ARTICLE TEMPLATE

On Approximating the Shape of One-Dimensional Posterior Marginals using the Low Discrepancy Points

Chaitanya Joshi[a] and Paul T. Brown [a] and Stephen Joe [a]

[a]Dept. of Mathematics, University of Waikato, Hamilton, New Zealand 3240.

ARTICLE HISTORY
Compiled November 18, 2021

ABSTRACT
A method to approximate Bayesian posterior by evaluating it on a low discrepancy sequence (LDS) point set has recently been proposed. However, this method does not focus on finding the posterior marginals. Finding posterior marginals when the posterior approximation is obtained using LDS is not straightforward, and as yet, there is no method to approximate one dimensional marginals using an LDS. We propose an approximation method for this problem. This method is based on an s-dimensional integration rule together with fitting a polynomial smoothing function. We state and prove results showing conditions under which this polynomial smoothing function will converge to the true one-dimensional function. We also demonstrate the computational efficiency of the new approach compared to a grid based approach.

KEYWORDS
Bayesian inference; low discrepancy sequences; quasi-Monte Carlo; Integrated Nested Laplace Approximation; interpolating polynomials

1. Introduction

1.1. *Motivation*

This work is motivated by an application in Bayesian statistical inference where there is an interest in the one-dimensional posterior distributions. While, Monte Carlo based methods such as the Markov Chain Monte Carlo or the Approximate Bayesian Computation are more widely used to approximate posterior distributions, these can be computationally expensive. Methods such as the Integrated Nested Laplace Approximation (INLA) (Rue, Martino, and Chopin 2009) that instead explore the posterior distribution using a deterministic set of grid points or using other deterministic set of points, such as central composite design (CCD) points (Martins et al. 2013) — have been proposed as computationally efficient alternatives. However, since the number of grid points increases exponentially with s, grid based methods can only be used when the (hyper) parameter space has very few dimensions (Rue, Martino, and Chopin 2009). Using CCD points is more efficient however, finding one dimensional distributions is then not straightforward. Existing numerical integration free methods can only approximate uni-modal distributions (Martins et al. 2013).

CONTACT Chaitanya Joshi. Email: cjoshi@waikato.ac.nz

Therefore, there is potential to explore the use of LDS to approximate the posterior distributions instead since such approximations could be more computationally efficient as well as accurate compared to those obtained using grid points or CCD points. Recently, a new approach to implement INLA using LDS (Brown et al. 2021) was indeed proposed. However, this approach does not focus on finding the posterior marginals. Finding posterior marginals when the posterior approximation is obtained using LDS is not straightforward, and as yet, there is no method to approximate one dimensional marginals using an LDS.

In this paper we develop a method to approximate the shape of the one-dimensional functions when an s-dimensional function is evaluated using N LDS points. However, the focus of this paper is purely mathematical. It is not expected that the method developed here can be used to approximate Bayesian posterior distributions in its existing form. We expand more on this point in Section 5. In this paper we simply develop a method and prove the convergence theorems for the approximations.

1.2. *Integration Rules and Low Discrepancy Sequences*

Suppose we have an integrable function $g : [\boldsymbol{a}, \boldsymbol{b}] \to \mathbb{R}$, where $\boldsymbol{a} = (a_1, \ldots, a_s)$ and $\boldsymbol{b} = (b_1, \ldots, b_s)$ with $a_j < b_j$ for $1 \leq j \leq s$. Without loss of generality, we may take the region of interest to be the unit hypercube $[0, 1]^s$ since a linear transformation may be used to map a function g defined over $[\boldsymbol{a}, \boldsymbol{b}]$ to a function f defined over $[0, 1]^s$.

Now consider the s-dimensional integral

$$I = \int_{[0,1]^s} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

The standard approach taken to find an approximation to I is typically to make use of an integration rule. These integration rules are of the form

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i), \tag{1.1}$$

where the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are sampled from the unit hypercube $[0, 1]^s$. There are a number of choices for the integration rules. One can use Monte Carlo (MC) rules in which the points are chosen randomly. However, such points do suffer from large gaps and clusters and this can affect the accuracy of the estimate for a given set of points Lemieux (2009). If the point set was taken to be the regular n-point grid for which the point set consists of the points $((i_1 - 1)/(n - 1), (i_2 - 1)/(n - 1), \ldots, (i_s - 1)/(n - 1))$, where $1 \leq i_\ell \leq n$ for $1 \leq \ell \leq s$, then the total number of points is $N = n^s$. If s is large, then the number of points increases rapidly as n increases.

A large class of integration rules is the class of *quasi-Monte Carlo* (QMC) rules. These are equal weight integration rules of the form (1.1) that use deterministic point sets, specifically, the low discrepancy sequences (LDS). These point sets have low *discrepancy* with respect to the Lebesgue measure on a unit hypercube. One of the most commonly used discrepancy measure is called the star discrepancy. Let $\mathcal{P}_N$ be an N element point set in $[0, 1]^s$. For $\boldsymbol{a} \in (0, 1]^s$, the star discrepancy D_N^* of this point

set is defined as

$$D_N^* = \sup_{\boldsymbol{a}\in[0,1]^s} \left| \frac{\alpha([0,\boldsymbol{a}),\mathcal{P}_N,N)}{N} - \prod_{j=1}^{s} a_j \right|,$$

where, $\alpha([0,\boldsymbol{a}),\mathcal{P}_N,N) = \#\left\{n \in \mathbb{N} : 1 \le n \le N, \boldsymbol{x}_n \in [0,\boldsymbol{a})\right\}$. For an infinite sequence $\mathcal{P}$, the star discrepancy D_N^* is the discrepancy of the first N elements of $\mathcal{P}$. A sequence of points is said to be *low discrepancy* if $D_N^* \in O(N^{-1}\log(N)^s)$. The widely stated *Koksma-Hlawka theorem* states that if the function f has a variation $V(f)$ in the sense of *Hardy and Krause* that is finite, then we have that $|I - \hat{I}_N| \le V(f)D_N^*$. For a general introduction to LDS, QMC rules and their applications, refer to Leobacher and Pillichshammer (2014), Dick and Pillichshammer (2010) or Lemieux (2009). In this paper, the main QMC rules used in the numerical experiments are rank-1 lattice rules. These are rules in which

$$\boldsymbol{x}_i = \left\{ \frac{i\boldsymbol{z}}{N} \right\}, \quad 1 \le i \le N. \tag{1.2}$$

Here the s components of $\boldsymbol{z}$ are integers in $\{1,2,\ldots,N-1\}$ and $\{x\} = x - \lfloor x \rfloor$ denotes the fractional part of $x \in \mathbb{R}$ which is applied component-wise for vectors. Although these are finite point sets and not sequences, the convergence rate of $O(N^{-1}\log(N)^s)$ is still guaranteed (Leobacher and Pillichshammer 2014) . More information about lattice rules is also available in Niederreiter (1992) or Sloan and Joe (1994).

The three types of point sets that we discuss in this paper (grids, random points, LDS) can all be described using a common general description that we give below.

