smaller than $n^{-1/2}$ whenever $n \ge n_0$.

It turns out that there is a big difference between reducing a bound on the error and reducing the error itself. The quantity $\log(n)^d/n$ increases with n until $n = e^d$ then it decreases. If reducing a bound always gave better results then $n-1$ would be a better sample size than n when $n < e^d$. It seems clear that the bound is quite loose until n is very large. [Actually some of these bounds are also tight. This means that there is a function f for which $|\hat{I} - I| = D_n^* V_{HK}(f)$. A bound can be tight for some function without being at all close for another.]

Empirical investigations repeatedly find that QMC is better than MC. Depending on the function tried, QMC can be far better or slightly better, or sometimes slightly worse, than MC. Because the points are deterministic, Bahvalov's theorem shows that we can construct pathological examples in which QMC is far worse than MC.

The explanation for the success of QMC is that the QMC integration rules used in practice tend to have very uniform low dimensional projections. Then

when the integrand is dominated by it's low dimensional parts QMC performs far better than MC. When the integrand is dominated by high dimensional parts, QMC does not improve over MC by much.

2.4 (t, m, s)-nets and (t, s)-sequences

One problem with the Halton sequence is that as d increases a larger value of n is required to get meaningful stratification. For $d = 5$, consecutive blocks of $2 \times 3 \times 5 \times 7 \times 11 = 2310$ points have a full 5 dimensional stratification. For $d = 10$, the product of the first 10 primes is 6469693230, so that no 10 dimensional stratification appears until over 6 billion points have been used.

A second problem with the Halton sequence is that pairs (X^1, X^2) are stratified in consecutive blocks of 6 points while pairs (X^1, X^3) are stratified every 10 points and pairs (X^2, X^3) are stratified every 15 points. It would be better to have a rule where all pairs of variables are stratified on the same blocks of points, just as holds for randomized orthogonal arrays of strength $t \geq 2$.

What is needed is something like a Halton sequence with the same base b used for all dimensions. The solution is found in (t, m, s)-nets as described below.

Definition 2.8 *Let $s \geq 1$ and $b \geq 2$ be integers. An elementary interval in base b is a subinterval of $[0, 1)^s$ of the form*

$$E = \prod_{j=1}^{s} \Big[\frac{c_j}{b^{k_j}}, \frac{c_j + 1}{b^{k_j}} \Big)$$

for integers k_j, c_j with $k_j \geq 0$ and $0 \leq c_j < b^{k_j}$.

Figure xxx shows some elementary intervals in base $b = 3$ and dimension $s = 2$. Elementary intervals are also called b-ary boxes, b-adic intervals, or cells. Figure xxx shows the points of a $(0, 3, 2)$-net in base 5.

Definition 2.9 *Let $m \geq 0$, $b \geq 2$ and $s \geq 1$ be integers. Let $n = b^m$. The finite sequence $(X_i)_{i=1}^n$ of points from $[0, 1)^s$ is a $(0, m, s)$-net in base b if every elementary interval E in base b of volume b^{-m} contains exactly 1 of the points X_i.*

Definition 2.9 stipulates that every b-ary box that "should" have one point of the sequence is required to have exactly one point of the sequence. This is a very strong multiple stratification and by weakening it somewhat, constructions for more values of s and b become available.

Definition 2.10 *Let $t \leq m$ be a nonnegative integer. A finite sequence of b^m points from $[0, 1)^s$ is a (t, m, s)-net in base b if every elementary interval in base b of volume b^{t-m} contains exactly b^t points of the sequence.*

Cells that "should" have b^t points do have b^t points. Smaller values of t imply stronger equidistribution statements. The nearly vacuous case $t = m$ merely states that all points of the sequence are in $[0,1)^s$.

Definition 2.11 *For $t \geq 0$, an infinite sequence $(X_i)_{i\geq 1}$ of points from $[0,1)^s$ is a (t,s)-sequence in base b if for all $k \geq 0$ and $m \geq t$ the finite sequence $(X_i)_{i=kb^m+1}^{(k+1)b^m}$ is a (t,m,s)-net in base b.*

A (t,s)-sequence is more flexible than a (t,s)-net because we can take as many or as few points as we wish from it. In particular we may use n points and then decide whether we want to continue or stop. It is less flexible than a Halton sequence because there is no way to add an $s+1$st input variable.

