The nonzero gain coefficients of Sobol's sequences are always powers of two

Zexin Pan
Stanford University

Art B. Owen
Stanford University

June 2021

Abstract

When a plain Monte Carlo estimate on n samples has variance σ^2/n, then scrambled digital nets attain a variance that is $o(1/n)$ as $n \to \infty$. For finite n and an adversarially selected integrand, the variance of a scrambled (t, m, s)-net can be at most $\Gamma\sigma^2/n$ for a maximal gain coefficient $\Gamma < \infty$. The most widely used digital nets and sequences are those of Sobol'. It was previously known that $\Gamma \leqslant 2^t 3^s$ for Sobol' points as well as Niederreiter-Xing points. In this paper we study nets in base 2. We show that $\Gamma \leqslant 2^{t+s-1}$ for nets. This bound is a simple, but apparently unnoticed, consequence of a microstructure analysis in Niederreiter and Pirsic (2001). We obtain a sharper bound that is smaller than this for some digital nets. We also show that all nonzero gain coefficients must be powers of two. A consequence of this latter fact is a simplified algorithm for computing gain coefficients of nets in base 2.

1 Introduction

Numerical integration is a fundamental task in scientific computation. In high dimensional problems, Monte Carlo (MC) methods are widely used for integration because they are less affected by dimension than classical methods, such as those in [1]. The MC problems we study are to compute $\mu = \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for a dimension $s \geqslant 1$ and most of our attention is on $f \in L^2[0,1]^s$. This μ is the mathematical expectation $\mathbb{E}(f(\boldsymbol{x}))$ for $\boldsymbol{x} \sim \mathbb{U}[0,1]^s$. By using transformations from [2] we can greatly expand MC to expectations of quantities with non-uniform distributions over domains other than the unit cube, and so for this paper it suffices to work with $\boldsymbol{x} \sim \mathbb{U}[0,1]^s$.

The MC estimate of μ is

$$\hat{\mu}_{\mathrm{MC}} = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_i), \quad \boldsymbol{x}_i \overset{\mathrm{iid}}{\sim} \mathbb{U}[0,1]^s. \tag{1}$$

The independent uniform draws $\boldsymbol{x}_i$ will form clusters and leave gaps in $[0,1]^s$. This fact has lead to the development of quasi-Monte Carlo (QMC) methods,

beginning with [23], designed to cover the unit cube more evenly. See [3] for a recent survey. A QMC estimate $\hat{\mu}_{\mathrm{QMC}}$ has the same form as $\hat{\mu}_{\mathrm{MC}}$ from (1) except that n distinct points $\boldsymbol{x}_i \in [0,1]^s$ are chosen deterministically so as to make the discrete uniform distribution on $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}\}$ close to the continuous uniform distribution on $[0,1]^s$, by minimizing a measure of the discrepancy (see [5]) between those distributions.

Using the Koksma-Hlawka inequality [7] it is possible to show that some QMC constructions attain

$$|\hat{\mu}_{\mathrm{QMC}} - \mu| = O(n^{-1}\log(n)^{s-1}) \tag{2}$$

when f has bounded variation in the sense of Hardy and Krause, which we write as $f \in \mathrm{BVHK} = \mathrm{BVHK}[0,1]^s$. See [21] for a description of this variation. A drawback of QMC points is that they do not support a practical strategy to compute the bound in (2). Randomized QMC (RQMC) points $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}$ are constructed so that individually $\boldsymbol{x}_i \sim \mathbb{U}[0,1]^d$ while collectively these points have the low discrepancy that makes (2) hold. Then we can estimate our error statistically, using independent replicates of the randomization procedure. See [9] for a survey of RQMC.

In this paper we focus on perhaps the most widely used QMC method, the Sobol' sequences of [28]. We consider randomizing them with the RQMC method known as scrambled nets from [18]. The MC estimate satisfies

$$\mathbb{E}((\hat{\mu}_{\mathrm{MC}} - \mu)^2) = \frac{\sigma^2}{n}. \tag{3}$$

Thus MC has a root mean squared error (RMSE) of $\sigma/n^{1/2}$. The QMC error in (2) is asymptotically better, but for large s the $\log(n)^{s-1}$ factor leaves room for doubt about QMC at feasible sample sizes.

For scrambled nets we have

$$\mathbb{E}((\hat{\mu}_{\mathrm{RQMC}} - \mu)^2) \leqslant \frac{\Gamma\sigma^2}{n} \tag{4}$$

for a maximal gain coefficient $\Gamma < \infty$, removing the powers of $\log(n)$. If $f \in \mathrm{BVHK}$, then (2) also holds for $\hat{\mu}_{\mathrm{RQMC}}$, so RQMC gets the asymptotic benefit of QMC while (4) bounds how much worse RQMC could be compared to MC for finite n (with an adversarily chosen integrand).

When scrambling the nets taken from Faure sequences [6], it is known from [19] that $\Gamma \leqslant \exp(1) \doteq 2.718$. The nets of Sobol' [28] appear to be more widely used. For them it is known from [20] that $\Gamma \leqslant 2^t 3^s$ where t is the quality parameter that we describe below. In this paper we improve that bound to show that $\Gamma \leqslant 2^{t+s-1}$. This bound can also be deduced from the results of Niederreiter and Pirsic [15], but to our knowledge this has not been remarked on before. We further show that all the nonzero gain coefficients are powers of two and we provide a slight improvement in the microstructure gain bounds from [15].

An outline of this paper is as follows. Section 2 defines digital nets and sequences, and reviews properties of scrambled digital nets. Section 3 proves our bound 2^{t+s-1}. Section 4 proves that nonzero gain coefficients must be powers of 2. Both of these results hold for any scrambled nets in base 2 including those of Sobol' [28] as well as those of Niederreiter and Xing [16, 17] and base 2 nets constructed via polynomial lattice rules, as described in [4]. Section 5 shows that an improved exponent of 2 is possible by sharpening the usual notion of the t parameter for a subset of variables. One concrete example with an improved exponent is provided using shift nets of Schmid [24]. Section 6 has a discussion.

2 Notation and background

In this section we define digital nets and sequences. Then we describe methods of scrambling them and their properties. The key property in this paper is the set of gain coefficients of a digital net.

Throughout this paper we have a dimension $s \geqslant 1$. We write $1{:}s$ for $\{1, 2, \dots, s\}$. We use $\mathbb{Z}$ for the integers, $\mathbb{N}_0$ for non-negative integers, and for integers $n \geqslant 1$ we let $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$. For $u \subseteq 1{:}s$ and $\boldsymbol{x} = (x_1, \dots, x_s) \in [0,1]^s$ we write $\boldsymbol{x}_u$ for the tuple $(x_j)_{j\in u}$. The cardinality of u is written $|u|$. For a set with a lengthy definition, # may be used for cardinality. For a statement S we use $\mathbf{1}_S$ or $\mathbf{1}\{S\}$, depending on readability, to denote a variable that is 1 when S holds and 0 otherwise.

2.1 Digital nets and sequences

We let $b \geqslant 2$ be an integer base in which to represent integers and points in $[0,1)$. We work with half-open intervals because we will need to partition $[0,1)^s$ into congruent subsets. Note that the problems are still defined as $\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ because QMC is strongly connected to Riemann integration [13] and the notion of bounded variation that we use is also defined on closed unit cubes. We begin with some standard definitions.

Definition 1. An s-dimensional elementary interval in base b has the form

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

where $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{Z}^s$ and $\boldsymbol{c} = (c_1, \dots, c_s) \in \mathbb{Z}^s$ satisfy $k_j \geqslant 0$ and $0 \leqslant c_j < b^{k_j}$.

Given $\boldsymbol{k}$, we define $|\boldsymbol{k}| = \sum_{j=1}^{s} k_j$. For a given vector $\boldsymbol{k}$, the $b^{|\boldsymbol{k}|}$ elementary intervals $E(\boldsymbol{k}, \boldsymbol{c})$ partition $[0,1)^s$ into congruent sub-intervals. Ideally they should all get the same number of our integration points $\boldsymbol{x}_i$ and digital nets defined next make this happen in certain cases.