The point set $\mathcal{P}_N$:
In (1.1), let the components of each $\boldsymbol{x}_i$ be denoted by $x_{i,j}$ for $1 \le j \le s$. Let us now assume that for a fixed j and $\forall i = 1,\ldots,N$, there are n distinct values of $x_{i,j}$ which we denote by z_k for $1 \le k \le n$. Here, for simplicity of notation, we have not included a j subscript. Further, let us assume that there are exactly m points that have the value z_k for their j^{th} subscript, for each k, $1 \le k \le n$. So the total number of points N satisfies $N = nm$. Note that this description of point sets, which, from now on, we refer to as $\mathcal{P}_N$, in fact, covers a number of point sets including random points used for the MC integration rule. In particular, it includes an n-point grid and the rank-1 lattice rule shown in Figure 1. As seen in Figure 1 [a], in an $n-$ point regular grid, the points are aligned in rows and columns, each containing n points. As a result, there are n distinct z_k's along each axis and $m = n$. On the other hand as illustrated in Figure 1 [b], in a rank-1 lattice, none of the points are aligned resulting in $n = N$ distinct z_k's along each axis and $m = 1$.

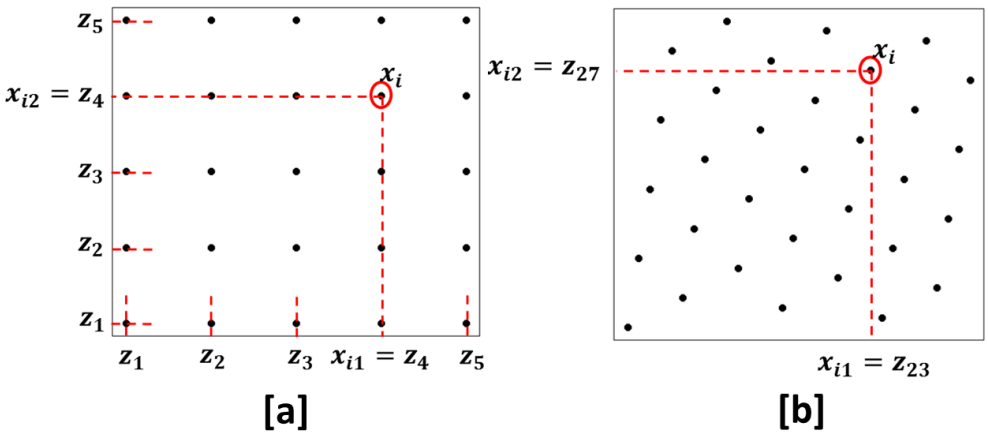

Figure 1. [a] 5−point grid ($m = 5$) and [b] 32−point rank-1 lattice ($m = 1$).

1.3. *Approximation to the one-dimensional functions using deterministic point sets*

Suppose that we are interested in approximating the functions

$$f_j(x) = \int_{[0,1]^{s-1}} f(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_s)\, \mathrm{d}\boldsymbol{x}_{-j}, \quad x \in [0,1],$$

where, for a vector $\boldsymbol{u} = (u_1, \ldots, u_s)$, $\boldsymbol{u}_{-j}$ denotes $(u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_s)$, for $1 \le j \le s$. So f_j is the function obtained by integrating out all the variables of f except the j-th one. The set of points $\{\boldsymbol{x}_i,\ 1 \le i \le N\}$ could be obtained either by sampling randomly (MC approach) or using a n-point regular grid or using a QMC approach. An integration rule of the form (1.1) can be used to approximate the one dimensiona functions. However, note that this approach does not approximate the *shape* of the one dimensional function. By *shape* we mean the graph of the one dimensional function (see Figure 2, columns 2,3, (a),(c)).

Example 1.1. As mentioned previously, the regular n-point grid consists of the points $((i_1-1)/(n-1), (i_2-1)/(n-1), \ldots, (i_s-1)/(n-1))$, where $1 \le i_\ell \le n$ for $1 \le \ell \le s$. For the j-th coordinate of these $N = n^s$ points, we have the n distinct values $(i_j-1)/(n-1)$, $1 \le i_j \le n$. As $N = n^s = nm$, it follows that $m = n^{s-1}$.

Example 1.2. As mentioned previously, the points of an ℓ-point rank-1 lattice rule are given by $\{i\boldsymbol{z}/\ell\}$, where $\boldsymbol{z} \in \{1, 2, \ldots \ell - 1\}^s$. Now let r be relatively prime with ℓ. Then one can obtain the lattice rule with point set given by

$$\left\{ \frac{i\boldsymbol{z}}{\ell} + \frac{(k_1, k_2, \ldots, k_s)}{r} \right\}, \quad 1 \le i \le \ell,\ 0 \le k_1, k_2, \ldots, k_s \le r - 1.$$

Such a lattice rule has $N = \ell r^s$ points and is an example of a maximal rank lattice rule (for example, see Sloan and Joe (1994)). Assuming that all the components of $\boldsymbol{z}$ are relatively prime with ℓ, then it may be shown that the j-th coordinate of these N points consists of the $n = \ell r$ distinct values $(i-1)/(\ell r)$ for $1 \le i \le \ell r$ with each value repeated $m = r^{s-1}$ times. We note that in the $r = 1$ case, the lattice rule is just

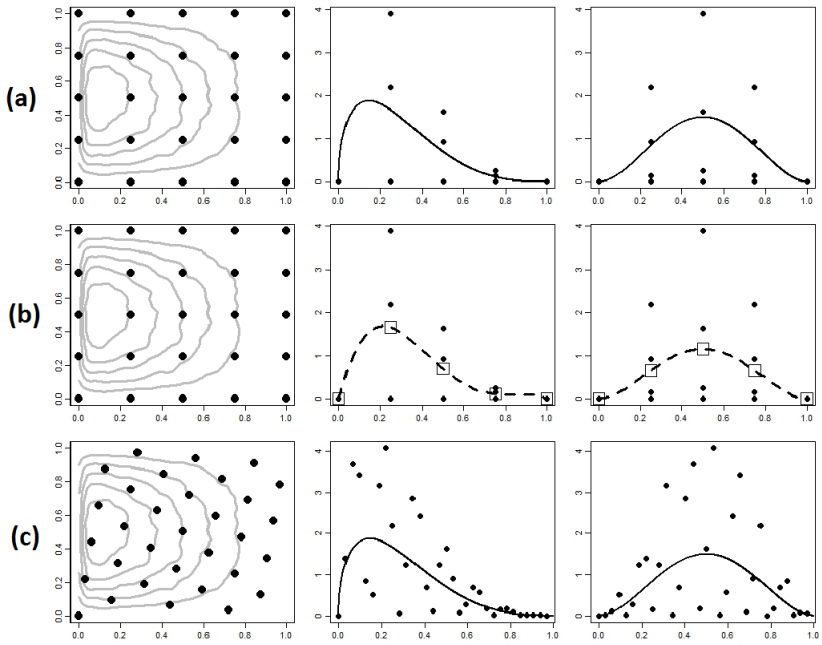

Figure 2. *First column:* Bi-variate Beta distribution contours along with the points used to approximate the one-dimensional functions: (a),(b) 5−point grid ($m = 5$) and (c) 32−point rank-1 lattice ($m = 1$). *Second and third columns*: the true-one dimensional functions along with the *unique* orthogonal projections of the bi-variate Beta distribution for a 5−point grid ($m = 5$) in Row (a) and a 32−point rank-1 lattice ($m = 1$) in Row (c). Row (b) shows an interpolant fit through the point-wise means (squares) for the 5−point grid. Because some of the function projections are the same, the number of points in the function projections are fewer than the total number of points shown in the first column on which the bi-variate function is evaluated.

a rank-1 lattice rule having a total of $\ell = N$ points. Moreover, the j-th coordinate of these points has the N distinct values $z_k = (k-1)/N$ for $1 \le k \le N$ with each value occurring just once (so that $N = nm$ with $n = N$ and $m = 1$). In the terminology of lattice rules, the lattice rule is said to be *fully projection regular* (see Lemieux (2009), Sloan and Joe (1994)). This property is also clearly illustrated in Figure 1.

