Available online at www.sciencedirect.com
SCIENCE DIRECT®
Reliability Engineering and System Safety 91 (2006) 1266–1280
RELIABILITY ENGINEERING & SYSTEM SAFETY
www.elsevier.com/locate/ress

Comparison of pure and "Latinized" centroidal Voronoi tessellation against various other statistical sampling methods

Vicente J. Romero[a,*], John V. Burkardt[b], Max D. Gunzburger[b], Janet S. Peterson[b]

[a]*Model Validation and Uncertainty Quantification Department, Sandia National Laboratories, Albuquerque, NM 87185-0828, USA*
[b]*School of Computational Science and Information Technology, Florida State University, Tallahassee, FL 32306-4120, USA*

Available online 19 January 2006

Abstract

A recently developed centroidal Voronoi tessellation (CVT) sampling method is investigated here to assess its suitability for use in statistical sampling applications. CVT efficiently generates a highly uniform distribution of sample points over arbitrarily shaped *M*-dimensional parameter spaces. On several 2-D test problems CVT has recently been found to provide exceedingly effective and efficient point distributions for response surface generation. Additionally, for statistical function integration and estimation of response statistics associated with uniformly distributed random-variable inputs (uncorrelated), CVT has been found in initial investigations to provide superior points sets when compared against latin-hypercube and simple-random Monte Carlo methods and Halton and Hammersley quasi-random sequence methods. In this paper, the performance of all these sampling methods and a new variant ("Latinized" CVT) are further compared for *non*-uniform input distributions. Specifically, given uncorrelated normal inputs in a 2-D test problem, statistical sampling efficiencies are compared for resolving various statistics of response: mean, variance, and exceedence probabilities.

Keywords: Centroidal Voronoi tessellation; Statistical sampling methods; Uncertainty propagation

1. Introduction and background

It is often beneficial in statistical sampling and function integration to sample "uniformly" over the applicable parameter space. Such uniformity, while conceptually simple and intuitive on a qualitative level, is on a quantitative level somewhat complicated to describe and characterize mathematically. Quantitative aspects of uniformity involve: (1) the equality with which points are spaced relative to one another in the parameter space (are they all nominally the same distance from one another?); (2) uniformity of point density over the entire domain of the parameter space (i.e., uniform "coverage" of the whole domain by the set of points, and not just good uniformity within certain regions of the space); and (3) isotropy (no preferential directionality) in the point placement pattern. Each of these aspects of uniformity can be quantified by several mathematical measures as described in Ref. [1]. We will not discuss these measures further here, but mention them to indicate that quantitative mathematical measures do exist for the intuitive notion of uniformity.

We find that for 2-D data sets the eye is an excellent discriminator of the different aspects of uniformity listed above. The intuitive sense of uniformity obtained from viewing sample sets in a unit square (2-D hypercube) usually correlates very strongly with the quantitative measures. Thus, for 2-D data sets like the ones we present later, fairly accurate visual judgements can be made about whether one particular layout of sample points is mathematically more uniform than another, or whether the uniformity varies significantly over the parameter space.

Achieving high sampling uniformity over generic domains is an area of active research. Much effort has been applied to the problem of achieving uniform placement of N samples over M-dimensional hypercubes, where M and N are both arbitrary. It is well recognized that

*Corresponding author. Tel.: +1 505 844 5890; fax: +1 505 844 8251.
E-mail addresses: vjromer@sandia.gov (V.J. Romero), burkardt@csit.fsu.edu (J.V. Burkardt), gunzburg@csit.fsu.edu (M.D. Gunzburger), jspeters@csit.fsu.edu (J.S. Peterson).

0951-8320/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2005.11.023

simple-random sampling (SRS) Monte Carlo does not do a particularly good job of uniformly spreading out the sample points. The popular latin hypercube sampling (LHS) method [2–4] generally does a much better job of uniformly spreading out the points. This is due to the greater sampling regularity over each individual parameter dimension before the individually generated parameter values are randomly combined into parameter sets which define the coordinates of the sampling points.

Recent efforts to modify LHS to get an even more uniform distribution of points over the parameter space have included distributed hypercube sampling (DHS [5]) and improved distributed hypercube sampling (IHS [6]). The fundamentals and history of these are reviewed briefly in [7]. Though the quantitative measure of uniformity used for comparisons in [5] and [6] was somewhat flawed, it does appear that DHS gives better sampling uniformity than LHS, and IHS gives better sampling uniformity than DHS (but is increasingly more computationally expensive as the dimensionality of the parameter space increases). We have recently become aware of another LHS variant, "optimal symmetric LHS" (OSLHS [8]) which also seems to improve the spatial uniformity of LHS samples. Its computational cost and performance relative to DHS and IHS are not yet known, however.

A number of other potential approaches for achieving uniform point placement that are not evolved from an LHS basis are reviewed (and some new ones are presented) in [9]. There, some quantitative metrics related to visual/sensory perception of point uniformity in 2-D are reviewed and some new metrics are presented. Many of these non-LHS-based approaches appear to work very well in 2-D, but it is said in [9] that some of the methods may not be applicable or may not perform well in more than two dimensions, and some clearly will not scale up to high dimensions affordably. Others seem more promising for high dimensions, but have not yet been investigated enough.

The so-called "quasi-random" (QR, see e.g., [10]) low-discrepancy sequence methods can often achieve reasonably uniform sample placement in hypercubes. A strength of some of these sequence methods (Halton, Sobol, etc.), is that they can produce fairly uniform point distributions even though samples are added one at a time to the parameter space. The one-at-a-time incremental sampling of certain QR methods (and SRS) enables these methods to have improved efficiency prospects versus CVT- and LHS-type methods in the area of error estimation and control (see e.g., [11]). The results achieved with QR sampling are often quite good. For instance, when resolving the mean and standard deviation of response measures, Hammersley sequences were found in [12] to converge to within 1% of exact results 3–100 times faster than LHS over a large range of test problems. For resolving response probabilities, Hammersley and a modified-Halton method were found in [13] to perform roughly the same as LHS on balance over several test problems.

However, when the hyperspace dimension becomes moderate to large and/or the sampling density becomes high, some (perhaps all?) sequences suffer from spurious correlation of the samples. This is shown for standard Halton sequences in 16-D (Ref. [5]) and 40-D (Ref. [13]). Sometimes a modification can be found to suppress or delay the onset of spurious correlation—as a fix from the literature implemented in [13] shows for Halton sequences.