Definition 2. For integers $m \geqslant t \geqslant 0$ and $b \geqslant 2$ and $s \geqslant 1$, a (t, m, s)-net in base b is a sequence $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1} \in [0,1)^s$ for $n = b^m$ where

$$\sum_{i=0}^{n-1} \mathbf{1}\{\boldsymbol{x}_i \in E(\boldsymbol{k}, \boldsymbol{c})\} = nb^{-|\boldsymbol{k}|} = b^{m-|\boldsymbol{k}|}$$

for every elementary interval $E(\boldsymbol{k}, \boldsymbol{c})$ with $|\boldsymbol{k}| \leqslant m - t$.

Other things being equal, we would prefer smaller t and $t = 0$ is the best. For a given m and s and b, the smallest attainable t might be larger than 0. The minT project [26, 27] keeps track of the minimum achieved values of t for given m and s and b along with known lower bounds. When we refer to the value of t for a sequence of points, we mean the smallest value of t for which the sequence is a (t, m, s)-net.

Definition 3. For integers $t \geqslant 0$ and $b \geqslant 2$ and $s \geqslant 1$, a (t, s)-sequence in base b is an infinite sequence $\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots \in [0,1)^s$ such that for any integer $m \geqslant t$ and any integer $r \geqslant 0$ the subsequence

$$\boldsymbol{x}_{rb^m}, \boldsymbol{x}_{rb^m+1}, \dots, \boldsymbol{x}_{(r+1)b^m-1} \in [0,1)^s$$

is a (t, m, s)-net in base b.

The value of (t, s)-sequences is that they provide an extensible set of (t, m, s)-nets. The first b^m points are a (t, m, s)-net in base b and if we increase the sample to b^{m+1} points then we have included $b - 1$ more (t, m, s)-nets and they're carefully constructed to fill the gaps that each other leave, so that taken together they now comprise a $(t, m + 1, s)$-net in base b. Taking b of those $(t, m + 1, s)$-nets yields a $(t, m + 2, s)$-net, and so on. The (t, m, s)-nets that we study are taken to be the first b^m points of a (t, s)-sequence.

The first (t, m, s)-nets and (t, s)-sequences are those of Sobol' [28]. They are all in base $b = 2$. Sobol's construction actually defines a whole family of point sequences, determined by one's choice of 'direction numbers'. Joe and Kuo [8] made an extensive search for good direction numbers and their choices are widely used.

The smallest value of t that one can attain is nondecreasing in s. The most favorable growth rates for t as a function of s are in the (t, s)-sequences of Niederreiter and Xing [16, 17]. These are ordinarily implemented in base 2.

The (t, s)-sequences of Faure [6] have $t = 0$ but they require a prime base $b \geqslant s$. The modern notion of digital nets and sequences is based on the synthesis in [14]. That reference also generalizes Faure's construction to bases $b = p^r$ for a prime number p and integer $r \geqslant 1$.

If $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}$ is a (t, m, s)-net in base b then $\boldsymbol{x}_{0,u}, \dots, \boldsymbol{x}_{n-1,u} \in [0,1)^{|u|}$ form a $(t, m, |u|)$-net in base b. It is common that the quality parameter of these projected digital nets is smaller than the one for the original net. We let t_u be the smallest such t for which $\boldsymbol{x}_{0,u}, \dots, \boldsymbol{x}_{n-1,u} \in [0,1)^{|u|}$ is a $(t, m, |u|)$-net in base b. For theory about t_u see [25], for its use defining direction numbers,

see [8], and for computational algorithms, see [11]. The quality parameter for the first b^m points of a (t,s)-sequence may also be smaller than the value of t that holds for the entire sequence. We will introduce a second quality parameter for a projected (t,m,s)-net in Section 5.

2.2 Scrambling nets

A scrambled net is one where the base b digits of a (t,m,s)-net in base b have been randomly permuted in such a way that the resulting points satisfy $\boldsymbol{x}_i \sim \mathbb{U}[0,1]^s$ individually while the ensemble $\boldsymbol{x}_0,\dots,\boldsymbol{x}_{n-1}$ is still a (t,m,s)-net in base b with probability one. See [18] for the details of a nested uniform scramble and [12] for a random linear scramble of Matoušek that requires less storage.

The nested uniform scrambling has the following properties:

$$\begin{aligned}
&\mathbb{E}(\hat{\mu}_{\mathrm{RQMC}}) = \mu && f \in L^1[0,1]^s,\\
&\Pr\bigl(\lim_{n\to\infty} \hat{\mu}_{\mathrm{RQMC}} = \mu\bigr) = 1 && f \in L^{1+\epsilon}[0,1]^s, \quad \text{some } \epsilon > 0,\\
&\mathrm{var}(\hat{\mu}_{\mathrm{RQMC}}) = o(1/n) && f \in L^2[0,1]^2,\\
&\mathrm{var}(\hat{\mu}_{\mathrm{RQMC}}) \leqslant \Gamma\sigma^2/n && \mathrm{var}(f(\boldsymbol{x})) = \sigma^2, \quad \text{some } \Gamma < \infty,\\
&\mathrm{var}(\hat{\mu}_{\mathrm{RQMC}}) = O(n^{-3}\log(n)^{s-1}) && \partial^u f \in L^2[0,1]^s \quad \text{all } u \subseteq 1{:}s, \quad \text{and}\\
&\mathrm{var}(\hat{\mu}_{\mathrm{RQMC}}) = O(n^{-2}\log(n)^{2(s-1)}) && f \in \mathrm{BVHK}[0,1]^s.
\end{aligned}$$

See [22] for references. It is likely that the random linear scrambling has these moment properties too. The rate $O(n^{-3}\log(n)^{s-1})$ is established under somewhat weaker conditions than stated above by Yue and Mao [29]. There is also a central limit theorem for nested uniform sampling when $t = 0$ due to Loh [10].

If f is singular then $f \notin \mathrm{BVHK}$, so most QMC theory does not apply to it. Many singular integrands of interest are in L^2, and so RQMC theory applies to them. Similarly, step discontinuities and discontinuites in the derivative of f typically lead to $f \notin \mathrm{BVHK}$ [21] while not ruling out $f \in L^2$.

The constant Γ above is the maximum gain coefficient of the digital net. It is the key quantity that we study here.

2.3 Gain coefficients

The gain coeffiicents we study are defined with respect to a different parameterization of elementary intervals. For $u \subseteq 1{:}s$, $\boldsymbol{k} \in \mathbb{N}_0^{|u|}$ and $\boldsymbol{c} \in \mathbb{N}_0^{|u|}$ with $c_j < b^{k_j}$ let

$$E(u,\boldsymbol{k},\boldsymbol{c}) = \prod_{j\in u}\Bigl[\frac{c_j}{b^{k_j}}, \frac{c_j+1}{b^{k_j}}\Bigr) \prod_{j\notin u}[0,1). \tag{5}$$

In this representation $\mathrm{vol}(E(u,\boldsymbol{k},\boldsymbol{x})) = b^{-|u|-|\boldsymbol{k}|}$.

Using a base b Haar wavelet decomposition of $L^2[0,1]^s$ in [19] we can write $f \in L^2[0,1]^s$ as

$$f(\boldsymbol{x}) = \sum_{u\subseteq 1:s}\sum_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \nu_{u,\boldsymbol{k}}(\boldsymbol{x})$$

where the function $\nu_{u,\boldsymbol{k}}$ is constant within the elementary intervals (5). These functions are defined there in a way that makes them mutually orthogonal. For $u = \varnothing$ there is just one of these functions, and it is constant over $[0,1]^s$ with $\nu_{\varnothing,()}(\boldsymbol{x}) = \mu$ for all $\boldsymbol{x}$.

From the orthogonality of $\nu_{u,\boldsymbol{k}}$ we find that

$$\sigma^2 \equiv \mathrm{var}(f(\boldsymbol{x})) = \sum_{u\neq\varnothing}\sum_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \sigma^2_{u,\boldsymbol{k}}$$

where for $u \neq \varnothing$ we let $\sigma^2_{u,\boldsymbol{k}} = \mathrm{var}(\nu_{u,\boldsymbol{k}}(\boldsymbol{x})) = \int_{[0,1]^s} \nu_{u,\boldsymbol{k}}(\boldsymbol{x})^2$. Therefore with plain MC,

$$\mathrm{var}(\hat{\mu}_{\mathrm{MC}}) = \frac{1}{n}\sum_{u\neq\varnothing}\sum_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \sigma^2_{u,\boldsymbol{k}}.$$

If instead of plain MC we use scrambled nets, then from [19] the sample averages of $\nu_{u,\boldsymbol{k}}$ are still uncorrelated and

$$\mathrm{var}(\hat{\mu}_{\mathrm{RQMC}}) = \frac{1}{n}\sum_{u\neq\varnothing}\sum_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \Gamma_{u,\boldsymbol{k}}\sigma^2_{u,\boldsymbol{k}}$$

for gain coefficients $\Gamma_{u,\boldsymbol{k}}$ defined at (6) below. The maximal gain coefficient is

$$\Gamma = \max_{u\neq\varnothing}\max_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \Gamma_{u,\boldsymbol{k}},$$

and then $\mathrm{var}(\hat{\mu}_{\mathrm{RQMC}}) \leqslant \Gamma\sigma^2/n = \Gamma\mathrm{var}(\hat{\mu}_{\mathrm{MC}})$.

