THE SCRAMBLES OF HALTON SEQUENCE AND THIER WEAKNESSES

A. MOGHARRABI OSTADKALAYEH AND B. FATHI VAJARGAH*

ABSTRACT. So far, many scrambles have been introduced to break the correlation between Halton's sequence points and improve its two-dimensional designs. In this paper, some of the most important scrambles that are available to scrambling the Halton sequence are evaluated, and describe their weaknesses. Also, we introduce a new method that, despite it's simplicity of execution, has good two-dimensional designs.

1. INTRODUCTION

Quasi-Monte Carlo methods are a variant of ordinary Monte Carlo methods that employ highly uniform quasirandom numbers in place of Monte Carlos pseudorandom numbers. Clearly, the generation of appropriate high-quality quasirandom sequences is crucial to the success of using quasi-Monte Carlo methods. The halton sequence, which is one of the standard low-discrepancy sequences, and one of it's important advantages is that the halton sequence is easy to implement due to its definition via the radical inverse function. However,the original halton sequence suffers from correlations between radical inverse functions with different bases used for different dimensions. These correlations result in poorly distributed two-dimensional projections. A standard solution

to this problem is to use a randomized (scrambled) version of the halton sequence which have both good two-dimensional projections and a smaller discrepancy, which is a measure of deviation from uniformity [1].

So far, many scrambles have been proposed for the halton sequence, which we intend to examine in some of the most important ones here. These scrambles are listed here in chronological order, and each has tried to upgrade them by fixing any deficiencies.

2. The halton sequence

Let $p \geq 2$ be an integer, then any integer $n \geq 0$ can be written in the form

$$(2.1) \qquad n = a_0 + a_1 p + \cdots + a_m p^m$$

where, $0 \leq a_j \leq p$ and $m = \lceil log_p n \rceil$ is the maximum number of digits needed to represent all n-values.

The radical inverse function ϕ for base p is defined by

$$(2.2) \qquad \phi_p(n) = \frac{a_0}{p} + \frac{a_1}{p^2} + \cdots + \frac{a_m}{p^{m+1}}$$

The Van der Corput sequence in base p is defined as the one-dimensional point set $\{\phi_p(n)\}_{n=0}^{\infty}$. Halton (1960) extended this definition to the s-dimensional sequence as

$$(2.3) \qquad X_n = (\phi_{p_1}(n), \phi_{p_2}(n), \ldots, \phi_{p_s}(n))$$

where $n = 0, 1, 2, \ldots$ and the dimensional bases $p_1, p_2, \cdots, p_s$ are pairwise coprime.In practice, we always use the first S primes as the bases.

As mentioned, the Halton sequence has poor two-dimensional projections (for dimensions greater than 10) because of the correlation between the radical inverse functions used for different dimensions [1]. Examples of these poor projections can be seen in Fig.1 (b, c and d).

In Fig.1 it can be seen that for small bases, the dispersion of the sequence points is acceptable. As the dimensions increase, the points fall into clusters of parallel lines with the $y = x$ line [1], and as these dimensions become larger, the number of clusters decreases, but are closer together, and the points accumulate in them, and the space becomes wider. All of the scrambles introduced are intended to break these lines and thus fill the space more uniformly.

The correlation between points of the Halton sequence can be broken by scrambling the digits of the sequence in a way that preserves the low-discrepancy properties. This was first formally described by Braaten and Weller [3], who defined the scrambled radical inverse function $S_p(n)$ as

$$(2.4) \qquad S_p(n) = \frac{\pi_p(a_0)}{p} + \frac{\pi_p(a_1)}{p^2} + \cdots + \frac{\pi_p(a_m)}{p^{m+1}}$$

here, $\pi_p(a_j)$ is a permutation on the digits $0,1,\ldots,p-1$ which holds the digit 0 fixed. There are $(p-1)!$ such permutations. The scrambled halton sequence is then given by

$$(2.5) \qquad X_n = (S_{p_1}(n), S_{p_2}(n), \ldots, S_{p_s}(n))$$

Note that since each set of permutations (one permutation for each dimension) always leads to the same scrambled version of the Halton sequence, the scrambling defined by 2.4 and 2.5 is a deterministic scrambling. If one is searching for the best possible deterministic scrambling under some criterium, then the search space should only contain permutations $\pi_p(a_j)$ [3]. Fig.2 show how the permutations break up the correlation between the coordinates of the Halton sequence.