Recently, a long-recognized approach for achieving uniformity of point placement in M-dimensional volumes, called "Centroidal Voronoi Tessellation" (CVT), has been made computationally efficient [14] for implementing the principles of centroidal Voronoi diagrams [15,16]. These diagrams subdivide arbitrarily shaped domains in arbitrary-dimensional space into arbitrary numbers of nearly uniform subvolumes, or Voronoi cells/regions. Given a set of N points $\{z_i\}$ $(i = 1, \ldots, N)$ in an M-dimensional hypercube, the Voronoi region or Voronoi cell V_j $(j = 1, \ldots, N)$ corresponding to z_j is defined to be all points in the space that are closer to z_j than to any of the other z_i's. The set $\{V_i\}$ $(i = 1, \ldots, N)$ is called a Voronoi tessellation or Voronoi diagram of the hypercube, the set $\{z_i\}$ $(i = 1, \ldots, N)$ being the generating points or generators. A CVT is a special Voronoi tessellation with the property that each generating point z_i is itself the mass centroid of the corresponding Voronoi region V_i.

Although CVTs are deterministic, they can be converged to with probabilistic sampling methods. In [14], new probabilistic CVT construction algorithms were introduced, implemented, and tested. These methods are generally much more computationally efficient than previous deterministic and probabilistic methods for constructing CVTs.

The CVT concept and the algorithms in [14] for their construction can be generalized in many ways (see [16,17]). For example, instead of a hypercube, general regions in M-dimensional space can be treated. This feature has been exploited with great success for discretizing arbitrary 2-D and 3-D domain volumes for computational mechanics analysis with meshless analogues of finite element methods (see [17]). Furthermore, points can be distributed non-uniformly according to a prescribed density function over the space (like the bi-normal density function that Fig. 7 in this paper corresponds to).

In initial investigations [1] for 2-D, 3-D, and 7-D test cases, CVT has provided greater sampling uniformity than Halton, Hammersley, Sobol, SRS, LHS, DHS, and IHS according to a meaningful subset of non-flawed quantitative quality measures. Additionally, no degradation of sampling uniformity has been detected in higher dimensions (i.e., for the 7-D case).

It is therefore natural to ask whether CVT can be applied for: (A) statistical sampling over arbitrary-dimensional spaces of input random variables to calculate various statistics of output response behavior; (B) function integration over arbitrarily shaped domains; and

(C) whether it can serve as a method for generating favorable point distributions for improved response-surface accuracy.

A preliminary positive indication regarding item (C) for response surface generation is presented in [7]. There, CVT was shown on several 2-D test problems to provide superior point distributions for generating locally conforming moving least-squares response surfaces. Point distributions by CVT, SRS, LHS, and a structured sampling method with deterministically uniform point placement [18] were tried in the study.

Ref. [19] compared sampling method performance in 2-D test problems of statistical function integration and estimation of response statistics associated with uniformly distributed random-variable inputs (uncorrelated). By the same weighted measure of sampling effectiveness defined and used in Section 3.3 of this paper, CVT handily outperformed SRS, LHS, Halton, and Hammersley in resolving various statistics of response: mean, variance, and probability quantiles.

In this paper we take a first step toward examining the potential of CVT for improved statistical sampling given *non*-uniform inputs. Specifically, the performance of the above sampling methods and a new CVT variant (''Latinized'' CVT) are compared for uncorrelated Normal input variables in a 2-D test problem. Statistical sampling efficiencies are compared for calculating response mean, variance, and ''exceedence'' probabilities.

2. Uniformly distributed test point-sets and their mapping to binormal joint densities

Fig. 1 shows three LHS and three corresponding CVT point sets for 100 samples in a 2-D unit hypercube. The three LHS point sets were generated with the software [20] for different initial seeds (Seed 1 = 123456789, Seed 2 = 192837465, Seed 3 = 987654321) and a uniform joint probability density function (JPDF) over the unit-square parameter space. The three corresponding CVT point sets were generated with the software [21] by using the LHS sets as initial conditions (starting point locations) from which the CVT iterations begin. In all cases each CVT set is much more uniform visually (and quantitatively, see [1]) than its associated LHS set. All three CVT sets are relatively similar visually and quantitatively, even though starting from three very different initial conditions given by the LHS sets.

The LHS sets exhibit significantly more clustering and non-uniformity of the points than the CVT sets. For a visual indicator of sampling uniformity, Fig. 2 compares a 25-sample LHS set and a 25-sample CVT set started from the LHS set. Non-overlapping circles are drawn in each domain, where each sample point has a circle centered about it having a radius proportional to the distance from the point to its nearest neighbor. The surrounding circles for the CVT set are all fairly uniform in size, whereas the variance in circle size is very large for the LHS set. Thus, the LHS point sets are relatively non-uniform in their ''coverage'' of the domain.

Besides the three LHS sets and three corresponding CVT sets shown in Fig. 1, three SRS sets generated from initial Seeds 1, 2, and 3 will also be tested here. These point sets are presented in Ref. [7]. They exhibit even less uniformity than the LHS sets in Fig. 1 here. Three CVT sets derived from the three different SRS sets as initial conditions can also be seen in [7]. The different LHS and SRS initial conditions do not have much of an impact on final CVT point uniformity, so the CVT algorithms appear to be robust in this regard.

Figs. 3 and 4 show Halton and Hammersley point sets and the corresponding CVT sets derived from them. Again, the resulting CVT sets are of essentially equivalent uniformity. The Halton point set is noticeably and quantitatively more uniform than any of the LHS sets; the Hammersley set is even more uniform than the Halton set; and the CVT sets in Figs. 1, 3, and 4 are even more uniform than the Hammersley set.

In Ref. [19] we compared the mentioned 2-D point sets for effectiveness in statistical function integration and estimation of response statistics for the case of uniformly distributed input random variables (uncorrelated). The CVT point sets performed best, as will be summarized in Section 4 of this paper. In this paper we focus on comparing the performance of the above sampling methods and a new CVT variant (''Latinized'' CVT, see [1]) as starting sets for mapped non-uniform point distributions intended to reflect a JPDF of uncorrelated normal inputs.

Our 2-D test problem has two random inputs $p1$ and $p2$ from independent normal distributions having means 0.5 and standard deviations $\sigma = 0.5/3$. The corresponding JPDF is shown in Fig. 5 after truncation of the function beyond the unit $p1$–$p2$ parameter space and renormalization to integrate to one over the space.

The following procedure is used to map a set of uniformly distributed points to a set that reflects the desired non-uniform JPDF (Fig. 5). First, for each random variable p in the problem, we consider its cumulative distribution function CDF(p), where

$$\mathrm{CDF}(p) = \int_0^{p \leqslant 1} \mathrm{PDF}(p')\,\mathrm{d}p' \tag{1}$$

and PDF(p) is the probability density function of the random input p. We note that the value of the CDF ranges from 0 to 1 as the coordinate p ranges from 0 to 1. It can also be shown that realizations $\{p_i\}$ drawn at random from a density function PDF(p) map through Eq. (1) (setting $p = p_i$) into a uniformly distributed set of realizations $\{\mathrm{CDF}(p_i)\}$. This set is therefore distributed uniformly between 0 and 1.