For a scrambled (t,m,s)-net in base b, if $m - t \geqslant |u| + |\boldsymbol{k}|$, then all of $E(u,\boldsymbol{k},\boldsymbol{c})$ contain the same number of points of the net. As a result $\nu_{u,\boldsymbol{k}}$ is integrated without error and $\Gamma_{u,\boldsymbol{k}} = 0$.

The general formula for gain coefficients when scrambling points $\boldsymbol{x}_0,\dots,\boldsymbol{x}_{n-1}$ is

$$\Gamma_{u,\boldsymbol{k}} = \frac{1}{n(b-1)^{|u|}}\sum_{i=0}^{n-1}\sum_{i'=0}^{n-1}\prod_{j\in u}\bigl(b\mathbf{1}_{\lfloor b^{k_j+1}x_{ij}\rfloor=\lfloor b^{k_j+1}x_{i'j}\rfloor} - \mathbf{1}_{\lfloor b^{k_j}x_{ij}\rfloor=\lfloor b^{k_j}x_{i'j}\rfloor}\bigr). \tag{6}$$

Here $\mathbf{1}_{\lfloor b^k x_{ij}\rfloor=\lfloor b^k x_{i'j}\rfloor}$ means that $x_{ij}, x_{i'j} \in [0,1)$ match in their first k base b digits. The bounds from [20] are based on equation (6). Equation (6) holds for whatever points we might choose to scramble, not just digital nets. However, the way digital nets are constructed tends to give them small values of $\Gamma_{u,\boldsymbol{k}}$.

When $b = 2$, the factors being multiplied in (6) can only take three distinct values, 0, -1, or 1, according to whether x_{ij} and $x_{i'j}$ match to fewer than k bits, exactly k bits, or more than k bits. Also the factor $(b-1)^{-|u|}$ reduces to 1. From this we get the simple bound

$$\Gamma_{u,\boldsymbol{k}} \leqslant \frac{1}{n}\sum_{i=0}^{n-1}\sum_{i'=0}^{n-1}\prod_{j\in u}\mathbf{1}_{\lfloor b^{k_j}x_{ij}\rfloor=\lfloor b^{k_j}x_{i'j}\rfloor} \tag{7}$$

when scrambling in base 2. We will see below that the bound in (7) is a power of two. More surprisingly the exact gain in (6) is either 0 or a power of two.

2.4 Prior gain bounds

From Lemma 3 in [20] we get

$$\Gamma_{u,\boldsymbol{k}} \leqslant b^t \frac{b^{|u|} + (b-2)^{|u|}}{2(b-1)^{|u|}}, \quad \text{when } m - t \leqslant |\boldsymbol{k}|, \tag{8}$$

for a slight generalization of (t, m, s)-nets in base b. The statement of that Lemma has $m - t < |\boldsymbol{k}|$ but the proof technique also applies when $m - t = |\boldsymbol{k}|$. In the case $b = 2$, the bound simplifies to $2^{t+|u|-1}$. In Section 3 we extend this bound to all $\Gamma_{u,\boldsymbol{k}}$. Lemma 4 of [20] gives

$$\Gamma_{u,\boldsymbol{k}} \leqslant b^t \Big(\frac{b+1}{b-1}\Big)^{|u|} \quad \text{when } |\boldsymbol{k}| < m - t < |u| + |\boldsymbol{k}|.$$

It is that Lemma that yields the bound $\Gamma \leqslant 2^t 3^s$ for nets in base 2.

When $t = 0$, [19] shows that $\Gamma_{u,\boldsymbol{k}} \leqslant (b/(b-1))^{s-1}$. Because such nets are only possible when $b \geqslant s$ we get $\Gamma_{u,\boldsymbol{k}} \leqslant (b/(b-1))^{b-1} \leqslant \exp(1)$. Despite this very low upper bound on worst case $\mathrm{var}(\hat{\mu}_{\mathrm{RQMC}})/\mathrm{var}(\hat{\mu}_{\mathrm{MC}})$, nets in base 2 are most used in practice.

Niederreiter and Pirsic [15] improved on the bounds of [20] by looking at the microstructure of digital nets. Microstructure refers to the placement of points within elementary intervals of volume smaller than b^{m-t}. For example the fact that Sobol' points have $t_{\{j\}} = 0$ is an aspect of their microstructure.

For $\boldsymbol{k} \in \mathbb{N}_0^s$ they introduce

$$A(k_1, \dots, k_s) = \left\lceil \max_{\boldsymbol{c} \in \mathbb{Z}_{b^{\boldsymbol{k}}}} \log_b\Big(\sum_{i=0}^{n-1} \mathbf{1}\{\boldsymbol{x}_i \in E(\boldsymbol{k}, \boldsymbol{c})\} \Big) \right\rceil$$

where the condition on $\boldsymbol{c}$ is interpreted componentwise. They also use

$$A_K = \max_{|\boldsymbol{k}|=K} A(k_1, \dots, k_s).$$

These quantities are well defined whether $\boldsymbol{x}_i$ are a (t, m, s)-net in base b or not, but they simplify for nets. The reference [15] provides several interesting upper and lower bounds on $A(\cdot)$ based on the t parameter of a net, or based on having all $\boldsymbol{x}_i \in \mathbb{Z}_{b^m}/b^m$ or knowing that one or more of the one dimensional projections of the net has $t_{\{j\}} = 0$.

From Proposition 5.1 of [15]

$$\Gamma_{u,\boldsymbol{k}} \leqslant b^{A_{|\boldsymbol{k}|}} \frac{b^{|u|} + (b-2)^{|u|}}{2(b-1)^{|u|}}.$$

This improves upon (8) by reducing the lead exponent of b and by applying to all gain coefficients. We are most interested in $b = 2$ for which their bound yields

$$\Gamma_{u,\boldsymbol{k}} \leqslant 2^{A_{|\boldsymbol{k}|}+|u|-1}.$$

Their Theorem 4.1 shows that for (t, m, s)-nets where $t_{\{1\}} = \cdots = t_{\{s\}} = 0$ that $A_{|\boldsymbol{k}|} \leqslant t$ when $|\boldsymbol{k}| > m - t$. Because $\Gamma_{1:s,\boldsymbol{k}} = 0$ whenever $|\boldsymbol{k}| \leqslant m - t$ we then get $\Gamma_{u,\boldsymbol{k}} \leqslant 2^{t+|u|-1}$ and hence $\Gamma \leqslant 2^{t+s-1}$ for Sobol' nets.

2.5 Constructions of nets

Here we describe the algorithms to construct digital nets in base 2, following [8]. We will describe how to compute 2^m points $\boldsymbol{x}_i \in [0,1)^s$ to m bits each. That is enough to get points that are a (t, m, s)-net in base 2. If one is planning to extend the points from $n = 2^m$ to some larger sample size $n = 2^M$, then it is best to use $m = M$. The points we generate actually belong to $\{0, 1/n, 2/n, \ldots, (n-1)/n\}^s$. After scrambling, one ordinarily adds random offsets $\boldsymbol{u}_i \overset{\text{iid}}{\sim} \mathbb{U}[0, 1/n)^s$ to the $\boldsymbol{x}_i$.

A digital net is defined in terms of s matrices $C_j \in \{0,1\}^{m\times m}$ for $j = 1, \ldots, s$. For Sobol' sequences

$$C_j = \begin{pmatrix} 1 & v_{2,j,1} & v_{3,j,1} & \cdots & v_{m,j,1} \\ 0 & 1 & v_{3,j,2} & \cdots & v_{m,j,2} \\ 0 & 0 & 1 & \cdots & v_{m,j,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

defined in terms of direction numbers $v_{k,j}$ that equal $0.v_{k,j,1}v_{k,j,2}v_{k,j,3}\ldots$ in their base 2 representation. Note especially that the matrix C_j is upper triangular and has 1s on its diagonal. Sobol' points ordinarily have $C_1 = I_m$.