We have that

$$f_j(z_k) = \int_{[0,1]^{s-1}} f(x_1, \ldots, x_{j-1}, z_k, x_{j+1}, \ldots, x_s)\, \mathrm{d}\boldsymbol{x}_{-j}$$

can be approximated using numerical integration by

$$\hat{f}_j(z_k) = \frac{1}{m} \sum_{\boldsymbol{x}_i : x_{i,j} = z_k} f(\boldsymbol{x}_i). \tag{1.3}$$

So $\hat{f}_j(z_k)$ is the point-wise mean obtained by averaging out over the m points, for each of whom, $x_{i,j} = z_k$. With these approximations to $f_j(z_k)$ for $1 \le k \le n$, one can then approximate the shape of f_j by fitting an interpolant through these n approximations. Note that, $\hat{f}_j(z_k)$ can be considered as the pointwise mean of the orthogonal projections of $f(\cdot)$ on the j^{th} axis. This is illustrated in Figure 2. An interpolant through the point-wise means of the orthogonal projections of

the bi-variate Beta distribution can approximate the shape of the one dimensional functions reasonably accurately for the 5−point grid ($m = 5$) (Figure 2 (b)). But the rank-1 lattice is fully projection regular, i.e., $m = 1$. Although such a property is advantageous for the numerical integration of integrands over $[0,1]^s$, it is not so advantageous when trying to approximate the shape of the one dimensional functions. We would not expect the approximation to the shape of f_j obtained by fitting an interpolant through the point-wise means (1.3) to be an accurate one when $m = 1$. For the 32−point rank-1 lattice, the point-wise means of the orthogonal projections of the bi-variate Beta distribution are the projections themselves (Figure 2 (c)) and one can see that an interpolant that passes through each one of them would not approximate the shape of the one dimensional function very accurately at all.

1.4. *Structure of This Paper*

In Section 2, we propose a new method that involves use of an integration rule as well as fitting of a polynomial smoothing function to approximate the shape of the one dimensional function. The theoretical results will be presented in Section 3. In Section 4, we provide some numerical results illustrating the efficiency and accuracy of the approximations produced by our new method as compared to those produced by a grid based method. Finally, we close in Section 5 giving a summary of the work and discuss further challenges.

2. New Method

Here we propose a method for approximating the shape of the one-dimensional functions

$$f_j(x) = \int_{[0,1]^{s-1}} f(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_s)\, \mathrm{d}\boldsymbol{x}_{-j} \quad x \in [0,1],$$

when, an s-dimensional function $f(\boldsymbol{x})$ has been evaluated at N distinct points $\boldsymbol{x} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ given by a point set $\mathcal{P}_N$. As discussed in Section 1.3, an interpolant through the point-wise means may not provide an accurate approximation when using the QMC integration rules. However, a smoothing function such as a least square polynomial fitted to the projected points may be a better option. Thus, the basic algorithm we propose is as follows:

Algorithm I ($m > 1$)

(1) Evaluate the function f at N points $\boldsymbol{x}_i$.
(2) For $j = 1, \ldots, s$, do:
    (a) Project the function evaluations $f(\boldsymbol{x}_i)$ on the j^{th} axis.
    (b) Fit a polynomial of degree $(n-1)$ to the projections.
(3) Repeat for each j.

As in Section 1.3, let the components of each $\boldsymbol{x}_i$ be denoted by $x_{i,j}$ for $1 \leq j \leq s$, $1 \leq i \leq N$. These components together with the function evaluations may be conveniently

represented in a matrix form as

$$\boldsymbol{\Psi}_{N\times(s+1)} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,s} & f(\boldsymbol{x}_1) \\ x_{2,1} & x_{2,2} & \cdots & x_{2,s} & f(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,s} & f(\boldsymbol{x}_N) \end{bmatrix}.$$

To approximate the shape of the j^{th} one-dimensional $f_j(x)$, we first orthogonally project $f(\boldsymbol{x}_i)$ on the j^{th} axis to obtain

$$\psi_j = \begin{bmatrix} x_{1,j} & f(\boldsymbol{x}_1) \\ x_{2,j} & f(\boldsymbol{x}_2) \\ \vdots & \vdots \\ x_{N,j} & f(\boldsymbol{x}_N) \end{bmatrix},$$

More formally, we can write $\psi_j = \boldsymbol{\Psi} P_j$, where P_j is the $(s+1) \times 2$ matrix with zeros everywhere except for ones in the j-th position of the first column and the last position of the second column.

Example 2.1. When $j = 2$, we have

$$P_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T.$$

Since the spread of the projected function points is not constant (as illustrated by Figure 2), a weighted least square polynomial may be required where the weights are proportional to the variances. However, we prove that in this case, a weighted least square polynomial of degree $(n-1)$ is equal to the ordinary least square polynomial of the same degree.

Let $\underline{M}$ be the design matrix when fitting a least squares polynomial of degree $(n-1)$ through the orthogonal projections of $f(\boldsymbol{x})$ on x_j. Such a projection has n unique abscissa points z_k, $k = 1, \ldots, n$, as described in Section 1.3. Then $\underline{M}$ is of size $N \times n$, and has a block structure,

$$\underline{M} = \begin{bmatrix} \mathbf{1} & \boldsymbol{t}_1 & \boldsymbol{t}_1^2 & \ldots & \boldsymbol{t}_1^{n-1} \\ \mathbf{1} & \boldsymbol{t}_2 & \boldsymbol{t}_2^2 & \ldots & \boldsymbol{t}_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1} & \boldsymbol{t}_n & \boldsymbol{t}_n^2 & \ldots & \boldsymbol{t}_n^{n-1} \end{bmatrix},$$

where each element block $\boldsymbol{t}_k^p \in \underline{M}, (p = 0, \ldots, n-1)$ is an $m \times 1$ column vector containing only the element z_k^p. We can also express $\underline{M}$ as a Kronecker product of the Vandermonde matrix M and the $m \times 1$ column vector of $1's$,

$$\underline{M} = M \otimes \mathbf{1}_{(m\times 1)},$$

where, M is a square Vandermonde matrix of size n, which is of full rank and is invertible since all elements z_k are unique.