Hence, we recognize that the above properties can be used to inverse-map numbers uniformly distributed between 0 and 1 (produced, e.g., by a random number generator) into realizations $\{p_i\}$ that would appear to be

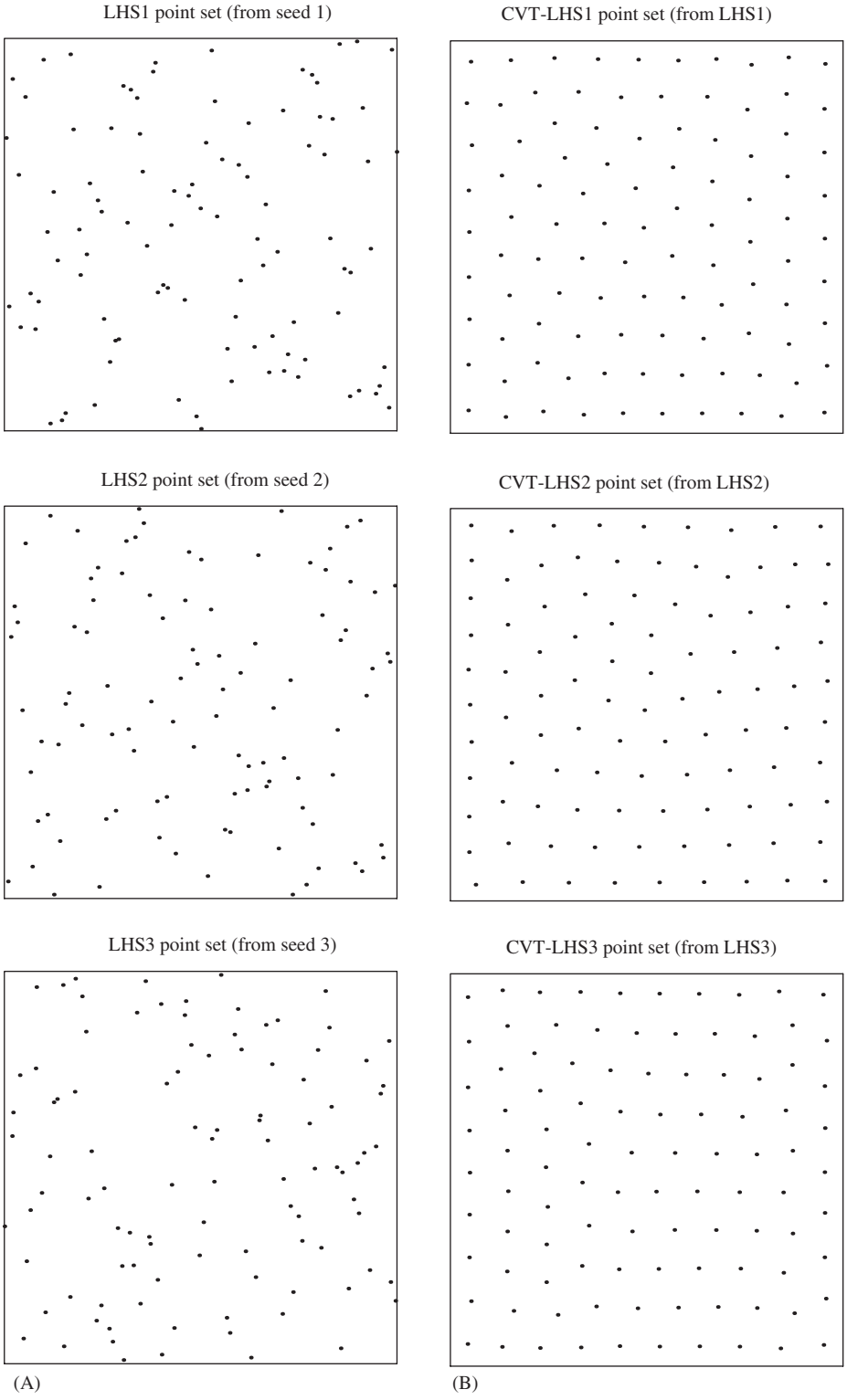

Fig. 1. One-hundred-point sample sets on a 2-D unit hypercube for: (A) left column—uniform JPDF LHS Monte Carlo with three different initial seeds; and (B) right column—corresponding uniform JPDF CVT sets starting from LHS sets as initial conditions.

drawn from the density function PDF(p). In a multi-dimensional problem, we inverse-map the coordinates of points uniformly distributed in the hypercube into transformed point sets of a non-uniform JPDF. Fig. 2 helps visualize the multidimensional mapping process. The tick marks on the coordinate axes indicate the projections of

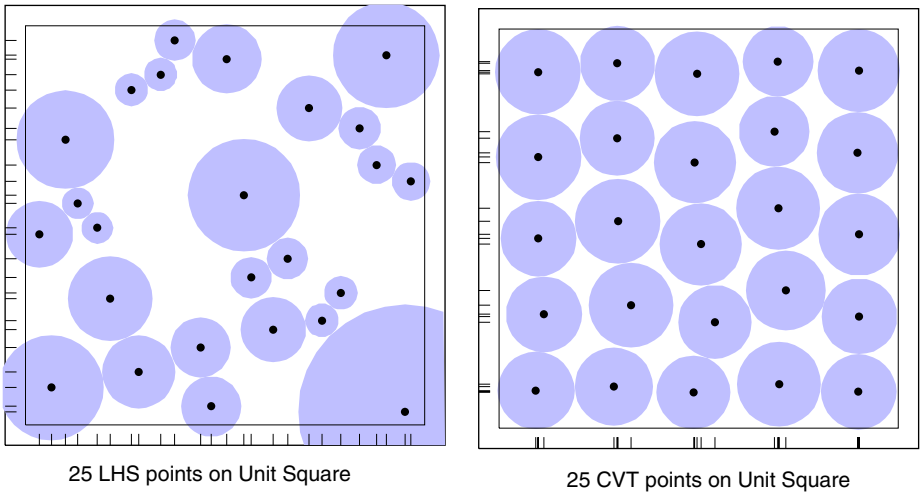

Fig. 2. LHS and CVT sample sets showing relative uniformities of point spacing and "discrepancies" of point projections onto coordinate axes.