The digital net construction works as follows. For $0 \leqslant i < 2^m$ write $i = \sum_{\ell=1}^m i_\ell 2^{\ell-1}$ for bits $i_\ell \in \{0,1\}$. Similarly, write $x_{ij} = \sum_{\ell=1}^m x_{ij\ell}2^{-\ell}$ for bits $x_{ij\ell} \in \{0,1\}$. Then the net $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{2^m-1}$ is defined by

$$\begin{pmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijm} \end{pmatrix} = C_j \begin{pmatrix} i_1 \\ i_2 \\ \vdots \\ i_m \end{pmatrix} \quad \mathsf{mod}\ 2.$$

To define t we describe a process of forming new matrices by combining some of the rows of $C_1, \ldots, C_s$. Let $C_j^{(k)} \in \{0,1\}^{k\times m}$ be the first k rows of C_j. Then for a non-empty $u = (r_1, r_2, \ldots, r_{|u|}) \subseteq 1{:}s$ and a vector $\boldsymbol{k} = (k_{r_1}, k_{r_2}, \ldots, k_{r_{|u|}}) \in \{0, 1, \ldots, m\}^s$, let

$$C_{u,\boldsymbol{k}} = \begin{pmatrix} C_{r_1}^{(k_1)} \\ C_{r_2}^{(k_2)} \\ \vdots \\ C_{r_{|u|}}^{(k_{|u|})} \end{pmatrix} \in \{0,1\}^{|\boldsymbol{k}|\times m}.$$

The t value of a digital net in base 2, constructed from $C_1, \dots, C_m$ is the smallest value of t such that $C_{u,\boldsymbol{k}}$ has linearly independent rows over $\mathbb{Z}_2$ whenever $|\boldsymbol{k}| \leqslant m - t$. This value is the smallest t for which the definition in terms of $E(\boldsymbol{k}, \boldsymbol{c})$ holds. The description above applies to any binary matrices $C_1, \dots, C_m \in \{0,1\}^{m\times m}$, not just upper triangular ones.

3 Bound on Γ

In this section we prove that $\Gamma_{u,\boldsymbol{k}} \leqslant 2^{t+|u|-1}$. It follows that $\Gamma \leqslant 2^{t+s-1}$. We make extensive use of the following elementary fact.

Proposition 1. *Let $A \in \{0,1\}^{K\times m}$ have rank r over $\mathbb{Z}_2$ and let $y \in \{0,1\}^K$. Then the set of solutions $x \in \{0,1\}^m$ to $Ax = y \bmod 2$ has cardinality 0 or 2^{m-r}.*

We need to keep track of the number of bits where $x, x' \in [0,1)$ match. For this we define

$$M(x, x') = \max\bigl\{k \in \mathbb{N}_0 \mid \lfloor 2^k x\rfloor = \lfloor 2^k x'\rfloor\bigr\} \in \mathbb{N}_0 \cup \{\infty\}.$$

Now for points that are scrambled in base 2 we get

$$\Gamma_{u,\boldsymbol{k}} = \frac{1}{n}\sum_{i=0}^{n-1}\sum_{i'=0}^{n-1}\prod_{j\in u} N_{i,i',j} \tag{9}$$

for

$$N_{i,i',j} = \begin{cases} 0, & M(x_{ij}, x_{i'j}) < k_j \\ -1, & M(x_{ij}, x_{i'j}) = k_j \\ 1, & M(x_{ij}, x_{i'j}) > k_j. \end{cases}$$

We use arrows to denote bit vectors derived from values in $[0,1)$ or in $\mathbb{N}_0$. For an integer $i = \sum_{\ell=1}^m i_\ell 2^{\ell-1}$ with $i_\ell \in \{0,1\}$ we write $\vec{i} = (i_1, i_2, \dots, i_m)^{\mathsf{T}}$ and for $x = \sum_{\ell=1}^m x_\ell 2^{-\ell}$ we write $\vec{x} = (x_1, x_2, \dots, x_m)^{\mathsf{T}}$. In either usage, $\vec{0} = (0, 0, \dots, 0)^{\mathsf{T}}$ and there are no nonzero values in $[0,1)\cap\mathbb{N}_0$, so the mapping to $\{0,1\}^m$ is well defined. We only need to represent the bits of 2^m integers in $\mathbb{Z}_{2^m}$ and 2^m of the points in $[0,1)$. Some points in $x \in [0,1)$ have two binary representations, such as $1/4 = 0.010000\cdots = 0.001111\cdots$. We use the choice that ends in a tail of 0s, via $x_\ell = \lfloor 2^\ell x\rfloor \bmod 2$.

We will also need to represent some sets of integers as bit vectors. Given a set $u \subseteq 1{:}s$ and $v \subseteq u$, we let $\vec{v} = \vec{v}[u] \in \{0,1\}^{|u|}$ have bits 1 for indices corresponding to elements of v and 0 for indices corresponding to elements of $u \setminus v$.

Arithmetic on bit vectors is done componentwise modulo 2. We write $\vec{i}\oplus\vec{j}$ and $\vec{i}\ominus\vec{j}$ for the componentwise sum and difference of bit vectors.

For non-empty $u = \{r_1, \dots, r_{|u|}\} \subseteq 1{:}s$ and $\boldsymbol{k} \in \mathbb{N}_0^{|u|}$ we define

$$C_{u,\boldsymbol{k}+\mathbf{1}} = C_{u,\boldsymbol{k}'} \quad \text{where } k'_{r_j} = k_{r_j} + 1 \text{ for } j = 1, \dots, |u|.$$

Thus $C_{u,\boldsymbol{k}+\mathbf{1}}$ has $|u|$ additional rows in it beyond those in $C_{u,\boldsymbol{k}}$. We write the matrix with just these $|u|$ additional rows as

$$\nabla C_{u,\boldsymbol{k}} = \begin{pmatrix} C_{r_1}(k_{r_1}+1,:) \\ C_{r_2}(k_{r_2}+1,:) \\ \vdots \\ C_{r_{|u|}}(k_{r_{|u|}}+1,:) \end{pmatrix}.$$

With the above setup, we are ready to establish our bounds. Within the proof of the next theorem we show that

$$\sum_{i'=0}^{2^m-1} \prod_{j\in u} N_{i,i',j} = \sum_{i'=0}^{2^m-1} \prod_{j\in u} N_{0,i',j}$$

by symmetry and then bound that sum using Proposition 1.

Theorem 1. *For integers $m \geqslant 1$ and $s \geqslant 1$, let $C_1, \dots, C_s \in \{0,1\}^{m\times m}$ generate the digital net $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{2^m-1}$ via $\vec{x}_{ij} = C_j\vec{i}$ for $0 \leqslant i < 2^m$ and $j = 1, \dots, s$. Then for nonempty $u \subseteq 1{:}s$ and $\boldsymbol{k} \in \mathbb{N}_0^{|u|}$ the gain coefficient $\Gamma_{u,\boldsymbol{k}}$ from (9) satisfies*

$$\begin{aligned}\Gamma_{u,\boldsymbol{k}} &= \sum_{i\in\mathbb{Z}_{2^m}} \mathbf{1}_{C_{u,\boldsymbol{k}}\vec{i}=0} \prod_{j\in u} N_{0,i,j} \\ &= \sum_{v\subseteq u} \#\big\{i \in \mathbb{Z}_{2^m} \mid C_{u,\boldsymbol{k}}\vec{i} = 0,\ \nabla C_{u,\boldsymbol{k}}\vec{i} = \vec{v}[u]\big\}(-1)^{|v|}. \end{aligned} \tag{10}$$

Proof. For any $i \in \mathbb{Z}_{2^m}$, there is some $\boldsymbol{c}$ with $c_j \in \mathbb{Z}_{2^{k_j}}$ for which $\boldsymbol{x}_i \in E(u, \boldsymbol{k}, \boldsymbol{c})$. Then for $i' \in \mathbb{Z}_{2^m}$ with $\boldsymbol{x}_{i'} \notin E(u, \boldsymbol{k}, \boldsymbol{c})$ we have $N_{i,i',j} = 0$ for some $j \in u$. As a result, $\prod_{j\in u} N_{i,i',j} = 0$ unless $\boldsymbol{x}_{i'} \in E(u, \boldsymbol{k}, \boldsymbol{c})$ too. Having both points in the same $E(u, \boldsymbol{k}, \boldsymbol{c})$ happens if and only if $C_{u,\boldsymbol{k}}\vec{i} = C_{u,\boldsymbol{k}}\vec{i'}$, and so only i' with $C_{u,\boldsymbol{k}}(\vec{i'} \ominus \vec{i}) = 0$ have $\prod_{j\in u} N_{i,i',j} \neq 0$.