3. THE HALTON SEQUENCE SCRAMBLES

In this section, we give a chronological overview of the different deterministic scramblings that appear in the literature.

3.1. **Warnock's folded radical inverse.** Warnock used the folded radical inverse function

$$(3.1) \ \ \psi_p(n) = \frac{(a_0 + 0) \, mod \, p}{p} + \frac{(a_1 + 1) \, mod \, p}{p^2} + \cdots + \frac{(a_m + m) \, mod \, p}{p^{m+1}}$$

instead of 2.2 to define a scrambled version of thehalton sequence.Fig.3 shows the two-dimensional projection implemented for the original halton sequence and the Warnock FRI sequence for several dimensions for comparison.This figure shows that the Warnock scramble did not make much of a change in the original sequence and poor two-dimensional projection still remains. Therefore, the performance of this scramble is not satisfactory.

3.2. **The permutations of Braaten and Weller.** Perhaps the most important scramble created for the halton sequence, which many other researchers have cited and come up with, is that of Bratton and Weller. Braaten and Weller used Algorithm 1 to define the π_p from the scrambled radical inverse function 2.4. They tabulated their permutations up to the first 16 primes.The L_2-star discrepancy plots(Fig.4) for 8, 12 and 16-dimensional sequences demonstrate that the BW-sequence has a lower L_2-star discrepancy than the halton sequence for the first 1000 points.As the dimension increases, this superiority seems to increase, but the difference between the discrepancies decreases with the number of points and the L_2-star discrepancy of the BW sequence approaches the halton sequence [3].

Although,it seems that the BW sequence has much more uniformity than the halton sequence and breaks the correlations between dimensions, the following drawbacks can be noted for this sequence:

- This sequence is tested for the first 1000 points only. They did not show what would happen after these 1000 points[3].
- The permutations are only specified up to the first 16 dimensions.Of course, Vandewoestyne has tabulated these permutations up to 64.
- It's difficult to implement. That is, it is difficult to obtain permutations by assuming that the star-discrepancy is minimized, and much computational effort is needed because of the calculation of many one-dimensional discrepancies[3].
- This method gets rid of some poor projections of the original halton sequence. However, the permutation focuses on optimal halton sequence only in one dimension, and the usage of such a measure makes the selected permutations still highly related for two dimensions whose bases are twin primes[1].

Algorithm 1 Permutation search of Braaten and Weller[3].

Set $\pi_p(0) = 0$
Set *ChoiceSet*=1,2,...,$p-1$
   for $i = 1$ to $p - 1$ **do**
      choose $\pi_p(i)$ from *ChoiceSet* so that it minimizes the one-
         dimensional discrepancy $T_N^* \left(\left\{ \frac{\pi_p(1)}{p}, \ldots, \frac{\pi_p(i)}{p} \right\} \right)$
      Set *ChoiceSet*= *ChoiceSet*\$\pi_p(i)$
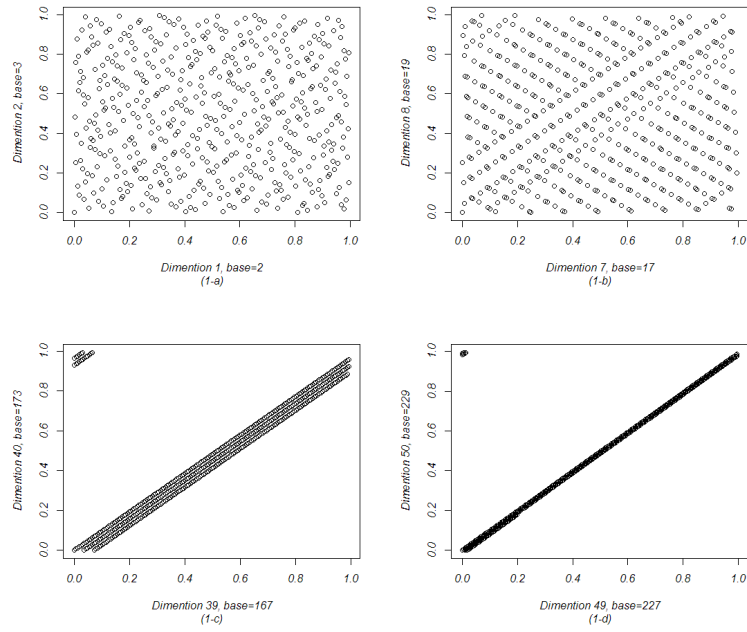   **end for**