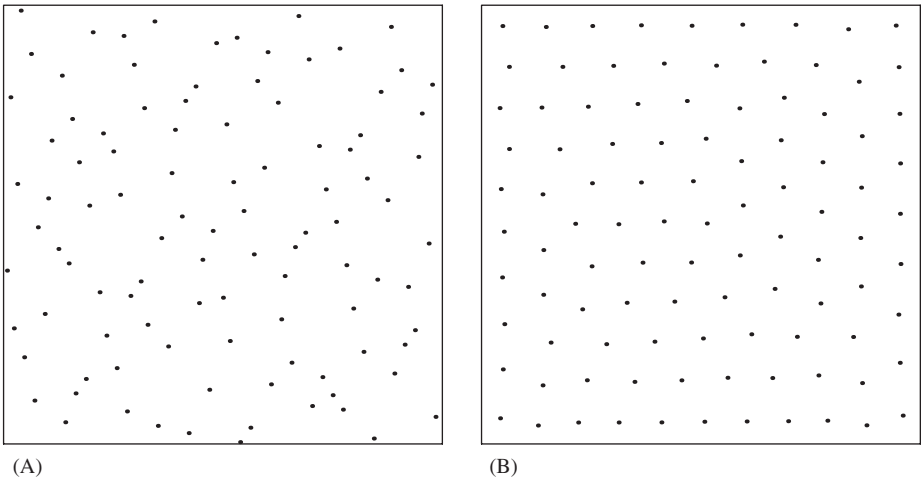

Fig. 3. One-hundred-point sample sets on 2-D unit hypercube for: (A) left plot—Halton QR sequence; and (B) right plot—corresponding CVT set starting from the Halton set as initial condition.

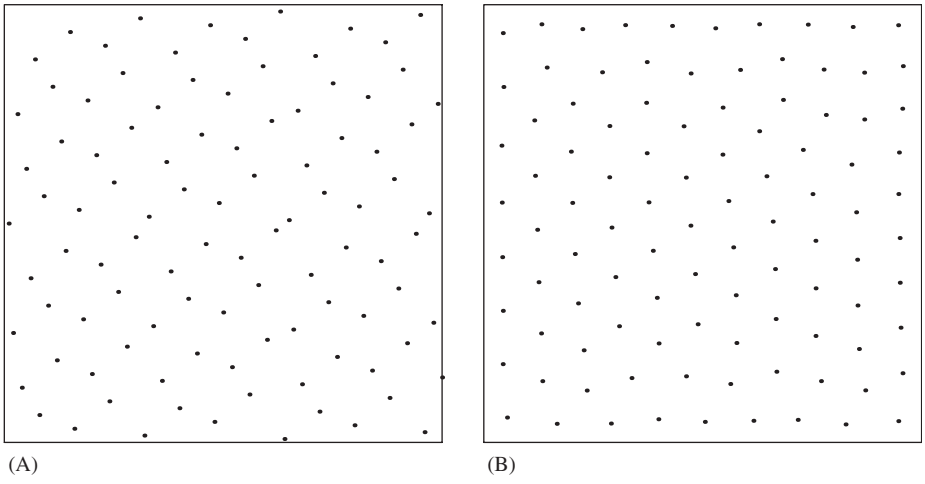

Fig. 4. One-hundred-point sample sets on 2-D unit hypercube for: (A) left plot—Hammersley QR sequence; and (B) right plot—corresponding CVT set starting from the Hammersley set as initial conditions.

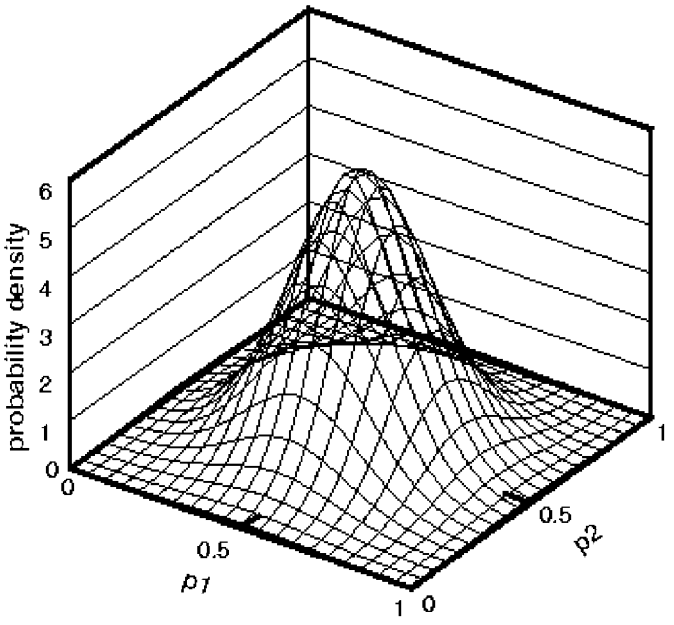

Fig. 5. Joint probability density function describing the random variables in the problem: normally distributed parameters $p1$ and $p2$ with means 0.5, standard deviations $\sigma = 0.167$, and truncation of the unit square parameter space at 3σ above and below the mean values.

the 2-D uniformly distributed LHS points onto the coordinate axes. The approximately uniformly distributed tick marks on each coordinate axis give the random values that are inverse-mapped through the marginal CDFs of the input variables into transformed tick locations ranging from 0 to 1 on the coordinate axes of the non-uniform JPDF space. These new tick locations tell the locations of the points in the non-uniform JPDF space, as translated from their initial locations in the uniform JPDF space. Thus, initially uniformly distributed point sets in the unit hypercube are transformed to non-uniformly distributed point sets in the unit hypercube. (Our transformation algorithm for mapping uncorrelated uniformly distributed points sets to bi-normally distributed point sets was verified as described later.) Correlation between input random variables can be imparted with the rank correlation procedure described in [22].

The mapping transformation presupposes a point set in an M-dimensional unit hypercube with point locations that project with uniform spacing onto all coordinate axes. However, consider the point sets in Fig. 2. Though the CVT set is more uniform *volumetrically* than the LHS set, the LHS points project more uniformly onto the coordinate axes. The LHS point set would sample each input variable at 25 nearly uniformly spaced values over its range. In contrast, the projections of the CVT points occur in clusters that portray a "banded-uniform" distribution over the 0–1 range on each axis. In the limit of a perfectly volumetrically uniform distribution of points over the domain, say a 5×5 rectangular array of points on the unit square, the points would project onto the coordinate axes making 5 uniformly spaced tick marks. These marks would inverse map through the marginal CDFs into only 5 different values or samples of each input variable. Thus, out of 25 sampling opportunities, each input variable is sampled at only five values. However, this is not *automatically* bad; the 25 particular *sets* or combinations of the five values of each input variable (when the uniform 5×5 grid of points is mapped to the JPDF space) may pose certain advantages over other point layouts. We are presently striving to understand the particular benefits and disadvantages that arise here.

One measure of a point set's uniformity of projection onto all the coordinate axes is called its *discrepancy*. As projection uniformity increases, discrepancy decreases. LHS is a lower-discrepancy sampling method than CVT is. Methods specifically designed with low discrepancy in mind are the quasi- or sub-random low-discrepancy sequence methods Halton, Hammersley, Sobol, etc. [10]. These can have both lower discrepancy than standard LHS *and* higher volumetric uniformity. Though CVT tends to have better volumetric uniformity than the sequence methods, which helps its relative performance in other areas (cf. [7,19]), it also has much higher discrepancy, which hurts its relative performance as a sampling basis for mapping to non-uniform JPDFs. Therefore, a hybrid of CVT and LHS has recently been formulated [1] which appears to have both lower discrepancy than pure CVT and higher volumetric uniformity than pure LHS. In the next section, we compare the performance of this hybrid "Latinized" CVT (LCVT) against pure CVT and other sampling methods.