Now for $\boldsymbol{x}_{i'} \in E(u, \boldsymbol{k}, \boldsymbol{c})$ it remains to find the sign of $\prod_{j\in u} N_{i,i',j}$. In that case

$$N_{i,i',j} = \begin{cases} -1, & M(x_{i'j}, x_{ij}) = k_j \\ 1, & M(x_{i'j}, x_{ij}) > k_j. \end{cases}$$

Suppose that $N_{i,i',j} = -1$ for $j \in v \subseteq u$ and $N_{i,i',j} = 1$ for $j \in u \setminus v$. This happens when and only when $\nabla C_{u,\boldsymbol{k}}(\vec{i'} \ominus \vec{i}) = \vec{v}[u]$. Then $\prod_{j\in u} N_{i,i',j} = (-1)^{|v|}$. It follows that

$$\begin{aligned}\Gamma_{u,\boldsymbol{k}} &= \frac{1}{n}\sum_{i=0}^{2^m-1} \sum_{v\subseteq u} \#\big\{i' \in \mathbb{Z}_{2^m} \mid C_{u,\boldsymbol{k}}(\vec{i'} \ominus \vec{i}) = 0, \nabla C_{u,\boldsymbol{k}}(\vec{i'} \ominus \vec{i}) = \vec{v}[u]\big\}(-1)^{|v|} \\ &= \sum_{v\subseteq u} \#\big\{i' \in \mathbb{Z}_{2^m} \mid C_{u,\boldsymbol{k}}\vec{i'} = 0, \nabla C_{u,\boldsymbol{k}}\vec{i'} = \vec{v}[u]\big\}(-1)^{|v|} \end{aligned}$$

where the second step follows because $\vec{i}' \ominus \vec{i}$ runs over the set $\{0,1\}^m$ for any $i \in \mathbb{Z}_{2^m}$. □

The next corollary is already known from the definition of (t,m,s)-nets. We include it to show how it follows from Theorem 1 and because we need it below.

Corollary 1. *If $C_{u,\boldsymbol{k}+\mathbf{1}}$ has full row rank $|u|+|\boldsymbol{k}|$ for non-empty $u \subseteq 1{:}s$, then $\Gamma_{u,\boldsymbol{k}} = 0$.*

Proof. When $C_{u,\boldsymbol{k}+\mathbf{1}}$ has full row rank then $C_{u,\boldsymbol{k}}\vec{i} = 0$ and $\nabla C_{u,\boldsymbol{k}}\vec{i} = \vec{v}[u]$ has $2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}})}$ solutions for all $v \subseteq u$. Now Theorem 1 yields $\Gamma_{u,\boldsymbol{k}} = 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}})}\sum_{v\subseteq u}(-1)^{|v|} = 0$. □

Corollary 2. $\Gamma_{u,\boldsymbol{k}} \leqslant 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}$.

Proof. We can rewrite equation (10) as

$$\begin{aligned}
\Gamma_{u,\boldsymbol{k}} &= \sum_{v\subseteq u} \#\bigl\{\vec{i} \in \{0,1\}^m \mid C_{u,\boldsymbol{k}}\vec{i} = 0, \nabla C_{u,\boldsymbol{k}}\vec{i} = \vec{v}[u]\bigr\}(-1)^{|v|}\\
&\leqslant \sum_{v\subseteq u} \#\bigl\{\vec{i} \in \{0,1\}^m \mid C_{u,\boldsymbol{k}}\vec{i} = 0, \nabla C_{u,\boldsymbol{k}}\vec{i} = \vec{v}[u]\bigr\}\\
&= \#\bigl\{\vec{i} \in \{0,1\}^m \mid C_{u,\boldsymbol{k}}\vec{i} = 0\bigr\}\\
&= 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}.
\end{aligned}$$

The last step follows from Proposition 1 after noting that there is at least one solution because $\vec{0}$ is a solution. □

Corollary 3. *Let $\boldsymbol{x}_i \in [0,1)^s$ for $i \in \mathbb{Z}_{2^m}$ be a (t,m,s)-net in base 2. Then*

$$\Gamma_{u,\boldsymbol{k}} \leqslant 2^{t+|u|-1}.$$

Proof. If $C_{u,\boldsymbol{k}+\mathbf{1}}$ has full row rank then $\Gamma_{u,\boldsymbol{k}} = 0$ by Corollary 1.

Suppose next that $C_{u,\boldsymbol{k}}$ has full row rank but $C_{u,\boldsymbol{k}+\mathbf{1}}$ does not. The matrix $C_{u,\boldsymbol{k}+\mathbf{1}}$ has $|u|+|\boldsymbol{k}|$ rows and this must be at least $m-t+1$ by the definition of t for a (t,m,s)-net. Because $C_{u,\boldsymbol{k}}$ has full row rank we get $\mathrm{rank}(C_{u,\boldsymbol{k}}) = |\boldsymbol{k}|$ and then by Corollary 2,

$$\Gamma_{u,\boldsymbol{k}} \leqslant 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})} = 2^{m-|\boldsymbol{k}|} \leqslant 2^{t+|u|-1}.$$

The remaining case is that $C_{u,\boldsymbol{k}}$ does not have full row rank. In that case $|\boldsymbol{k}| \geqslant m-t+1$. The matrix $C_{u,\boldsymbol{k}}$ must have a subset of $m-t$ rows defined by $C_{u,\boldsymbol{k}'}$ with $\boldsymbol{k}' \leqslant \boldsymbol{k}$ componentwise for which $C_{u,\boldsymbol{k}'}$ has full row rank. Then by Corollary 2,

$$\Gamma_{u,\boldsymbol{k}} \leqslant 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})} \leqslant 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}'})} = 2^t \leqslant 2^{t+|u|-1}. \quad \square$$

Sobol' matrices are upper triangular with ones on their diagonal. Corollary 3 does not require upper triangular matrices or ones on the diagonal. Those properties are important but their benefit comes through t.

The largest bounds on gain coefficients come from the case where $C_{u,\boldsymbol{k}}$ has full rank but $C_{u,\boldsymbol{k}+\mathbf{1}}$ does not. Among these, the largest are the ones for large $|u|$.

Remark 1. The strategy above can be extended to (t, m, s)-nets in base p for prime numbers p to show that $\max_{u,\boldsymbol{k}} \Gamma_{u,\boldsymbol{k}} \leqslant p^{t+|u|-1}$. That will not generally improve on the bound $p^t((p+1)/(p-1))^{|u|}$ from [20]. Even for $p = 3$ it brings improvements only for $|u| \leqslant 2$ and raises the bound for $|u| \geqslant 3$.

4 Γ is a power of 2

Here we prove that the upper bound is actually tight, so that $\Gamma_{u,\boldsymbol{k}}$ is either 0 or $2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}$, making the maximal gain a power of 2 (because it is impossible to have every $\Gamma_{u,\boldsymbol{k}} = 0$). We need some further notation. For $w \subseteq u \subseteq 1{:}s$ and $\boldsymbol{k} \in \mathbb{N}_0^{|u|}$ let $\boldsymbol{k} + \mathbf{1}_w$ be the vector $\boldsymbol{k}' \in \mathbb{N}_0^{|u|}$ with $k'_j = k_j + 1$ for $j \in w$ and $k'_j = k_j$ for $j \in u \setminus w$. We then introduce a generalized gain coefficient

$$\Gamma^w_{u,\boldsymbol{k}} = \sum_{i \in \mathbb{Z}_{2^m}} \mathbf{1}\{C_{u,\boldsymbol{k}}\vec{i} = 0\} \prod_{j \in w} N_{0,i,j}. \tag{11}$$

As usual, an empty product is 1. Also the matrix with all the rows of $C_{u,\boldsymbol{k}+\mathbf{1}_w}$ that are not in $C_{u,\boldsymbol{k}}$ is denoted $\nabla^w C_{u,\boldsymbol{k}} \in \{0,1\}^{|w|\times m}$.