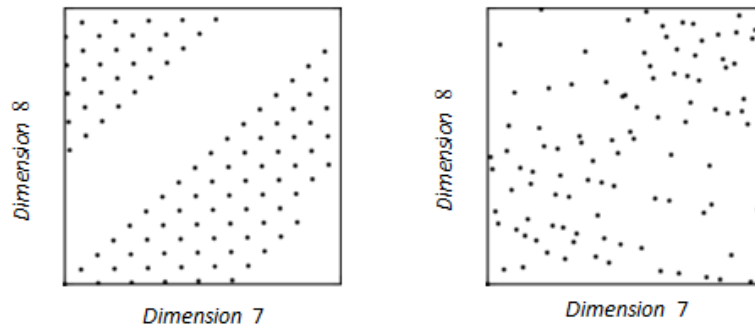FIGURE 1. Different two-dimensional projections of the halton sequence

FIGURE 2. Halton points before (left) and after (right) scrambling.

3.3. **Faure's algorithm for constructing the permutations.** Faure proposed his algorithm (Algorithm 2) to create π_p permutations for the one-dimensional sequence of Van der Corput. The permutations shown in table1 can easily be obtained by using the faure algorithm.

TABLE 1.   Permutations by Faure

π_2	(0 1)	π_6	(0 2 4 1 3 5)
π_3	(0 1 2)	π_7	(0 2 5 3 1 4 6)
π_4	(0 2 1 3)	π_8	(0 4 2 6 1 5 3 7)
π_5	(0 3 2 1 4)	$\vdots$	$\vdots$

The advantage of the faure algorithm is its ease of execution. But for cases where p is even, in permutations always the initial digits are all even and the last digits are all odd.

Algorithm 2 Faure's algorithm to determine π_p [3].

Set $p = 2$
Set $\pi_p = (0, 1)$
repeat
$p = p + 1$
   if p is even then $\pi_p = (2\pi_{\frac{p}{2}}, 2\pi_{\frac{p}{2}} + 1)$
      else
            $\eta = \pi_{p-1}$
            $k = \frac{p-1}{2}$
            add 1 to each element of η which is $\geq k$
            $\pi_p = \eta$ with k added in the middle
   end if
until all necessary permutations are found

3.4. **Warnock's PhiCf sequence.** After the first failed attempt, warnock combined the initial behavior of the *Weyl* sequence with the asymptotic behavior of the halton sequence to construct what he called the PhiCf sequence. The name of this sequence is due to the use of continued fraction expansion in its structure. He replaces each a_i in 2.1 with $S(p)\, a_i$ mod p to obtain a new kind of radical inverse function