For weighted least squares, we assign a weight w_k to all projections corresponding to a unique abscissa point z_k. We define the weights matrix $\underline{W}$ of size $N \times n$ by

$$\underline{W} = \begin{pmatrix} w_1 I_m & 0I_m & \cdots & 0I_m \\ 0I_m & w_2 I_m & \cdots & 0I_m \\ \vdots & \vdots & \ddots & \vdots \\ 0I_m & 0I_m & \cdots & w_n I_m \end{pmatrix},$$

where, I_m is the identity matrix with size $m \times m$. $\underline{W}$ can also be expressed as a Kronecker product

$$\underline{W} = W \otimes I_m,$$

where W is the $n \times n$ diagonal matrix of weights

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}.$$

In what follows, we will make use of the following well known properties (Gentle 2007) of the Kronecker product.

(1) **Scalar property**: For matrices A and B, and scalar k

$$(kA) \otimes B = A \otimes (kB) = k(A \otimes B).$$

(2) **Mixed product property**: For matrices A, B, C, and D, such that AC and BD exist, then

$$(A \otimes B)(C \otimes D) = AC \otimes BD.$$

(3) **Inverse property**: If matrices A and B are invertible, then $(A \otimes B)^{-1}$ exists, and can be expressed as

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

(4) **Transposition**: For matrices A and B

$$(A \otimes B)^T = A^T \otimes B^T.$$

Let $\hat{f}_j^{WLS}$ be the weighted least square polynomial approximation of degree $(n-1)$ to the j^{th} one-dimensional function f_j and $\hat{f}_j^{LS}$ be the least square polynomial approximation of the same degree. Further, let $\underline{\hat{f}_j^{LS}}$ be the values taken by $\hat{f}_j^{LS}$ for the elements in the design matrix $\underline{M}$. Similarly, $\underline{\hat{f}_j^{WLS}}$.

Theorem 2.1. *For any* $j \in \{1, \ldots, s\}$, $\hat{f}_j^{WLS} = \hat{f}_j^{LS}$.

Proof. We have

$$\begin{aligned}\underline{\hat{f}_j^{LS}} &= \underline{M}(\underline{M}^T\underline{M})^{-1}\underline{M}^T\boldsymbol{f} \\ &= (M\otimes\mathbf{1})\left[(M\otimes\mathbf{1})^T(M\otimes\mathbf{1})\right]^{-1}(M\otimes\mathbf{1})^T\boldsymbol{f} \\ &= (M(M^TM)^{-1}M^T)\otimes(\mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T)\boldsymbol{f}.\end{aligned}$$

Since M is a square Vandermonde matrix and invertible, and $\mathbf{1}^T\mathbf{1} = m$, we have

$$\begin{aligned}\underline{\hat{f}_j^{LS}} &= (MM^{-1}(M^T)^{-1}M^T)\otimes(\frac{1}{m}\mathbf{1}\mathbf{1}^T)\boldsymbol{f} \\ &= \frac{1}{m}I_n\otimes(\mathbf{1}\mathbf{1}^T)\boldsymbol{f}.\end{aligned} \tag{2.1}$$

We have

$$\begin{aligned}\underline{\hat{f}_j^{WLS}} &= \underline{M}(\underline{M}^T\underline{W}\underline{M})^{-1}\underline{M}^T\underline{W}\boldsymbol{f} \\ &= (M\otimes\mathbf{1})\left((M\otimes\mathbf{1})^T(W\otimes I_m)(M\otimes\mathbf{1})\right)^{-1}(M\otimes\mathbf{1})^T(W\otimes I_m)\boldsymbol{f} \\ &= (M(M^TWM)^{-1}M^TW)\otimes(\mathbf{1}(\mathbf{1}^TI_m\mathbf{1})^{-1}\mathbf{1}^TI_m)\boldsymbol{f}.\end{aligned}$$

Since W is also square and invertible (W is a diagonal matrix, with $w_{i,i} > 0$), and $(\mathbf{1}^TI_m\mathbf{1}) = m$, we have

$$\begin{aligned}\underline{\hat{f}_j^{WLS}} &= (MM^{-1}W^{-1}(M^T)^{-1}M^TW)\otimes(\frac{1}{m}\mathbf{1}\mathbf{1}^TI_m)\boldsymbol{f} \\ &= (I_nW^{-1}I_nW)\otimes(\frac{1}{m}\mathbf{1}\mathbf{1}^TI_m)\boldsymbol{f} \\ &= \frac{1}{m}I_n\otimes(\mathbf{1}\mathbf{1}^T)\boldsymbol{f} = \hat{f}_j^{LS}(z_k).\end{aligned}$$

□

We can further show that $\hat{f}_j^{LS}$ will pass through $\hat{f}_j(z_k)$ for each k.

Theorem 2.2. *For any $j \in \{1,\ldots,s\}$, $\hat{f}_j^{LS}$ will pass through $\hat{f}_j(z_k)$ for $1 \le k \le n$.*

Proof. Using Equation (2.1) we have that

$$\begin{aligned}\underline{\hat{f}_j^{LS}} &= \frac{1}{m}(I_n\otimes\mathbf{1}\mathbf{1}^T)\boldsymbol{f} \\ &= \frac{1}{m}\begin{pmatrix}J_m & 0_m & \ldots & 0_m\\ 0_m & J_m & \ldots & 0_m\\ \vdots & \vdots & \ddots & \vdots\\ 0_m & 0_m & \ldots & J_m\end{pmatrix}\begin{pmatrix}\boldsymbol{f}_1\\ \boldsymbol{f}_2\\ \vdots\\ \boldsymbol{f}_n\end{pmatrix} = \begin{pmatrix}\tilde{f}_{j,1}\\ \tilde{f}_{j,2}\\ \vdots\\ \tilde{f}_{j,n}\end{pmatrix},\end{aligned}$$

where each element J_m or 0_m is a square matrix of size $m \times m$ that contains all 1's or all 0's respectively and $\boldsymbol{f}_k$, $k = 1, \ldots, n$ is the $m \times 1$ vector of function evaluations $f(\boldsymbol{x})$ corresponding to z_k. □

For fully projection regular point sets such as many of the LDS, including the rank-1 lattice rules, $m = 1$ and using Algorithm I in such cases will imply fitting a polynomial of degree $(N-1)$ passing through all of the N function projections. Such a polynomial will not approximate the desired shape accurately. Here, we propose a partitioning approach to overcome this problem. Suppose we partition $[0,1]$ into n partitions, with breakpoints given by $0 = z_0 < z_1 < z_2 < \ldots < z_{n-1} < z_n = 1$. As above, we assume that the total number of points N factorises as $N = nm$. Further, we assume the points are such that there are exactly m points whose j-th component belongs to $[z_k, z_{k+1})$ for $0 \leq k \leq n-1$. Note that these assumptions are not necessary for the validity of the theory, instead, they have been made only to simplify the notation. We provide below the modified algorithm with a partitioning step.

Algorithm II ($m = 1$)

(1) Evaluate the function f at N points $\boldsymbol{x}_i$.
(2) For $j = 1, \ldots, s$, do:
  (a) Project the function evaluations $f(\boldsymbol{x}_i)$ on the j^{th} axis.
  (b) Partition $[0,1]$ into n partitions, with breakpoints given by $0 = z_0 < z_1 < z_2 < \ldots < z_{n-1} < z_n = 1$.
  (c) Fit a polynomial of degree $(n-1)$ to the projections.
(3) Repeat for each j.