Lemma 1. *If $w \neq \varnothing$ and $\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}_w}) - \mathrm{rank}(C_{u,\boldsymbol{k}}) = |w|$ then $\Gamma^w_{u,\boldsymbol{k}} = 0$.*

Proof. Because $\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}_w}) - \mathrm{rank}(C_{u,\boldsymbol{k}}) = |w|$, the image of $\nabla^w C_{u,\boldsymbol{k}}\vec{i}$ for $\vec{i} \in \{0,1\}^m$ with $C_{u,\boldsymbol{k}}\vec{i} = 0$ has rank $|w|$. But $\nabla^w C_{u,\boldsymbol{k}}\vec{i} \in \{0,1\}^{|w|}$, so the image is the whole space. Therefore for any $v \subseteq w$ the system of equations

$$C_{u,\boldsymbol{k}}\vec{i} = 0 \quad \text{and} \quad \nabla^w C(u,\boldsymbol{k})\vec{i} = \vec{v}[w]$$

is consistent and has $2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}_w})}$ solutions. The rest of the proof is like that in Corollary 1. □

Lemma 2. *If $w \neq \varnothing$ and $\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}_w}) - \mathrm{rank}(C_{u,\boldsymbol{k}}) < |w|$ then there exists a nonempty $v \subseteq w$ such that $\prod_{j\in v} N_{0,i,j} = 1$ for any $i \in \mathbb{Z}_{2^m}$ with $C_{u,\boldsymbol{k}}\vec{i} = 0$.*

Proof. By hypothesis, there exist coefficients $a_{j,\ell}$ for $1 \leqslant \ell \leqslant k_j$ for $j \in u \setminus w$ and $1 \leqslant \ell \leqslant k_j + 1$ for $j \in w$ with at least one $a_{j,k_j+1} = 1$ for $j \in w$ such that

$$\sum_{j\in u\setminus w} \sum_{\ell=1}^{k_j} a_{j,\ell} C_j(\ell,:) \oplus \sum_{j\in w} \sum_{\ell=1}^{k_j+1} a_{j,\ell} C_j(\ell,:) = 0 \tag{12}$$

with $C_j(\ell,:) \in \{0,1\}^m$ equal to row ℓ of C_j.

We will show that $v = \{j \in w \mid a_{j,k_j+1} = 1\}$ satisfies the conditions of the Lemma. To do this we choose any $i \in \mathbb{Z}_{2^m}$ with $C_{u,\boldsymbol{k}}\vec{i} = 0$ and let $\vec{e} = \nabla^w C_{u,\boldsymbol{k}}\vec{i} \in \{0,1\}^{|w|}$. Multiplying both sides of (12) by $\vec{i}$ gives

$$\sum_{j\in w} a_{j,k_j+1} C_j(k_j+1,:)\vec{i} = \sum_{j\in v} e_j = 0 \bmod 2.$$

Because the bits of $\vec{e}$ sum to zero in $\mathbb{Z}_2$ there must be an even number of them that equal 1. There are then an even number of $j \in v$ with $N_{0,i,j} = -1$ and then $\prod_{j\in v} N_{0,i,j} = 1$. □

Theorem 2. *$\Gamma_{u,\boldsymbol{k}}$ is either 0 or $2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}$.*

Proof. We proceed by induction on $|w|$ to prove that $\Gamma^w_{u,\boldsymbol{k}} \in \{0, 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}\}$ for all $w \subseteq u$. The conclusion then follows because $\Gamma^u_{u,\boldsymbol{k}} = \Gamma_{u,\boldsymbol{k}}$.

We begin with $w = \varnothing$. Then from the definition (11) of generalized gain coefficients,

$$\Gamma^{\varnothing}_{u,\boldsymbol{k}} = \sum_{i\in\mathbb{Z}_{2^m}} \mathbf{1}_{C_{u,\boldsymbol{k}}\vec{i}=0} \prod_{j\in\varnothing} N_{0,i,j} = \sum_{i\in\mathbb{Z}_{2^m}} \mathbf{1}_{C_{u,\boldsymbol{k}}\vec{i}=0} \in \{0, 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}\}$$

by Proposition 1.

Now we suppose that $\Gamma^w_{u,\boldsymbol{k}} \in \{0, 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}\}$ holds whenever $0 \leqslant |w| < r$ for some $r \leqslant |u|$. If $|w| = r$ and $\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}_w}) - \mathrm{rank}(C_{u,\boldsymbol{k}}) = |w|$ then $\Gamma^w_{u,\boldsymbol{k}} = 0$ for $w \neq \varnothing$ by Lemma 1.

It remains to consider the case with $|w| = r$ and $\mathrm{rank}(C_{u,\boldsymbol{k}+\mathbf{1}_w}) - \mathrm{rank}(C_{u,\boldsymbol{k}}) < |w|$. In this case w is not empty and we let v be the non-empty subset of w from Lemma 2. Then because $\prod_{j\in v} N_{0,i,j} = 1$

$$\Gamma^w_{u,\boldsymbol{k}} = \sum_{i\in\mathbb{Z}_{2^m}} \mathbf{1}_{C_{u,\boldsymbol{k}}\vec{i}=0} \prod_{j\in w} N_{0,i,j} = \sum_{i\in\mathbb{Z}_{2^m}} \mathbf{1}_{C_{u,\boldsymbol{k}}\vec{i}=0} \prod_{j\in w\setminus v} N_{0,i,j} = \Gamma^{w\setminus v}_{u,\boldsymbol{k}}.$$

Now $|w \setminus v| < |w| = r$ so we can apply the induction hypothesis. □

Since there are only two possibilities for $\Gamma_{u,\boldsymbol{k}}$ we are able to get a computationally advantageous check for which of them holds.

Corollary 4. *$\Gamma_{u,\boldsymbol{k}} = 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}$ if and only if $\sum_{j\in u} C_j(k_j+1,:) \in \{0,1\}^m$ is in the row space of $C_{u,\boldsymbol{k}}$.*

Proof. First suppose that $\sum_{j\in u} C_j(k_j+1,:)$ is in the row space of $C_{u,\boldsymbol{k}}$. Then we can apply Lemma 2 with $v = w = u$ to get that $\prod_{j\in u} N_{0,i,j} = 1$ for any $i \in \mathbb{Z}_{2^m}$ with $C_{u,\boldsymbol{k}}\vec{i} = 0$. Conversely, suppose that $\Gamma_{u,\boldsymbol{k}} = 2^{m-\mathrm{rank}(C_{u,\boldsymbol{k}})}$. Then from details of the proof of Corollary 2,

$$\begin{aligned}
&\sum_{v\subseteq u} \#\big\{\vec{i} \in \{0,1\}^m \mid C_{u,\boldsymbol{k}}\vec{i} = 0, \nabla C_{u,\boldsymbol{k}}\vec{i} = \vec{v}[u]\big\}(-1)^{|v|} \\
&= \sum_{v\subseteq u} \#\big\{\vec{i} \in \{0,1\}^m \mid C_{u,\boldsymbol{k}}\vec{i} = 0, \nabla C_{u,\boldsymbol{k}}\vec{i} = \vec{v}[u]\big\},
\end{aligned}$$

which rules out having any solutions $\vec{i} \in \{0,1\}^m$ to

$$C_{u,\boldsymbol{k}}\vec{i} = 0, \nabla C_{u,\boldsymbol{k}}\vec{i} = \vec{v}[u]$$

for any v with an odd cardinality. Therefore $\sum_{j\in u} C_j(k_j+1,:)\vec{i} = 0$ whenever $C_{u,\boldsymbol{k}}\vec{i} = 0$ which then implies that $\sum_{j\in u} C_j(k_j+1,:)$ is in the row space of $C_{u,\boldsymbol{k}}$. □

5 Reduced upper bound

From Corollary 2 we have $\Gamma_{u,\boldsymbol{k}} \leqslant 2^{t+|u|-1}$. It follows immediately that $\Gamma_{u,\boldsymbol{k}} \leqslant 2^{t_u+|u|-1}$ because we could have formed a net out of only $\boldsymbol{x}_{i,u} \subset [0,1)^{|u|}$. In this section we show that it is possible to improve that bound.