Fig. 6 shows uniformly distributed point sets from SRS, LHS, CVT, LCVT, Halton, and Hammersley methods, and corresponding mapped bi-normal point sets. The SRS, LHS, CVT, and LCVT sets plotted in Fig. 6 correspond to random-number-generator Seed 1 described at the start of this Section. The SRS, LHS, CVT, and LCVT uniform and mapped sets are typical of the three sets obtained from the three different initial seeds listed earlier. Our mapping process was checked by verifying that our bi-normal results mapped from the Seed 1, 2, and 3 uniform LHS sets in Fig. 1 were essentially identical to bi-normal LHS sets generated directly from the LHS code [20] that produced the three uniform LHS sets. Thus, our mapping process corresponds almost exactly to the mapping process used in the well-pedigreed code [20].

The effect of high discrepancy in uniform CVT sets is immediately apparent in the mapped set in Fig. 6. The mapped CVT set has a rectangular-shaped layout of points rather than a circularly oriented layout seemingly more appropriate for the circularly symmetric bi-normal JPDF targeted (Fig. 5). Unexpectedly, we find in the next section that this non-intuitive rectangular-shaped set of points actually performs relatively well among the six types of mapped sets shown in Fig. 6. This rectangular-shaped set performs much better, in fact, than the much more likely

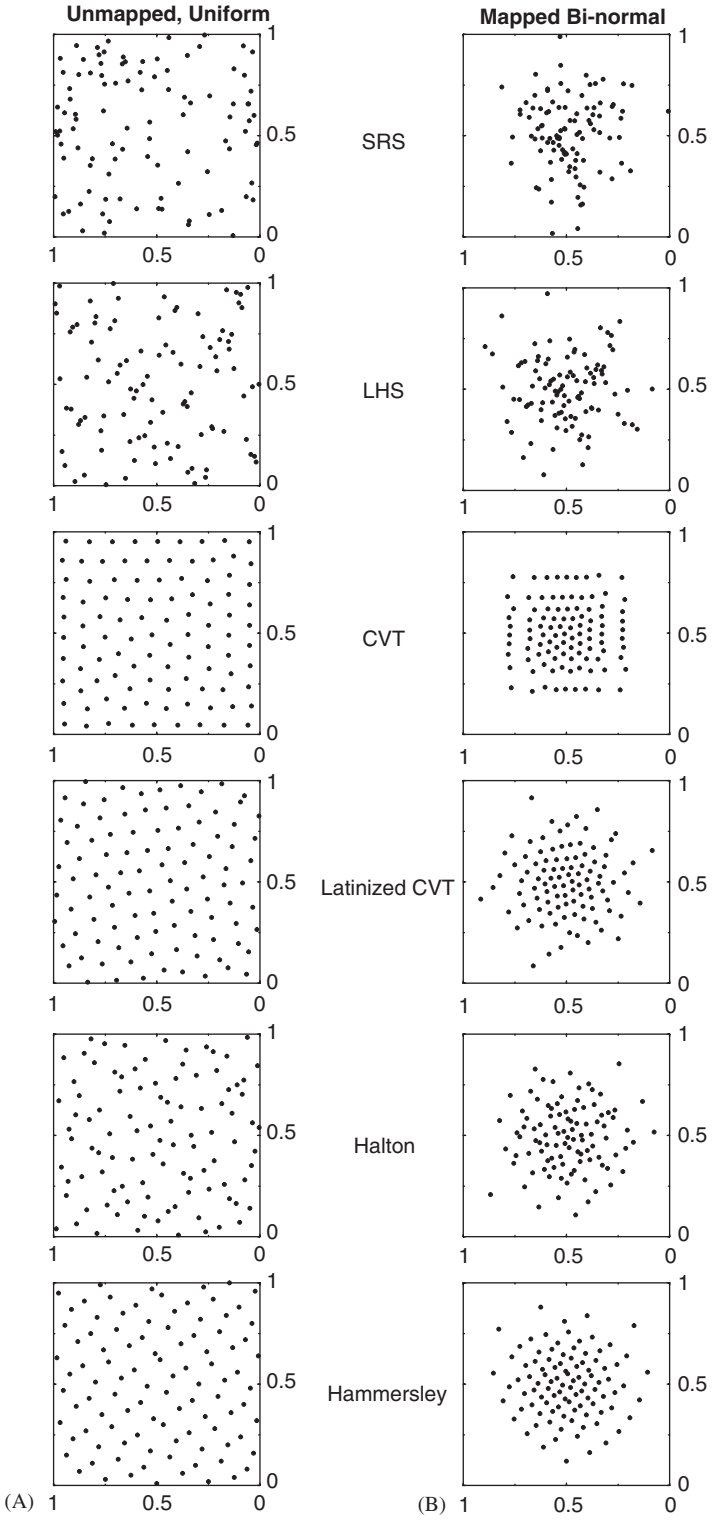

Fig. 6. One-hundred-point sample sets on a 2-D unit hypercube for: (A) left column—uniformly distributed point sets; and (B) right column—corresponding bi-normally distributed point sets mapped from uniform sets.

looking set shown in Fig. 7, which was generated directly with density-weighted CVT. The set mapped from uniform Latinized CVT appears much closer to a bi-normal density than the rectangular mapped CVT set, and performs even better. The performance of the various mapped sets is examined more closely in the next section.

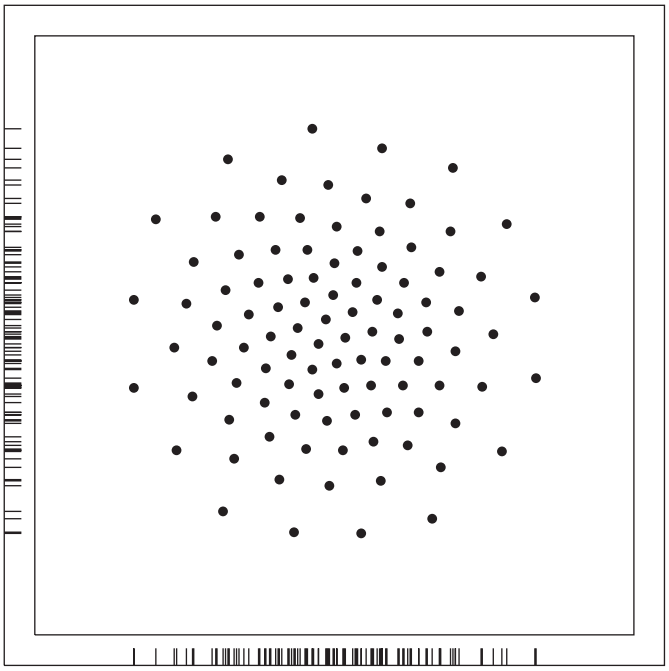

Fig. 7. One-hundred-point set in a unit square, generated directly with density-weighted CVT to model the bi-normal joint probability density function shown in Fig. 5.