$$(3.2) \quad \omega_p(n) = \frac{(S(p).a_0)\,mod\,p}{p} + \frac{(S(p).a_1)\,mod\,p}{p^2} + \cdots + \frac{(S(p).a_m)\,mod\,p}{p^{m+1}}$$
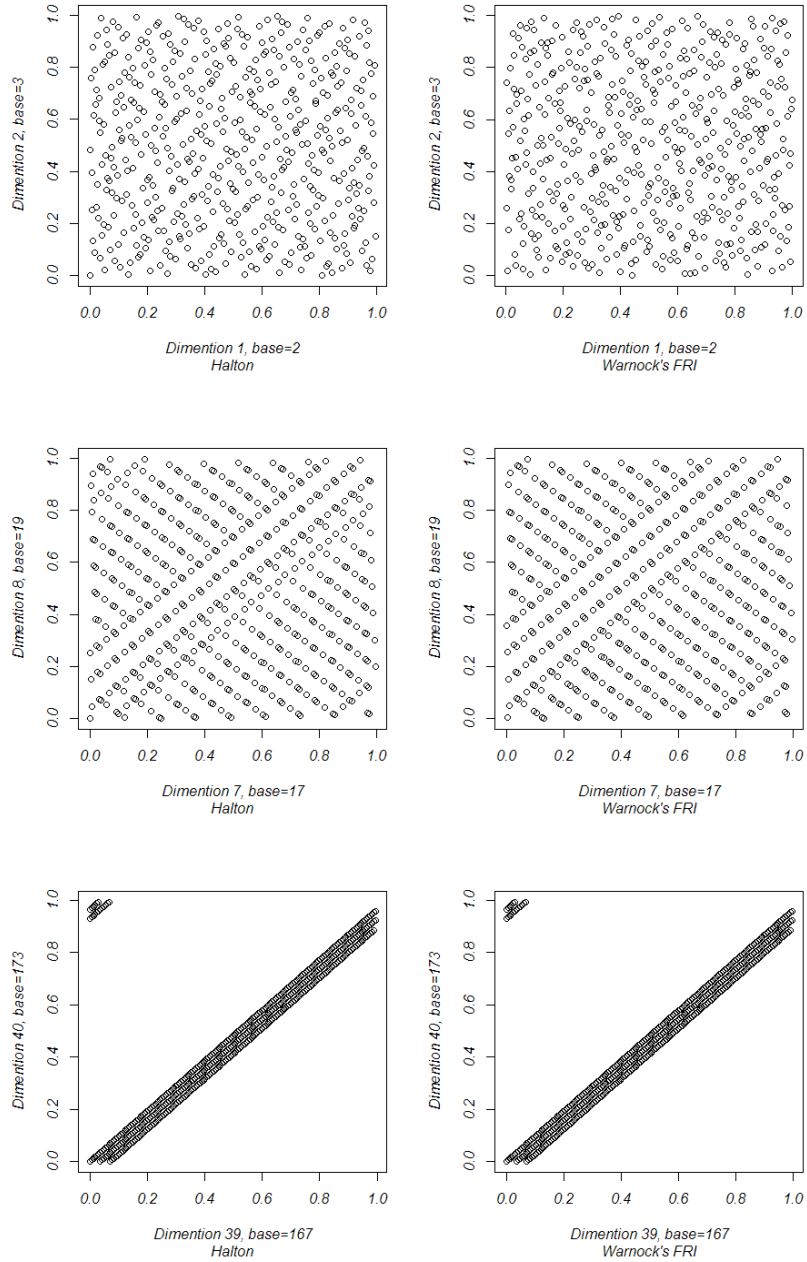
FIGURE 3. 500 point from Halton sequence (left) and Warnock sequence (right).

where $S(p)$ is defined to be a number such that $S(p)/p$ is close to the fractional part of $\sqrt{p}$. The details are given in Algorithm 3 [3].

Since p_i is different for each $S(p_i)$, then the scrambled version of a sequence generated by a different $S(p_i)$ is expected to be independent. This is due to the fact that the Weyl sequences, consisting of multiples of square roots of primes have good discrepancy in low dimension. The square roots of primes are independent[1].

Although the warnock algorithm is an interesting way to find $S(p_i)$, he did not provide a reason for choosing the $S(p_i)$ method in the final step of the algorithm.

Algorithm 3 Warnock's algorithm to determine $S(p)$[3].

Set $X_U = \lceil (p\sqrt{p}) \rceil$ (where $\{x\}$ denotes the fractional part of x)

Set $X_L = \lfloor (p\sqrt{p}) \rfloor$

$\frac{X_U}{p} = [0; a_1, a_2, \ldots, a_k]$ (continued fraction expansion of $\frac{X_U}{p}$)

$\frac{X_L}{p} = [0; e_1, e_2, \ldots, e_m]$ (continued fraction expansion of $\frac{X_L}{p}$)

Choose $S(p)$ as either X_U or X_L according to:
   1. The smaller sum of partial quotients
   2. The smallest largest partial quotient
   3. The nearest to the fractional part of $\sqrt{p}$

3.5. **Mascagni and Chi's optimal scrambling.** Mascagni and Chi [1] considered the linear scrambling $\pi_{p_i}(a_j) = w_i a_j \pmod{p_i}$ with w_i an integer. For a prime modulus p, and a primitive root W modulo p as multiplier, we have that the discrepancy, $D_N^{(2)}$, satisfies