We need to use a second notion of the t parameter specific to a subset $u \neq \varnothing$ of components of $\boldsymbol{x}_i$. This notion is denoted t^*_u. We define it below side by side with the prior t_u to make comparisons easier. We also need a quantity t_d for $1 \leqslant d \leqslant s$ to describe the quality of projected nets. The new and old quantities are

$$\begin{aligned}
t &= m+1-\min_{u\neq\varnothing,\boldsymbol{k}\in\mathbb{N}_0^{|u|}}\bigl\{|\boldsymbol{k}| \mid C_{u,\boldsymbol{k}} \text{ not of full rank}\bigr\},\\
t_d &= m+1-\min_{u:|u|\leqslant d,\boldsymbol{k}\in\mathbb{N}_0^{|u|}}\bigl\{|\boldsymbol{k}| \mid C_{u,\boldsymbol{k}} \text{ not of full rank}\bigr\},\\
t_u &= m+1-\min_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}}\bigl\{|\boldsymbol{k}| \mid C_{u,\boldsymbol{k}} \text{ not of full rank}\bigr\}, \quad \text{and}\\
t^*_u &= m+1-\min_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}}\bigl\{|\boldsymbol{k}| \mid C_{u,\boldsymbol{k}} \text{ not of full rank, } \boldsymbol{k} \geqslant \mathbf{1}_{\mathrm{u}} \text{ componentwise}\bigr\}.
\end{aligned}$$

Because t^*_u adds constraints $k_j \geqslant 1$ for $j \in u$, we have $t^*_u \leqslant t_u$. Likewise, if $v \subseteq u$, then t^*_v adds constraints $k_j \geqslant 1$ for $j \in v$ and $k_j = 0$ for $j \in u \setminus v$, and we have $t^*_v \leqslant t_u$. For any $\boldsymbol{k}$ that attains the minimum defined in t_u, define $v = \{j \in u \mid k_j \geqslant 1\}$ and $\boldsymbol{k}' \in \mathbb{N}_0^{|v|}$ to be the nonzero entries of $\boldsymbol{k}$. Then $C_{u,\boldsymbol{k}}$ is also $C_{v,\boldsymbol{k}'}$ and $t^*_v \geqslant t_u$. Therefore

$$t_u = \max_{v\subseteq u} t^*_v.$$

From the definitions of t_d and t, we have

$$\begin{aligned}
t_d &= \max_{u:|u|\leqslant d} t_u = \max_{u:|u|\leqslant d}\max_{v\subseteq u} t^*_v = \max_{|v|\leqslant d} t^*_v, \quad \text{and}\\
t &= \max_{1\leqslant d\leqslant s} t_d = \max_{u\neq\varnothing} t_u = \max_{v\neq\varnothing} t^*_v.
\end{aligned}$$

Theorem 3 below shows that we can replace the bound $2^{t_u+|u|-1}$ by $2^{t^*_u+|u|-1}$. It then follows that

$$\max_{u:|u|\leqslant d}\max_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \Gamma_{u,\boldsymbol{k}} \leqslant \max_{u:|u|\leqslant d} 2^{t^*_u+|u|-1} \leqslant 2^{t_d+d-1}.$$

For the next results we need to use the vector $\mathbf{1}_u = (1,1,\ldots,1) \in \mathbb{Z}^{|u|}$.

Theorem 3. *For any $u \subseteq 1{:}s$ where $C_{u,\mathbf{1}_u}$ has full row rank,*

$$\max_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \Gamma_{u,\boldsymbol{k}} \leqslant 2^{t_u^*+|u|-1} \quad \textit{and} \quad \max_{v\subseteq u} \max_{\boldsymbol{k}\in\mathbb{N}_0^{|v|}} \Gamma_{v,\boldsymbol{k}} = 2^{t_u^*+|u|-1}.$$

Proof. First we prove that $\Gamma_{u,\boldsymbol{k}} \leqslant 2^{t_u^*+|u|-1}$. By the definition of t_u^*, if a matrix $C_{u,\boldsymbol{k}}$ with $\boldsymbol{k} \geqslant \mathbf{1}_u$ does not have full row rank, it must satisfy $|\boldsymbol{k}| \geqslant m+1-t_u^*$. The proof in the case where $C_{u,\boldsymbol{k}}$ has full row rank is like that in Corollary 3. The remaining case is when $C_{u,\boldsymbol{k}}$ does not have full row rank.

Define $v = \{j \in u \mid k_j = 0\}$. Then $|\boldsymbol{k}| + |v| \geqslant m - t_u^* + 1$ because $C_{u,\boldsymbol{k}+\mathbf{1}_v}$ does not have full row rank and $\boldsymbol{k}+\mathbf{1}_v \geqslant \mathbf{1}_u$. The matrix $C_{u,\boldsymbol{k}+\mathbf{1}_v}$ must have a subset of $m-t_u^*$ rows defined by $C_{u,\boldsymbol{k}'}$ with $\boldsymbol{k}' \leqslant \boldsymbol{k}+\mathbf{1}_v$ for which $C_{u,\boldsymbol{k}'}$ has full row rank. Then by Corollary 2,

$$\Gamma_{u,\boldsymbol{k}} \leqslant 2^{m-\operatorname{rank}(C_{u,\boldsymbol{k}})} \leqslant 2^{m-\operatorname{rank}(C_{u,\boldsymbol{k}+\mathbf{1}_v})+|v|} \leqslant 2^{t_u^*+|v|} \leqslant 2^{t_u^*+|u|-1},$$

where the last inequality follows from v being a proper subset of u because $\boldsymbol{k}$ cannot be 0.

To prove the second statement, notice that for any $v \subseteq u$ and $\boldsymbol{k} \in \mathbb{N}_0^{|v|}$ such that $\boldsymbol{k} \geqslant \mathbf{1}_v$ and $C_{v,\boldsymbol{k}}$ is row rank deficient, we can define $\boldsymbol{k}' \in \mathbb{N}_0^{|u|}$ so that $k'_j = k_j$ for $j \in v$ and $k'_j = 1$ for $j \in u \setminus v$. Then $\boldsymbol{k}' \geqslant \mathbf{1}_u$ and $C_{u,\boldsymbol{k}'}$ is row rank deficient as well because it is made of $C_{v,\boldsymbol{k}}$ with $|u|-|v|$ extra rows. It follows that $t_v^* \leqslant t_u^* + |u| - |v|$ and

$$\max_{\boldsymbol{k}\in\mathbb{N}_0^{|v|}} \Gamma_{v,\boldsymbol{k}} \leqslant 2^{t_v^*+|v|-1} \leqslant 2^{t_u^*+|u|-1}.$$

It remains to show that there exists $v \subseteq u$ and $\boldsymbol{k}^v \in \mathbb{N}_0^{|v|}$ such that $\boldsymbol{k}^v \geqslant \mathbf{1}_v$ and $\Gamma_{v,\boldsymbol{k}^v} = 2^{t_u^*+|u|-1}$. First we choose any $\boldsymbol{k}^*$ that attains the minimum $|\boldsymbol{k}|$ defined in t_u^*. Because $C_{u,\boldsymbol{k}^*}$ is not of full rank, its row vectors must be linearly dependent. That is, there exist coefficients $a_{j,\ell} \in \{0,1\}$ for $1 \leqslant \ell \leqslant k_j^*$ and $j \in u$ such that

$$\sum_{j\in u} \sum_{\ell=1}^{k_j^*} a_{j,\ell} C_j(\ell,:) = 0 \bmod 2.$$

Define $v = \{j \in u \mid a_{j,k_j^*} = 1\}$ and $\boldsymbol{k}^v \in \mathbb{N}_0^{|v|}$ such that $k_j^v = k_j^*$ for $j \in v$. Because $\boldsymbol{k}^*$ attains the smallest $|\boldsymbol{k}|$ among $\boldsymbol{k} \geqslant \mathbf{1}_u$, $a_{j,k_j^*} = 1$ for all $j \in u$ such that $k_j^* \geqslant 2$. In other words, $j \in u \setminus v$ if and only if $k_j^* = 1$ and $a_{j,1} = 0$. Hence $|\boldsymbol{k}^v| = |\boldsymbol{k}^*| - |u \setminus v|$ and

$$\sum_{j\in v} C_j(k_j^v,:) + \sum_{j\in v} \sum_{\ell=1}^{k_j^v-1} a_{j,\ell} C_j(\ell,:) = 0.$$