Faure (1982, Theorem 1) provides a construction of $(0,m,s)$-nets and $(0,s)$-sequences in base p where $p \geq s$ is a prime number. Faure (1982, Theorem 4(ii)) also proves that for a $(0,s)$-sequence, in base $b \geq s \geq 2$, $D_n^* = O((\log n)^s/n)$. The Hammersley trick of adding one equispaced variable also works for Faure's $(0,m,s)$-net allowing the construction of a $(0,m,p+1)$-net in base p for prime p.

The $(0,3,2)$-net in base 5 shown in Figure xxx is in fact the first 2 dimensions of the points of a $(0,3,5)$-net in base 5. For each vector of scales $(k_1,\ldots,k_5)$ with $k_j \geq 0$ and $\sum_{j=1}^5 k_j = 3$, there are 125 rectangular cells of volume $1/125$ in $[0,1)^5$ that each contain exactly 1 of the 125 points. Some combinatorial arguments show that there are 35 such tilings, and so $n = 125$ points of the net manage to balance $35 \times 125 = 4375$ cells of volume 125. Of these only $5 \times 125 = 625$ would have been balanced in a Latin hypercube sample. There is considerable power in such stratification. It would have been impossible to take explicit account of 4375 control variates.

Proposition 2.5 *The star discrepancy of a (t,m,s)-net in base b with $m > 0$ satisfies*

$$D_n^* \leq B(s,b)b^t(\log n)^{s-1} + O((\log n)^{s-2})$$

where

$$B(s,b) = \begin{cases} \left(\frac{b-1}{2\log b}\right)^{s-1} & s = 2, \text{ or } b = 2, s = 3, 4 \\ \frac{1}{(s-1)!}\left(\frac{\lfloor b/2\rfloor}{\log b}\right)^{s-1} & \text{otherwise.} \end{cases}$$

Proof From Theorems 4.10 of Niederreiter (1992). □

As $s \to \infty$, the constant multiplying $(\log n)^{s-1}/n$ is much more favorable for a $(0,m,s)$-net in base s than for the Hammersley sequence. Suppose that the $(0,m,s)$-net has for its base the smallest prime $p \geq s-1$ using the Hammersley trick if necessary. Then it can be shown that the lead constant goes to ∞ faster than exponentially for the Halton sequence and it goes to 0 faster than exponentially for the $(0,,m,s)$-net.

In the 1960's Sobol' constructed and studied (t,m,s)-nets and (t,s)-sequences in base 2. See Sobol' (19xx) and references in Niederreiter (198x,1992). The Sobol' sequences are a widely used alternative to the Faure sequences. Points in

base 2 can be generated by bit manipulation methods that often result in faster generation.

Niederreiter (1987) extended Faure's constructions to bases $b \geq s$ that are prime powers, and to bases that are products of such prime powers. Niederreiter (1987) also merged Faure's and Sobol's concepts to produce the definitions given above.

In general, if $(X_i)_{i=1}^n$ is a (t, m, s)-net in base b then $D_n^* = O(\log(n)^{s-1}/n)$, and if $(X_i)_{i\geq 1}$ is a (t, s)-sequence in base b then $D_n^* = O(\log(n)^s/n)$. Niederreiter (1992, Theorems 4.10 and 4.17) gives more precise statements of these facts.

To conclude this section, suppose for some integer $1 \leq \lambda < b$ that (X_i) comprises the first λb^m points of a (t, s)-sequence in base b. Then each elementary interval of volume b^{t-m} has λb^t points in it. Thus (X_i) is equidistributed over the same set of elementary intervals as is a (t, m, s)-net, but if $\lambda > 1$, it is not a (t, m, s)-net because λb^m is not a power of b. Another equidistribution property of (X_i) is as follows: no elementary interval of volume b^{t-m-1} has more than b^t points in it. This holds because such an elementary interval has only b^t points of the first b^{m+1} points of the (t, s)-sequence.

Definition 2.12 *Let s, m, t, b, λ be integers with $s \geq 1$, $m \geq 0$, $0 \leq t \leq m$, $b \geq 2$ and $1 \leq \lambda < b$. A sequence (X_i) of λb^m points is called a (λ, t, m, s)-net in base b if every elementary interval in base b of volume b^{t-m} contains λb^t points of the sequence and no elementary interval in base b of volume b^{t-m-1} contains more than b^t points of the sequence.*