(3.3) $$(p-1)D_{p-1}^{(2)} \leq 2 + \Sigma_{i=1}^q a_i$$

where a_i is the ith digit in the continued fraction expansion of $\frac{W}{p}$ with $a_q = 1$. So our job is reduced to finding a primitive root W_p modulo p such that W_p has the smallest sum of continued fraction expansion digits with $\frac{W}{p} = [a_1, a_2, \ldots, a_q = 1]$, [4]. Using this criterion, they list the result of their search for the best primitive root modulo p for the first 40 dimentions of the halton sequence and empirically verify optimality by approximating a given integral.

The advantage of this method is that it is simple to calculate the primitive roots, but since each number may have several primitive roots and minimizing the sum of the fraction expansion digits may be hold for some of them, they did not specify which one to choose and why?

The other disadvantages of this method is the conditions for choosing the optimal primitive root w_i is only based on p_i. It could happen that for a different p_i, w_i may be close whenever p_i is close. This leads to very high correlations[3].

It is necessary to mention two points about reference [1]:

- It is stated in section 4 "Note $e = p$ is normally considered to be sufficient to ensure small correlation, however, that implies prime pairs of the form p and $p+p^2$". That is not true, since p is a prime number, then $p+p^2$ is a even and not a prime.
- In Table 2 of the optimal coefficients, the number 205 is written in fron of number 213, while 213 is not the prime number at all and should be replaced by 223. Or, for number 11 the value of 3 is written, while 3 is not the primitive root of 11 and the number 2 should be replaced.

3.6. **Vandewoestyne's reverse permutations.** Consider the permutations $\pi_p = (0 \quad p-1 \quad p-2 \quad \ldots \quad 1)$ for the prime bases p of the halton sequence,where each digit a_i is replaced by $p - a_i$, except when it is zero. The generation of these permutations is by far the simplest one among all algorithms mentioned in this paper.

Consider the following functions:

$$(3.4) \quad \delta(p - a_i) = \begin{cases} 0, & if \ a_i = 0 \\ p - a_i, & if \ a_i \neq 0 \end{cases} \quad , \Theta(\frac{1}{p^i}) = \begin{cases} 0, & if \ a_i = 0 \\ \frac{1}{p^i}, & if \ a_i \neq 0 \end{cases}$$

then the scrambled radical inverse function for the reverse halton sequence can be written as:

From these formulas, it is easy to see that a reverse halton point is just a standard halton point shifted by a certain amount. How far we shift for each n depends on the digits of n. A digit $d_i = 0$ means that the shift does not include the term $\Theta\left(\frac{1}{p^i}\right)$ Since different numbers n may contain zero-digits at different places, the shift-size can be different
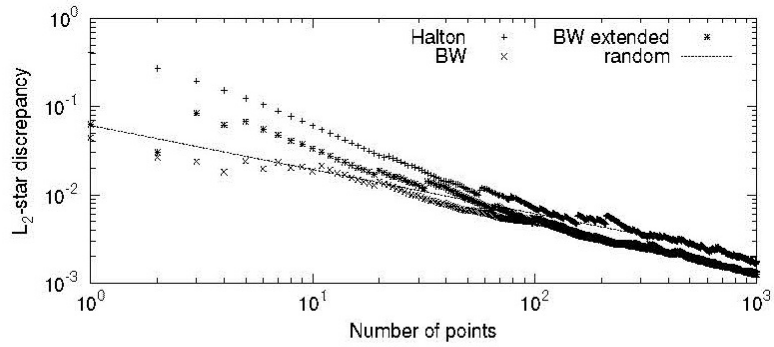
(a) 8 dimensions.
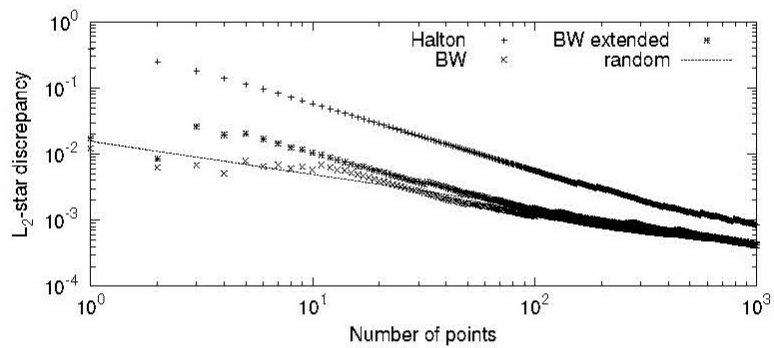
(b) 12 dimensions.
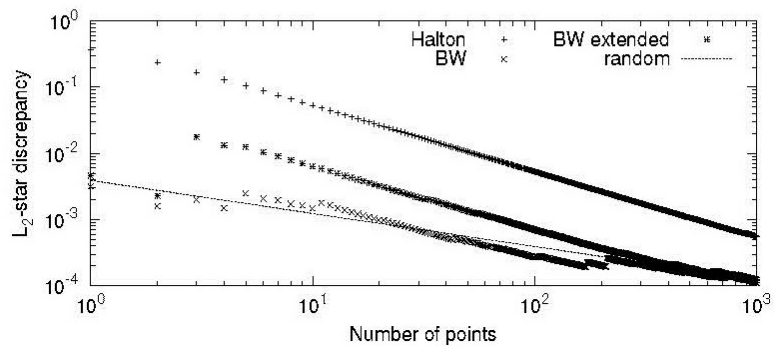
(c) 16 dimensions.