Similar to (1.3), one can calculate

$$\tilde{f}_{j,k}(z_k) = \frac{1}{m} \sum_{\boldsymbol{x}_i : x_{i,j} \in [z_k, z_{k+1})} f(\boldsymbol{x}_i). \tag{2.2}$$

Let $\tilde{f}_j^{LS}$ be the least square polynomial of degree $(n-1)$. Then, we can show that $\tilde{f}_j^{LS}$ will pass through $\tilde{f}_{j,k}(z_k)$ for each k.

Theorem 2.3. *For any $j \in \{1, \ldots, s\}$, $\tilde{f}_j^{LS}$ will pass through $\tilde{f}_{j,k}$ for $0 \leq k \leq n-1$.*

Proof. The proof is similar to that of Theorems 2.1 and 2.2. □

MC integration rules generate points that are fully projection regular *w.p.* (*with probability*) 1. Therefore Algorithm II, approximation (2.2) and Theorem 2.3 are also applicable when the function has been evaluated using a random point set.

3. Convergence theorems

3.1. *For point sets where $m > 1$*

The new approach described in the previous section essentially involves evaluating f on a set of N points in $[0,1]^s$ and then approximating the one-dimensional function f_j by fitting a least square polynomial through the orthogonal projections $f(\boldsymbol{x}_i)$ of $f(\cdot)$ on the j^{th} axis. Theorem 2.2 proves that $\hat{f}_j^{LS}$ passes through the n point-wise means

$\hat{f}_j(z_k)$. This implies that this approach is equivalent to the interpolating polynomial approach where a polynomial of degree $(n-1)$ is fitted to n function evaluations. Therefore the convergence properties can be studied using the relevant literature in numerical analysis. We assumed that there were $N = n \times m$ points in $[0,1]^s$ such that f_j is approximated at n distinct points z_k, $1 \leq k \leq n$, and that for each unique value of z_k, there is a subset of m points whose k-th co-ordinate is equal to z_k.

The choice of the points z_k is crucial and determines the convergence properties and the computational efficiency as discussed below. The next theorem gives the convergence result when the z_k are equidistant points (in a grid).

Theorem 3.1. *Suppose that f_j is infinitely differentiable such that*

$$\max_{\xi \in [0,1]} |f_j^{(n)}(\xi)| \leq C, \, \forall n,$$

for some $C < \infty$ such that $\frac{C}{(n-1)^n} \ll 1, \forall n$. If the z_k are equidistant points, then $\hat{f}_j^{LS} \to f_j$ as $m \to \infty$ and $n \to \infty$.

Proof. As $m \to \infty$,

$$\hat{f}_j(z_k) = \frac{1}{m} \sum_{\boldsymbol{x}_i : x_{i,j} = z_k} f(\boldsymbol{x}_i) \to f_j(z_k). \tag{3.1}$$

Equation (3.1) holds due to the Koksma-Hlawaka inequality (Niederreiter (1992)) if the $\boldsymbol{x}_i$ are sampled using a grid.
For the interpolating polynomial of degree $n-1$, it follows from a standard result in approximation theory (see for example, Cheney and Kincaid (1999), Kress (1998)), that

$$\max_{z \in [0,1]} |f_j(z) - \hat{f}_j^{LS}(z)| \leq \max_{\xi \in [0,1]} \frac{|f_j^{(n)}(\xi)|}{n!} \max_{z \in [0,1]} \prod_{k=1}^{n} |z - z_k|.$$

This implies that

$$\max_{z \in [0,1]} |f_j(z) - \hat{f}_j^{LS}(z)| \leq \frac{C}{n!} \max_{z \in [0,1]} \prod_{k=1}^{n} |z - z_k|. \tag{3.2}$$

It is known that (see for example, Cheney and Kincaid (1999)) that if the n points z_k are equidistant on $[0,1]$, then

$$\max_{z \in [0,1]} \prod_{k=1}^{n} |z - z_k| \leq \frac{(n-1)!}{4} \left(\frac{1}{n-1} \right)^n.$$

From (3.2), we then have

$$\max_{z \in [0,1]} |f_j(z) - \hat{f}_j^{LS}(z)| \leq \frac{C}{4n(n-1)^n}.$$

The assumption that $\frac{C}{(n-1)^n} \ll 1$ for all n then implies that as $m \to \infty$ and $n \to \infty$, $\hat{f}_j^{LS} \to f_j$. □

If the function f_j is n times differentiable then the result in Theorem 3.1 indicate that the approximation obtained using $\hat{f}_j^{LS}$ will still be good as long as the derivatives are sufficiently bounded.

3.2. *For fully projection regular point sets where* $m = 1$

Theorem 3.1 provides the conditions under which $\hat{f}_j^{LS} \to f_j$ for grids constructed using equidistant points. Now, we show that the polynomial approximation will converge to the shape of the true one dimensional function if the function was explored using LDS instead of a grid.

Theorem 3.2. *Let $h_k = z_{k+1} - z_k$ using the partitions defined in Algorithm II, and points $\boldsymbol{x}_i$ sampled using a QMC integration rule. If $\tilde{f}_{j,k}$ is as given in (2.2), then $\tilde{f}_{j,k} \to f_j(z_k)$ as $m \to \infty$ and $h_k \to 0$.*

Proof. One may consider $\tilde{f}_{j,k}$ as an approximation to the integral

$$\frac{1}{h_k} \int_{[0,1]^{s-1}} \int_{z_k}^{z_{k+1}} f(\boldsymbol{x}) \, \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j}. \tag{3.3}$$

As $m \to \infty$, $\tilde{f}_{j,k}$ converges to this integral due to the Koksma-Hlawaka inequality (Niederreiter (1992)). For the integral in (3.3), we can swap the order of integration by Fubini's theorem since f is integrable and Lebesgue measure is a σ-finite measure. So the integral becomes

$$\frac{1}{h_k} \int_{z_k}^{z_{k+1}} \int_{[0,1]^{s-1}} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{-j} \, \mathrm{d}x_j = \frac{1}{h_k} \int_{z_k}^{z_{k+1}} f_j(x_j) \, \mathrm{d}x_j.$$

Letting $h_k \to 0$, it follows from the definition of derivative that this integral converges to $f_j(z_k)$. □

Theorem 3.3. *Suppose that f_j is infinitely differentiable such that*

$$\max_{\xi \in [0,1]} |f_j^{(n)}(\xi)| \leq C, \, \forall n,$$

for some $C < \infty$ such that $\frac{C}{(n-1)^n} \ll 1$, $\forall n$. If the z_k are equidistant points and points $\boldsymbol{x}$ sampled using a QMC integration rule, then $\tilde{f}_j^{LS} \to f_j$ as $m \to \infty$ and $n \to \infty$.

Proof. The result follows from Theorems 3.1, 2.3 and 3.2. □

Note that if the function f_j is n times differentiable then the results in Theorem 3.3 indicate that the approximation obtained using $\tilde{f}_j^{LS}$ will still be good as long as the derivatives are sufficiently bounded.