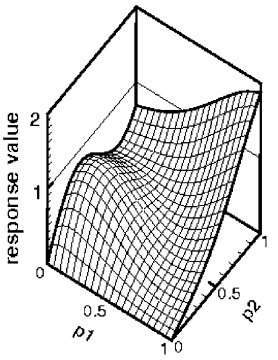

Fig. 8. 2-D model function for system response as a function of input parameters $p1$ and $p2$.

3. Statistical sampling effectiveness of the methods

3.1. 2-D model response function and statistical measures of response behavior

Fig. 8 shows an analytic multi-modal function describing deterministic system response r as a function of two system inputs $p1$ and $p2$:

$$r(p1,p2) = \left[0.8\kappa + 0.35\sin\left(2.4\pi\frac{\kappa}{\sqrt{2}}\right)\right][1.5\sin(1.3\theta)] \qquad (2)$$

on the domain $0 \leqslant p1 \leqslant 1$ and $0 \leqslant p2 \leqslant 1$, where $\kappa = \sqrt{(p1)^2 + (p2)^2}$, $\theta = a\tan(p2/p1)$.

A statistical problem arises if $p1$ and $p2$ are random variables. In that case, any particular realization $p1_i$ and $p2_i$ of the stochastic variables yields a response r_i as given by the above functional relationship. An ensemble of responses accompanies the different realizations of $p1$ and $p2$ as they vary stochastically or randomly according to their individual propensities.

The joint probability density function JPDF($p1$, $p2$) (Fig. 5) for attaining various input combinations of $p1$ and $p2$ maps through the response function $r(p1,p2)$ into a corresponding probability density function for response values, PDF (r). Operationally, the response probability density function can be approached closer and closer (for un-biased sampling methods) as more and more parameter sets or realizations $(p1,p2)_i$ are randomly generated from the governing input JPDF and are propagated through the response function $r(p1,p2)$ into response realizations r_i. The response realizations are distributed in the response space (i.e., along the response coordinate axis r) with a density that, as more and more samples are added, trends toward the exact PDF of response.

Very often, only certain statistical measures of the PDF of response are desired or can be reasonably estimated. Response mean, μ_r, and standard deviation, σ_r, can be estimated directly from the mean $\hat{\mu}_r$ and standard deviation $\hat{\sigma}_r$ of the population or set $\{r_i\}$ of realizations. We have the following definitions:

$$\hat{\mu}_r = \frac{1}{N}\sum_{i=1}^{N} r_i, \qquad (3)$$

$$\hat{\sigma}_r = \left[\frac{1}{N-1}\sum_{i=1}^{N}(r_i - \hat{\mu}_r)^2\right]^{1/2}, \qquad (4)$$

where N is the number of realizations or samples of response.

Also of interest is the probability of response exceeding (or not exceeding) some particular threshold value r_T. The former is equivalent to the volume integral of the joint probability density function Fig. 5 integrated over the region of the $p1$–$p2$ domain where response exceeds the stipulated threshold value r_T. Three such regions corresponding to thresholds $r_T = 1.0$, 0.5, and 0.2 are shown in Fig. 9 (as shaded areas on the cutting planes $r_T = 1.0$, 0.5, 0.2) for our response function Eq. (2).

Exceedence probability is very simply estimated as the ratio of the number of calculated response values at or above the given threshold value, to the total number of samples N of response. As the number of response realizations increases, the estimate tends toward greater accuracy, i.e., toward the value of the true exceedence probability if an un-biased sampling method is used. This is of course also true for the estimates $\hat{\mu}_r$ and $\hat{\sigma}_r$ of response mean and standard deviation.

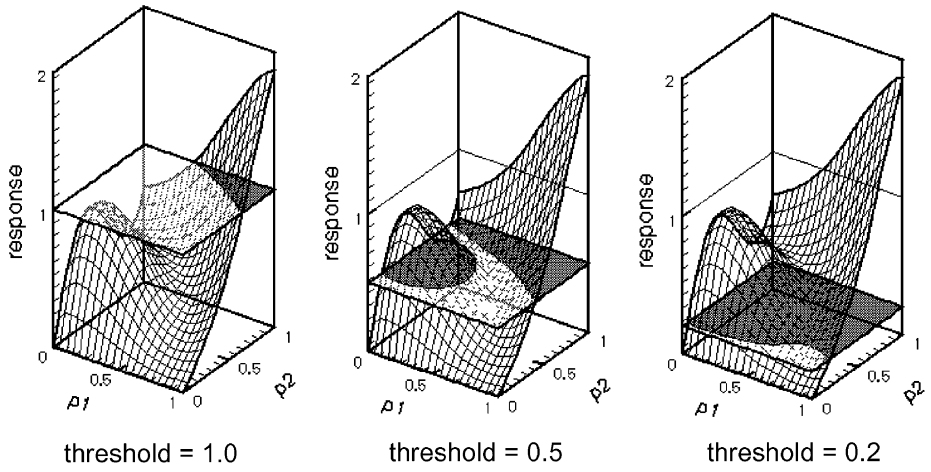

Fig. 9. Cutting planes through exact function showing associated exceedence (shaded) and complement (unshaded) regions of the $p1$–$p2$ parameter space for response threshold values of 1.0, 0.5, and 0.2, respectively.

3.2. Comparison of response statistics from various sampling methods

Here we compare estimates of response mean, standard deviation, and exceedence probabilities as obtained from the various sample sets represented by the right column in Fig. 6. We map these $\{(p1,p2)_i\}$ sets through our response function Eq. (2) to obtain corresponding response sets, and then calculate the aforementioned statistics of the response populations. We then compare the calculated statistics of each response set to reference values obtained from using three million SRS samples at parameter values generated by the SRS option of the sampling code [20]. The reference values are actually averages of three results, each obtained from one million samples generated from random initial seeds "X", "Y", and "Z" (different from Seeds 1, 2, and 3 used to generate the 100-sample sets).