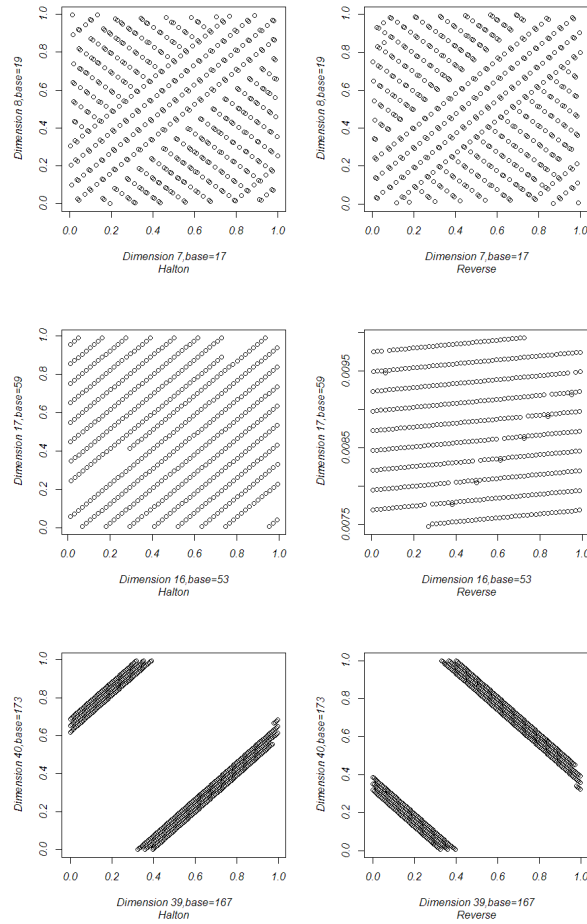FIGURE 4. L_2-star discrepancies for several sequences calculated for dimensions 8, 12 and 16.

FIGURE 5. 500 Halton points (left) and Vandewoestyne's reverse sequence (right)

for different values of n. For groups of n-values containing the same number of zero digits at the same places,the shift will be the same, thereby explaining why we still observe a certain degree of correlation between several two-dimensional projections.(See Fig.5)
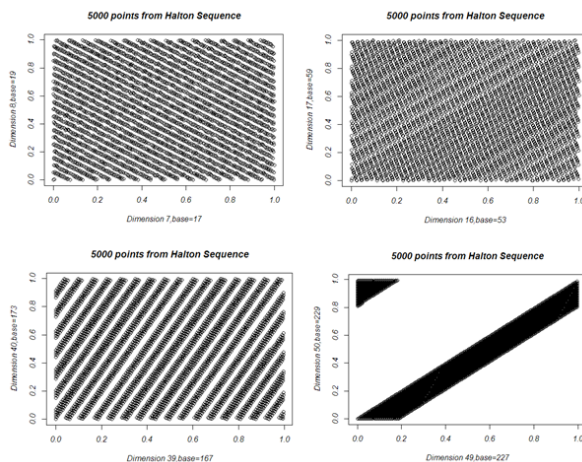
FIGURE 6.   5000 Halton points in two dimentions