3.3. *For random point sets*

As pointed out in Section 2, Algorithm II, approximation (2.2) and Theorem 2.3 are also applicable when the function has been evaluated using a random point set. We provide the corresponding result for this case.

Theorem 3.4. *Let $h_k = z_{k+1} - z_k$ using the partitions defined in Algorithm II, and points $\boldsymbol{x}$ sampled using a MC integration rule. If $\tilde{f}_{j,k}$ is as given in (2.2), then $\tilde{f}_{j,k} \to f_j(z_k)$ w.p. 1 as $m \to \infty$, and $h_k \to 0$.*

Proof. Proof is similar to Theorem 3.2 except that as $m \to \infty$, $\tilde{f}_{j,k}$ converges to the integral (3.3) *w.p.* 1 because of the law of large numbers. The rest of the proof is exactly the same. □

Theorem 3.5. *Suppose that f_j is infinitely differentiable such that*

$$\max_{\xi \in [0,1]} |f_j^{(n)}(\xi)| \leq C, \, \forall n,$$

for some $C < \infty$ such that $\frac{C}{(n-1)^n} \ll 1$, $\forall n$. If the z_k are equidistant points and points $\boldsymbol{x}$ sampled using a MC integration rule, then $\tilde{f}_j^{LS} \to f_j$ as $m \to \infty$ and $n \to \infty$.

Proof. The result follows from Theorems 3.1, 2.3 and 3.4. □

Note that if the function f_j is n times differentiable then the results in Theorem 3.5 indicate that the approximation obtained using $\tilde{f}_j^{LS}$ will still be good as long as the derivatives are sufficiently bounded.

4. Numerical Examples

The algorithms proposed in Section 2 can be used when a function is explored using a grid, MC or QMC integration rules. However, because this work was motivated by the need to develop a method for QMC integration rules (and no other method exists, to our best knowledge), we focus on QMC integration rules in the examples below. Wherever possible, we also compare the results against those obtained using a grid. Since this problem was motivated by a possible application in the Bayesian statistical inference, we illustrate using a few standard probability distributions.

The integration rules used are known as Korobov lattice rules. These are rank-1 lattice rules in which the generating vector $\boldsymbol{z}$ in (1.2) is of the form

$$\boldsymbol{z} = (1, \alpha, \alpha^2, \ldots, \alpha^{s-1}),$$

where α is an integer in $\{1, 2, \ldots, N-1\}$. Appropriate choices of α may be found by using the Lattice Builder software (see L'Ecuyer and Munger (to appear)).

4.1. *Exponential distribution*

Most statistical distributions are smooth with bounded derivatives and therefore satisfy the smoothness requirements of Theorems 3.1, 3.3 and 3.4. Here, we illustrate how the exponential distribution, for example, satisfies these smoothness conditions.

The Exponential distribution is slightly different since the derivative does not exist at zero. However, here we show that it still satisfies the smoothness conditions imposed by Theorems 3.1, 3.3 and 3.4. Suppose that the j-th one dimensional distribution is exponential with parameter λ. Then we have that,

$$f_j(x) = \lambda e^{-\lambda x};$$

the n^{th} derivative is given by

$$f_j^{(n)}(x) = (-1)^n \lambda^{n+1} e^{-\lambda x},$$

and

$$\sup_x |f_j^{(n)}(x)| = \lim_{x \to 0+} |f_j^{(n)}(x)| = \lambda^{n+1}.$$

We assume here that the interval of interest is $[0, b)$ for some $b < \infty$, b large enough so that $\int_0^b f_j(x)\, dx \approx 1$. Note that the convergence results proved in Section 3 are applicable here since the function can be linearly transformed to be defined over $[0, 1]$. Then, $\exists n' > 0$ and $c < 1$ such that $\forall n > n' + 1$, $\frac{b}{n-1} \leq \frac{1}{nc} < 1$. Further, for any $\lambda < \infty$, $\exists n'' > n'$ such that, $\forall n > n'', \lambda^{n+1} \left(\frac{1}{nc}\right)^n \ll 1$.

Thus, it can be seen that conditions for Theorem 3.3 are satisfied and $\tilde{f}_j^{LS} \to f_j$ as $m \to \infty$ and $n \to \infty$. This is illustrated in Figure 3. Here, the joint distribution is bi-variate and is a product of two Exponential distributions. We find the least squares approximations to the marginals using Korobov lattices with different n and m, the convergence is achieved as they both increase.

4.2. *Multi-modal and skewed distributions*

Figures 4 and 5 illustrate that a grid is quite inefficient at accurately capturing the shape of the distribution even in low dimensional problems, especially when it is multi-modal or heavily skewed. Here, we consider a multi-modal distribution and the Beta distribution, in four variables, and try to approximate the shape of the marginals using the grid points (and fitting the interpolant through pointwise means) as well as using LDS points and our new method of fitting the least squares polynomials of degree $(n-1)$ through the orthogonal projections of the joint distribution on the marginals proposed in this paper.

Figure 4 shows that the marginals approximated using the Korobov lattice with 4096 points are very accurate whereas the approximation using an $8-$point grid with the same number of points ($8^4 = 4096$) is not as accurate. Figure 5 illustrates that the approximations to Beta marginals using a 1024 point Korobov lattice are much more accurate than the approximations obtained using grids with $6^4 = 1296$ or even $8^4 = 4096$ points. Thus, using LDS enables efficient and more accurate approximation of the shape of the one-dimensional distributions.

4.3. *High-dimensional distributions*

To illustrate the real computational benefit of using low discrepancy sequences, we consider two distributions of dimensions 10 and 12 respectively. These distributions

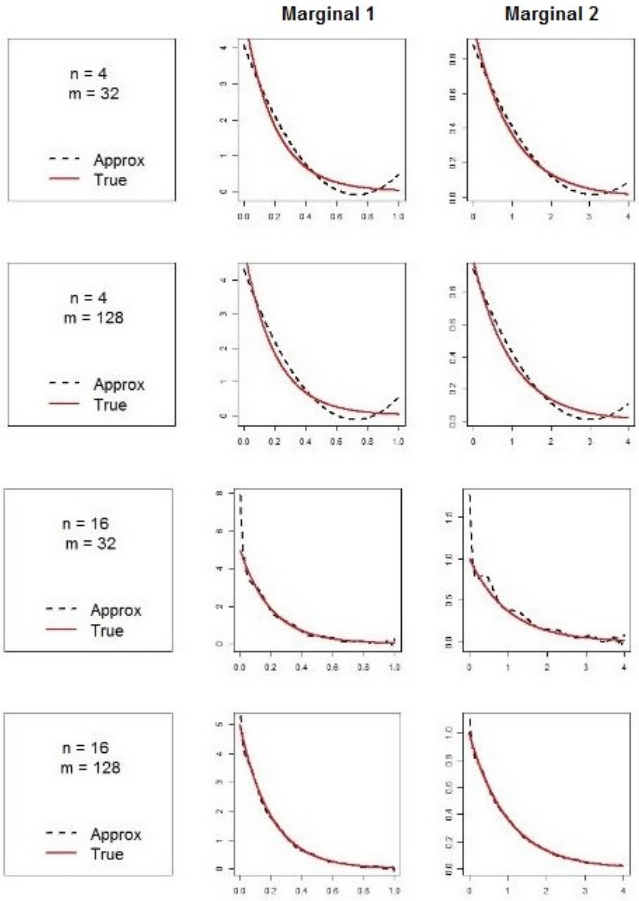

Figure 3. Least squares approximation to the Exponential marginals using Korobov lattices as n and m increase.

have been generated as products of independent Gamma distributions with different parameters. A 5-point grid will require $5^{10} = 9,765,625$ points in 10 dimensions and $244,140,625$ points in 12 dimensions and will likely still yield inaccurate estimates, as illustrated by an inability of $n-$point grids to capture various shapes when n is small in Figures 4 and 5.