Three "replicate" sets of one million samples each are used in preference to one set of three million samples so that *empirical* confidence intervals (CI) on the calculated averages could be compared against their classical CI to reaffirm or caveat them. (Recent research [23,24] has shown that for SRS, empirical CI appear to be somewhat more accurate than classical CI.) Empirical CI are formed by assuming the calculated statistic (response mean, standard deviation, or exceedence probability) is a random realization from a normal or nearly normal distribution about the exact result. Hence a T-distribution with $3-1=2$ degrees of freedom can be used to get CIs about the small-sample average of the three replicates. Thus, for 95% empirical CI the following formula is used:

$$95\%\ \text{confidence half-interval} = 4.303\frac{\hat{\sigma}_{\text{est}}}{\sqrt{3}}, \tag{5}$$

where $\hat{\sigma}_{\text{est}}$ is the standard deviation of the three estimates.

Table 1 shows various estimates of response mean, standard deviation, and exceedence probabilities calculated from the three one-million-sample SRS sets. The average and standard deviation of the estimates are also shown in the table.

Table 1
Calculated response statistics for reference values (10^6 samples, Bi-normal JPDF, SRS)

	Response statistic			
Realization	$\hat{\mu}_r$	$\hat{\sigma}_r$	$\hat{P}_{0.2}$	$\hat{P}_{0.5}$
1	0.511872	0.162834	0.984429	0.448457
2	0.511824	0.162733	0.984585	0.447915
3	0.511940	0.162737	0.984511	0.449029
Average	0.511879	0.162768	0.984508	0.448467
Std. dev.	5.829E−05	5.719E−05	7.803E−05	0.000557

3.2.1. Mean of response

The average of the three 10^6-sample estimates of mean response is taken as the reference value, $\hat{\mu}_{\text{ref}} = 0.511879$ from Table 1. Empirical CIs on this reference mean are obtained by substituting the standard deviation of the estimates, $\hat{\sigma}_{\text{est}} = 5.829\text{E}-05$ from Table 1, into Eq. (5). Thus, empirical 95% half-CI are 0.000145. When the reference mean is calculated based on the entire population of $N = 3\times10^6$ samples, the value does not change from the averaged value based on three separate 10^6-sample sets, but the classical CI can be computed. The classical 95% half-CI from standard statistical formulas is somewhat larger, at 0.000185. Using the larger (classical) CI here to be conservative, we say that with at least 95% certainty the true response mean μ lies within the range $\hat{\mu}_{\text{ref}} \pm 0.000185 = (0.511694, 0.512064)$. The CI range ±0.000185 is typically very small compared to the differences ($\hat{\mu}_r$ error) listed in Tables 2 and 3 between $\hat{\mu}_{\text{ref}}$ and the estimates of mean response from the 100-sample sets.

We take the differences from $\hat{\mu}_{\text{ref}}$ in Tables 2 and 3 as nominal measures of the error of the estimates produced from the 100-sample sets. For SRS, LHS, CVT, and LCVT

Table 2
Calculated response means (100 samples, Normal 2-D JPDF)

	SRS		LHS		Latinized CVT	
Realization	$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error
1	0.49980	−0.01208	0.51466	+0.00278	0.51238	+0.00050
2	0.51058	−0.00130	0.51110	−0.00079	0.50780	−0.00408
3	0.50315	−0.00873	0.50752	−0.00436	0.51182	−5.867E−05
Average	0.50451	−0.00737	0.51109	−0.00079	0.51067	−0.00121
Std. dev.	0.00552	0.00552	0.00357	0.00357	0.00250	0.00250
Avg. abs. error		0.00737		0.00264		0.00155
		Rank 6		Rank 4		Rank 1

Table 3
Calculated response means (100 samples, Normal 2-D JPDF)

	CVT		Halton		Hammersley	
Realization	$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error
1	0.50868	−0.00320	0.50565	−0.00623	0.51029	−0.00159
2	0.51148	−0.00040				
3	0.50840	−0.00348				
Average	0.50952	−0.00236	0.50565	−0.00623	0.51029	0.00159
Std. dev.	0.00170	0.00170				
Avg. abs. error		0.00236		0.00623		0.00159
		Rank 3		Rank 5		Rank 2

methods there is no unique 100-sample set. The SRS and LHS sets depend on the initial seed and the particular pairing of the 0–1 random variates on the $p1$ and $p2$ axes. The CVT and LCVT sets depend somewhat on the initial condition (particular uniform sample set started from), though fairly insensitively if enough iterations are performed to stabilize the guiding uniformity measures (see [1]). We therefore use three instantiations of SRS, LHS, CVT, and LCVT sets to begin to obtain a representative picture of the errors we might expect from each of these types of sampling. For each of these methods we average the individual absolute errors from the three instantiations to determine an average magnitude of error. This measure reflects contributions from both the average signed error of the three estimates (method bias or "accuracy"), as well as the variance of the three results (method repeatability or "precision"). This error measure is zero only if both the average signed error is zero and the variance of the estimates is zero. Furthermore, this error measure applies as well to the Halton and Hammersley results which consist of only one instantiation because they are deterministic sampling methods.

To also obtain a broad picture of each method's sampling efficacy across the different types of statistics calculated, we use an integer ranking score for method accuracy for each of the various calculated statistics (response mean, variance, and exceedence probabilities). Rank 1 indicates the method was the most accurate and therefore ranked first in performance. Rank 6 indicates the method was the least accurate among the sampling schemes tried. This allows us to compare method performance across the different types of statistics calculated (see Section 3.3). This is perhaps somewhat more satisfying than the piecemeal comparisons in, e.g., [13,19] that fail to give an explicit impression (quantitative balanced indicator) of the overall performance of the various sampling methods across a matrix of test problems. The accuracy ranking of each method according to average magnitude of error is given on the final lines in Tables 2 and 3.

3.2.2. Standard deviation of response

Tables 4 and 5 show the estimates of the standard deviation of response. Nominal errors from the reference value $\hat{\sigma}_{\text{ref}} = 0.162768$ are also shown. This value is the average of the three standard deviations in Table 1 calculated from the three 10^6 SRS sets. The standard deviation of these three estimates is $\hat{\sigma}_{\text{est}} = 5.719\text{E}-05$. Empirical 95% half-CI by Eq. (5) are 0.000142. Therefore we say that with 95% confidence the true response standard deviation σ lies within the range $\hat{\sigma}_{\text{ref}} \pm 0.000142 = (0.162626, 0.162910)$. The CI are negligibly small compared to the average absolute error magnitudes of the 100-sample estimates in Tables 4 and 5.