Corollary 4 then implies that $\Gamma_{v,\boldsymbol{k}^v-\mathbf{1}_v} = 2^{m-\operatorname{rank}(C_{v,\boldsymbol{k}^v-\mathbf{1}_v})}$. Again, because $\boldsymbol{k}^*$ attains the smallest $|\boldsymbol{k}|$, $C_{v,\boldsymbol{k}^v-\mathbf{1}_v}$ has full row rank and $\operatorname{rank}(C_{v,\boldsymbol{k}^v-\mathbf{1}_v}) = |\boldsymbol{k}^v| - |v| = |\boldsymbol{k}^*| - |u|$. Therefore

$$\Gamma_{v,\boldsymbol{k}^v-\mathbf{1}_v} = 2^{m-\operatorname{rank}(C_{v,\boldsymbol{k}^v-\mathbf{1}_v})} = 2^{m-|\boldsymbol{k}^*|+|u|} = 2^{t_u^*+|u|-1}. \quad \square$$

Corollary 5. *If there exists* $u \subseteq 1{:}s$ *with* $\mathrm{rank}(C_{u,\mathbf{1}_u}) < |u|$ *then the maximal gain coefficient* $\Gamma = 2^m$*. Otherwise* $\Gamma = 2^{t^*_{1:s}+s-1}$*.*

Proof. In the former case, we can choose the smallest size u whose $C_{u,\mathbf{1}_u}$ is not full rank. Then the row vectors of $C_{u,\mathbf{1}_u}$ are linearly dependent, but any proper subset of them are linearly independent. Therefore $\sum_{j\in u} C_j(1,:) = 0$. We can then apply Corollary 4 to $\Gamma_{u,0}$ and derive $\Gamma_{u,0} = 2^m$, which is the largest $\Gamma_{u,\boldsymbol{k}}$ in view of Corollary 2.

In the latter case, the conclusion immediately follows by applying Theorem 3 to $u = 1{:}s$. □

Example 1. Here we consider a $(1, 1, 4)$-net in base 2, known as a shift net [24] because the columns of the first generator matrix C_1 are shifted to create the other generator matrices. The top three rows of the generator matrices C_1 through C_4 are

$$\begin{pmatrix} 0&0&0&1\\ 0&1&1&0\\ 0&0&1&0 \end{pmatrix}, \begin{pmatrix} 0&0&1&0\\ 1&1&0&0\\ 0&1&0&0 \end{pmatrix}, \begin{pmatrix} 0&1&0&0\\ 1&0&0&1\\ 1&0&0&0 \end{pmatrix} \text{ and } \begin{pmatrix} 1&0&0&0\\ 0&0&1&1\\ 0&0&0&1 \end{pmatrix}.$$

The fourth rows could be anything without changing this example. From $m = s = 4$ and $t = 1$ we get a gain coefficient bound $\Gamma \leqslant 2^{t+s-1} = 16$. However $t^*_{1:s} = 0$ for this net after observing that

$$C_{1:4,\mathbf{1}_{1:4}} = \begin{pmatrix} 0&0&0&1\\ 0&0&1&0\\ 0&1&0&0\\ 1&0&0&0 \end{pmatrix}$$

has full rank. Therefore $\Gamma = 2^{t^*_{1:s}+s-1} = 8$.

6 Discussion

Our first contribution is to tighten the bounds on $\Gamma_{u,\boldsymbol{k}}$ and hence also the maximal gain $\Gamma = \max_{|u|>0} \max_{\boldsymbol{k}\in\mathbb{N}_0^{|u|}} \Gamma_{u,\boldsymbol{k}}$ for digital nets in base 2 of which the constructions of Sobol' [28] and Niederreiter-Xing [16, 17] are the most important. Our second contribution is to show that gain coefficients for base 2 digital nets must be either 0 or a power of 2, so the maximal correlation is a power of 2. Finally, a consequence of our results is a more efficient algorithm for computing gain coefficients.

Acknowledgments

This work was supported by the U.S. National Science Foundation under grant IIS-1837931.

References

[1] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration (2nd Ed.)*. Academic Press, San Diego, 1984.

[2] Luc Devroye. *Non-uniform Random Variate Generation*. Springer, 1986.

[3] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

[4] J. Dick and F. Pillichshammer. *Digital sequences, discrepancy and quasi-Monte Carlo integration*. Cambridge University Press, Cambridge, 2010.

[5] J. Dick and F. Pillichshammer. Discrepancy theory and quasi-Monte Carlo integration. In *A panorama of discrepancy theory*, pages 539–619. Springer, 2014.

[6] H. Faure. Discrépance de suites associées à un système de numération (en dimension s). *Acta Arithmetica*, 41:337–351, 1982.

[7] F. J. Hickernell. Koksma-Hlawka inequality. *Wiley StatsRef: Statistics Reference Online*, 2014.

[8] S. Joe and F. Y. Kuo. Constructing Sobol' sequences with better two-dimensional projections. *SIAM Journal on Scientific Computing*, 30(5):2635–2654, 2008.

[9] P. L'Ecuyer and C. Lemieux. A survey of randomized quasi-Monte Carlo methods. In M. Dror, P. L'Ecuyer, and F. Szidarovszki, editors, *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pages 419–474. Kluwer Academic Publishers, 2002.

[10] W.-L. Loh. On the asymptotic distribution of scrambled net quadrature. *Annals of Statistics*, 31(4):1282–1324, 2003.

[11] P. Marion, M. Godin, and P. L'Ecuyer. An algorithm to compute the t-value of a digital net and of its projections. *Journal of Computational and Applied Mathematics*, 371:112669, 2020.

[12] J. Matoušek. *Geometric Discrepancy : An Illustrated Guide*. Springer-Verlag, Heidelberg, 1998.

[13] H. Niederreiter. Quasi-Monte Carlo methods and pseudo-random numbers. *Bulletin of the American Mathematical Society*, 84(6):957–1041, 1978.

[14] H. Niederreiter. Point sets and sequences with small discrepancy. *Monatshefte fur mathematik*, 104:273–337, 1987.

[15] H. Niederreiter and G. Pirsic. The microstructure of (t,m,s)-nets. *Journal of Complexity*, 17(4):683–696, Dec. 2001.

[16] H. Niederreiter and C. Xing. Low-discrepancy sequences and global function fields with many rational places. *Finite Fields and Their Applications*, 2:241–273, 1996.

[17] H. Niederreiter and C. Xing. Quasirandom points and global function fields. In S. Cohen and H. Niederreiter, editors, *Finite Fields and Applications*, volume 233, pages 269–296, Cambridge, 1996. Cambridge University Press.

[18] A. B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In H. Niederreiter and P. J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 299–317, New York, 1995. Springer-Verlag.

[19] A. B. Owen. Monte Carlo variance of scrambled equidistribution quadrature. *SIAM Journal on Numerical Analysis*, 34(5):1884–1910, 1997.

[20] A. B. Owen. Scrambling Sobol' and Niederreiter-Xing points. *Journal of Complexity*, 14(4):466–489, 1998.

[21] A. B. Owen. Multidimensional variation for quasi-Monte Carlo. In J. Fan and G. Li, editors, *International Conference on Statistics in honour of Professor Kai-Tai Fang's 65th birthday*, 2005.

[22] A. B. Owen and D. Rudolf. A strong law of large numbers for scrambled net integration. *SIAM Review*, 63(2):360–372, 2021.

[23] R.D. Richtmyer. The evaluation of definite integrals and quasi-Monte Carlo method based on the properties of algebraic numbers. Technical report, Los Alamos Scientific Laboratory, Los Alamos, NM, 1951.

[24] W. Ch. Schmid. Shift—nets: a new class of binary digital (t, m, s)–nets. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1996*, pages 369–381. Springer, 1998.

[25] Wolfgang Ch. Schmid. Projections of digital nets and sequences. *Mathematics and Computers in Simulation*, 55:239–247, 2001.

[26] R. Schürer and W. Ch. Schmid. MinT: a database for optimal net parameters. In H. Niederreiter and D. Talay, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2004*. Springer, 2006.

[27] R. Schürer and W. Ch. Schmid. MinT: new features and new results. In P. L'Ecuyer and A. B. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2008*. Springer, 2009.

[28] I. M. Sobol'. The use of Haar series in estimating the error in the computation of infinite-dimensional integrals. *Dokl. Akad. Nauk SSSR*, 8(4):810–813, 1967.

[29] R.-X. Yue and S.-S. Mao. On the variance of quadrature over scrambled nets and sequences. *Statistics & probability letters*, 44(3):267–280, 1999.