Figure 6 shows that for $s = 10$, very accurate estimates can be obtained using LDS with as little as 2^{16} points (150 times fewer than a 5-point grid). Although estimates obtained using 2^{17} points are even more accurate, the difference between the two is very small suggesting that our estimates have started to converge to the true marginals. For 12-dimensional Gamma, 2^{16} points give reasonably accurate estimates and the convergence is achieved by $2^{19}(= 524,288)$ points as can be seen in Figure 7. However, this is negligible compared to the 244 million points required for a 5-point grid.

5. Summary and Discussion

This paper proposes a new method to approximate the shape of one dimensional functions f_j, where, f_j is the function obtained by integrating out all the variables of an $s-$dimensional function f except the j-th one and where the function has been ex-

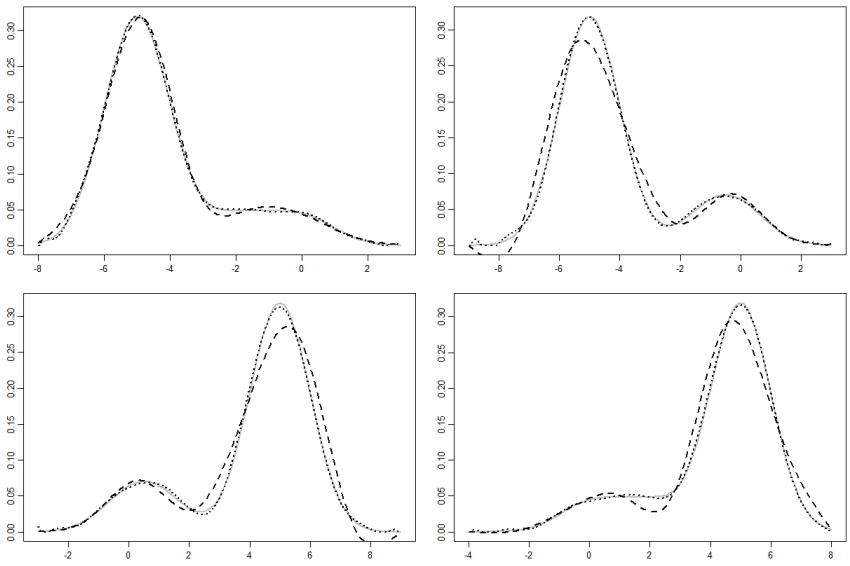

Figure 4. Approximating marginals of a four-dimensional multi-modal distribution (line) using: Korobov lattices with 4096 points (dotted) and an 8−point grid with 4096 points (dashed).

plored using a point set. Not only is this method easy and computationally efficient but also, it can be used when the function is evaluated using the grid, the MC or the QMC integration rules. To our best knowledge, a formal method to solve this problem has not been proposed yet, especially for QMC integration rules. The method uses a least squares polynomial smoother. We propose two algorithms - two versions of the method - one where the point set used are fully projection regular (or fully projection regular $w.p.\,1$, in case of MC rules) and the other when this is not the case. We prove the convergence properties for both these algorithms. We show that implementing our new method using LDS points only requires $O(mn)$ function evaluations, compared to the traditional grid based approaches that require $O(n^s)$ function evaluations. Typically, $m < n^{(s-1)}$ and therefore implementing our new method using LDS points is computationally more efficient than using an n point grid. Further, the examples illustrate that our method also produces more accurate approximation than using the traditional grid based approach.

In practice, the polynomials could be fitted to an appropriate transformation (for example, a log transformation) of the distributions. This may improve the computational stability of the algorithm. However, in this paper, we have considered fitting the polynomials to an untransformed probability distribution to show that the convergence exists in a more general case where a transformation may not be possible or desirable.

The need to develop such a method was motivated by a potential application in Bayesian statistics, specifically, in computational methods that explore the posterior distribution using a set of deterministic point sets as discussed in Section 1.1. However, practical challenges will need to be overcome before the method developed here can be incorporated within the computation Bayesian methods. For instance, the proposed method provides asymptotic guarantees as the number of points and the degree of the polynomial go to infinity. However, it cannot specify the number of points and the degree of the polynomial needed to achive a reasonable approximation for a given function or indeed for a wide range of functions (class of all continuous

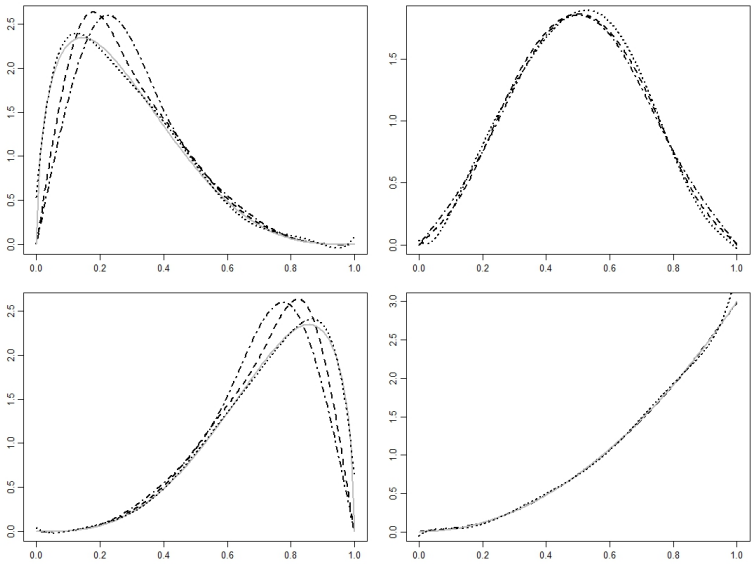

Figure 5. Approximating marginals of a four-dimensional Beta distribution (line) using: Korobov lattices with 1024 points (dotted), a 6−point grid with 1296 points (dash-dotted) and a 8−point grid with 4096 points (dashed).

probability distributions, for example). Thus further work will be required to develop a method that can potentially improve the computational efficiency of Bayesian methods using QMC integration rules.

However, to the best of our knowledge, this paper presents the first formal method developed to approximate the shape of the one-dimensional function obtained by integrating out all other variables using LDS.

References

Brown, Paul T., Chaitanya Joshi, Stephen Joe, and Hvard Rue. 2021. "A novel method of marginalisation using low discrepancy sequences for integrated nested Laplace approximations." *Computational Statistics & Data Analysis* 157: 107147.