4. OTHER METHODS OF IMPROVING THE HALTON SEQUENCE

As discussed, all methods of scrambling the halton sequence have drawbacks that make us hesitant to use them. So, we must look for ways that, despite the simplicity of execution, break the correlation between dimensions in the halton sequence. At least three other methods can be considered as follows:

4.1. **Increasing the number of points.** Previously, Fig.1, we implemented two-dimensional projections of the Halton sequence using 500 points in R software. The problem with these designs is that the points accumulate in some areas, leaving a lot of empty space. The number of points can be increased to fill this gap. Now we bring up the same designs again with more points. (Fig.6) We see that in two-dimensional projections with more points, empty spaces begin to fill. Of course, this method has several problems: First, no matter how many points we get, the correlation between the dimensions does not broken. Second, the larger the bases used, the more points are needed to fill in the gaps. Third, it should be kept in mind that this approach is not useful in high-dimensional problems and will increases costs.

4.2. **Increasing the difference between the bases.** This method is useful when the number of dimensions is small. When the number
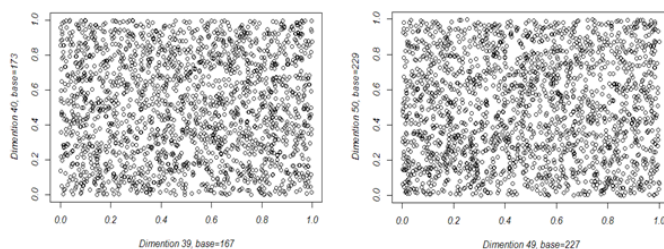
FIGURE 7. 2000 points of the Halton sequence in two dimensions after random sampling

of dimensions is large, it is hardly useful. Increasing the difference between the bases also raises a problem in the upper bound of the star-discrepancy for the halton sequence[1].

4.3. **Sampling from Sequence.** This seems to be the simplest and best way to break the correlation between dimensions. If one can reorder or shuffle the digits of each point in the halton sequence for different dimensions, the correlations between different dimensions can be made very small. This is due to the fact that there are gaps between the most significant digits of $\phi_p(n)$, $p < 10$, which have good two-dimensional projections. However, when $p > 10$, there are no gaps for the most significant digits of $\phi_p(n)$ and the most significant digits cycle from 1 to 9 without jumps.

Random selection of sequence points can partially solve this problem. Fig.7 shows the improvement of the two-dimensional projections mentioned in the previous sections using this method.

5. Conclusion

In this paper, the halton sequence and the various methods used to scramble it are studied. Because of the flaws in these methods, we have proposed methods such as increasing the number of points, Increasing the difference between the bases and Sampling from Sequence. We saw that the random sampling without replacement method, despite the simplicity of execution, significantly increases the uniformity of the distribution of the sequence points. Of course, we should consider other aspects of this method, such as low discrepancy or convergence rate, and so on.

References

[1] Chi, H., Mascagni, M., Warnock, T., *On the Optimal halton Sequences*, Math. Comput. in Simul., **70(1)** (2005), 9–21.

[2] Fathi Vajargah, B., Eskandari Chechaglou, A., *Optimal halton sequence via inversive Scrambling*, Communications in Statistics Simulation and Computation, **42** (2013), 476–484.

[3] Vandewoestyne, B., *Quasi-Monte Carlo techniques for the approximation of high-dimensional integrals*, [PhD theses] Advisor Ronald Cools,June 2008.

[4] M. Mascagni, H. Chi, *On the scrambled Halton sequences*, Monte Carlo Meth. Appl., **10** (2013), 435–442.

Ali Mogharrabi Ostadkalayeh
Department of Statistics, Science and Research branch, IAU Tehran, Iran
Email: mogharrabi.ali@malayeriau.ac.ir

Behrouz Fathi Vajargah
Department of Statistics, University of Guilan, Rasht, Iran
Email: fathi@guilan.ac.ir