Cheney, W., and D. Kincaid. 1999. *Numerical Mathematics and Computing*. 4th ed. ITP.

Dick, J., and F. Pillichshammer. 2010. *Digital Nets and Sequences*. Cambridge.

Gentle, J.E. 2007. *Matrix Algebra: Theory, Computations and Applications in Statistics*. Springer.

Kress, R. 1998. *Numerical Analysis*. Springer.

L'Ecuyer, P., and D. Munger. to appear. "LatticeBuilder: A general software tool for constructing rank-1 lattice rules." *ACM Transactions on Mathematical Software* .

Lemieux, C. 2009. *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer.

Leobacher, G., and F. Pillichshammer. 2014. *Introduction to Quasi-Monte Carlo Integration and Applications*. Birkhauser.

Martins, Thiago G., Daniel Simpson, Finn Lindgren, and Hvard Rue. 2013. "Bayesian computing with INLA: New features." *Computational Statistics & Data Analysis* 67: 68–83.

Niederreiter, H. 1992. "Random Number Generation and Quasi-Monte Carlo methods." *SIAM CBMS-NSF Regional conference series in Applied Mathematics* 63.

Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian inference for latent Gaussian models by using a integrated nesed Laplace approximations." *Journal of the Royal*

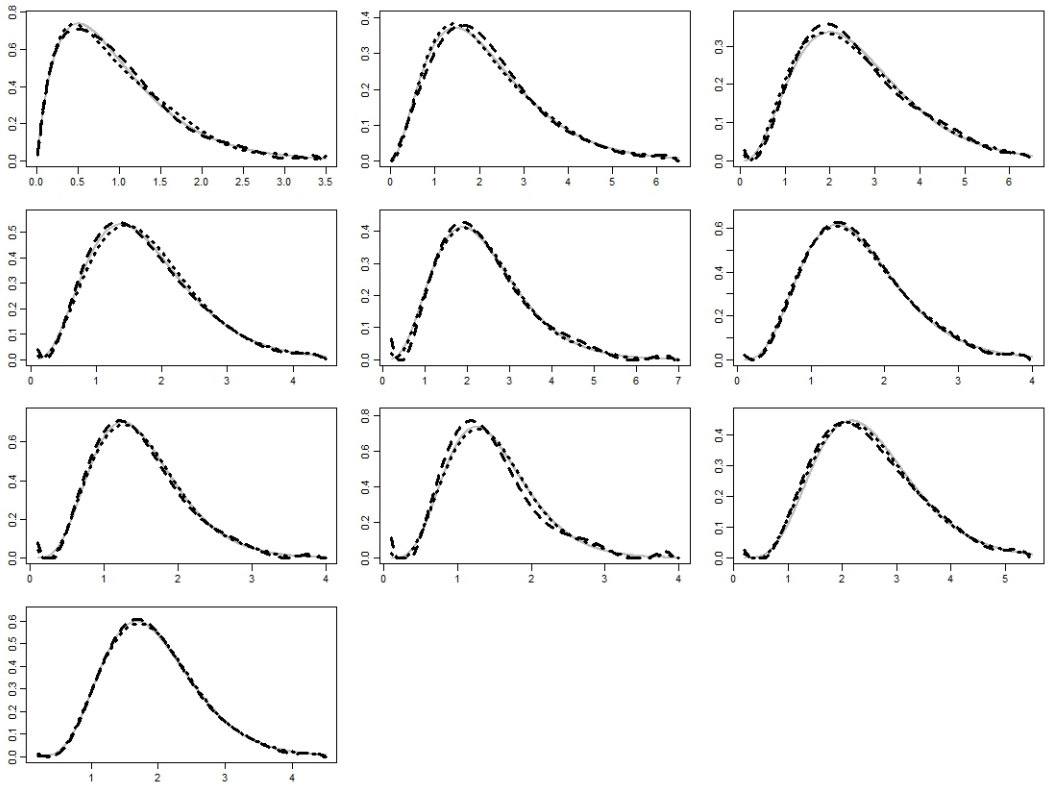

Figure 6. 10-dimensional Gamma using Korobov lattice with i) $2^{16} = 65,536$ (dashed) and ii) $2^{17} = 131,072$ points (dotted).

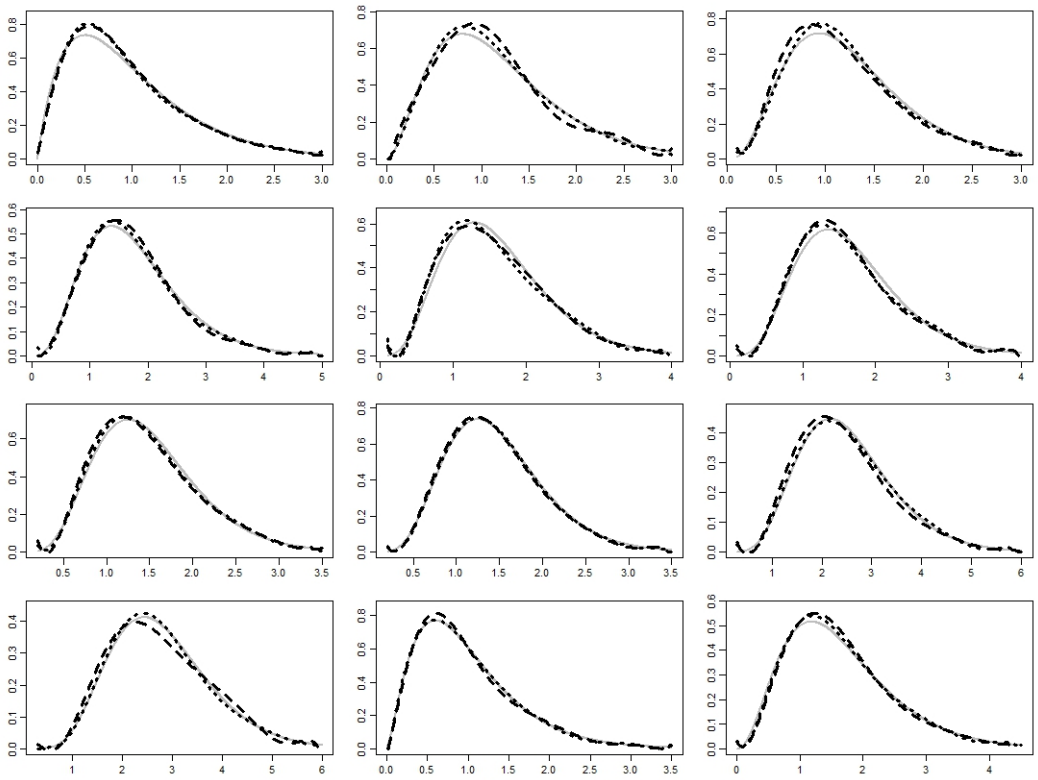

Figure 7. 12-dimensional Gamma using Korobov lattice with i) $2^{16} = 65,536$ (dashed) and ii) $2^{19} = 524,288$ points (dotted).

Statistical Society - Series B 71 (Part2): 318 – 392.
Sloan, I. H., and S. Joe. 1994. *Lattice Methods for Multiple Integration.* Oxford.