3.2.3. Exceedence probability for response threshold $r_{\text{T}} = 0.2$

Tables 6 and 7 show the estimates of the exceedence probability (EP) corresponding to a response threshold

Table 4
Calculated response standard deviations (100 samples, Normal 2-D JPDF)

	SRS		LHS		Latinized CVT	
Realization	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error
1	0.16874	+0.00597	0.18054	+0.01777	0.15782	−0.00495
2	0.16265	−0.00012	0.15570	−0.00707	0.15532	−0.00745
3	0.15699	−0.00578	0.14191	−0.02086	0.15554	−0.00723
Average	0.16279	2.533E−05	0.15938	−0.00339	0.15623	−0.00654
Std. dev.	0.00588	0.00588	0.01958	0.01958	0.00138	0.00138
Avg. abs. error		0.00396		0.01523		0.00654
		Rank 2		Rank 6		Rank 3

Table 5
Calculated response standard deviations (100 samples, Normal 2-D JPDF)

	CVT		Halton		Hammersley	
Realization	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error
1	0.14804	−0.01473	0.15415	−0.00862	0.16077	−0.00200
2	0.14906	−0.01371				
3	0.14615	−0.01662				
Average	0.14775	−0.01502	0.15415	−0.00862	0.16077	−0.00200
Std. dev.	0.00148	0.00148				
Avg. abs. error		0.01502		0.00862		0.00200
		Rank 5		Rank 4		Rank 1

level of $r_T = 0.2$. Nominal errors from the reference value $\hat{P}_{0.2,\mathrm{ref}} = 0.984508$ are also shown. This value is the average of the three EPs in Table 1 calculated from the three 10^6 SRS sets. The standard deviation of these three estimates is $\hat{\sigma}_{\mathrm{est}} = 7.803\mathrm{E}-05$. Empirical 95% half-CI by Eq. (5) are 0.000194. When the reference EP is calculated based on the entire population of $N = 3 \times 10^6$ samples, the value does not change from the averaged value based on three separate 10^6-sample sets, but classical CI can be computed. The classical 95% half-CI from standard statistical formulas is somewhat smaller, at 0.000140. Using the larger (empirical) 95% half-CI for conservatism, we say that to 95% confidence the true probability $P_{0.2}$ of response exceeding the threshold value $r_T = 0.2$ lies within the range $\hat{P}_{0.2,\mathrm{ref}} \pm 0.000194 = (0.984314, 0.984702)$. The CI are negligibly small compared to the error magnitudes of the 100-sample estimates in Tables 6 and 7. We note that the SRS and LHS results are both ranked at 4.5 because they both have the same error magnitude and together-speak for the 4th and 5th ranks.

3.2.4. Exceedence probability for response threshold $r_T = 0.5$

Tables 8 and 9 show the estimates of the exceedence probability corresponding to a response threshold of $r_T = 0.5$. Nominal errors from the reference value $\hat{P}_{0.5,\mathrm{ref}} = 0.448467$ are also shown. This value is the average of the three EPs in Table 1 calculated from the three 10^6 SRS sets. The standard deviation of these three estimates is $\hat{\sigma}_{\mathrm{est}} = 0.000557$. Empirical 95% half-CI by Eq. (5) are 0.001384. When the reference EP is calculated based on the entire population of $N = 3 \times 10^6$ samples, the value does not change from the averaged value based on three separate 10^6-sample sets, but classical CI can be computed. The classical 95% half-CI from standard statistical formulas is considerably smaller, at 0.000563. Using the larger (empirical) 95% half-CI for conservatism, we say that to 95% confidence the true probability $P_{0.5}$ of response exceeding the threshold value $r_T = 0.5$ lies within the range $\hat{P}_{0.5,\mathrm{ref}} \pm 0.001384 = (0.447083, 0.449851)$. The CI are negligibly small compared to the average absolute error magnitudes of the 100-sample estimates in Tables 8 and 9.

3.3. Weighted measure of statistical sampling merit

The performance rankings for the sampling schemes and statistical quantities tested are summarized in Table 10. The last column contains a normalized weighted figure of merit which is a broad measure of each method's sampling performance across the different types of statistics calculated. This figure of merit is obtained by first averaging the rankings for the (two) exceedence probabilities calculated, and then averaging this rank for EPs with the ranks for the mean and standard deviation calculations. These averages are then divided by the number of sampling methods involved. Hence, the normalized ranks in this column add up to unity. The bar chart in Fig. 10 helps visually assess the relative performance of the sampling

Table 6
Calculated response exceedence probabilities, threshold = 0.2 (100 samples, Normal 2-D JPDF)

	SRS		LHS		Latinized CVT	
Realization	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error
1	0.97	−0.01451	0.98	−0.00451	0.99	+0.00549
2	0.99	+0.00549	0.99	+0.00549	0.98	−0.00451
3	0.99	+0.00549	1.00	+0.01549	0.98	−0.00451
Average	0.98333	−0.00117	0.99	+0.00549	0.98333	−0.00118
Std. dev.	0.01155	0.01155	0.01	0.01	0.00577	0.00577
Avg. abs. error		0.00850		0.00850		0.00484
		Rank 4.5		Rank 4.5		Rank 2

Table 7
Calculated response exceedence probabilities, threshold = 0.2 (100 samples, Normal 2-D JPDF)

	CVT		Halton		Hammersley	
Realization	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error
1	0.99	+0.00549	0.97	−0.01451	0.98	−0.00451
2	0.99	+0.00549				
3	0.99	+0.00549				
Average	0.99	+0.00549	0.97	−0.01451	0.98	−0.00451
Std. dev.	0.0	0.0				
Avg. abs. error		0.00549		0.01451		0.00451
		Rank 3		Rank 6		Rank 1

Table 8
Calculated response exceedence probabilities, threshold = 0.5 (100 samples, Normal 2-D JPDF)

	SRS		LHS		Latinized CVT	
Realization	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error
1	0.43	−0.01847	0.44	−0.00847	0.46	+0.01153
2	0.41	−0.03847	0.45	+0.00153	0.43	−0.01847
3	0.43	−0.01847	0.45	−0.00847	0.47	+0.02153
Average	0.42333	−0.02513	0.44667	−0.0018	0.45333	+0.00487
Std. dev.	0.01155	0.01155	0.00577	0.00577	0.02082	0.02082
Avg. abs. error		0.02513		0.00384		0.01718
		Rank 6		Rank 1		Rank 5

Table 9
Calculated response exceedence probabilities, threshold = 0.5 (100 samples, Normal 2-D JPDF)

	CVT		Halton		Hammersley	
Realization	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error
1	0.44	−0.00847	0.46	+0.01153	0.44	−0.00847
2	0.46	+0.01153				
3	0.46	+0.01153				
Average	0.45333	+0.00487	0.46	+0.01153	0.44	−0.00847
Std. dev.	0.01155	0.01155				
Avg. abs. error		0.01051		0.01153		0.00847
		Rank 3		Rank 4		Rank 2

Table 10
Sampling method accuracy rankings for various calculated statistics of response

Sampling method	Response mean	Response standard deviation	Exceedence probability (0.2 threshold)	Exceedence probability (0.5 threshold)	Normalized weighted average
SRS	6	2	4.5	6	0.21
LHS	4	6	4.5	1	0.20
CVT	3	5	3	3	0.17
LCVT	1	3	2	5	0.12
Halton	5	4	6	4	0.22
Hammersley	2	1	1	2	0.07

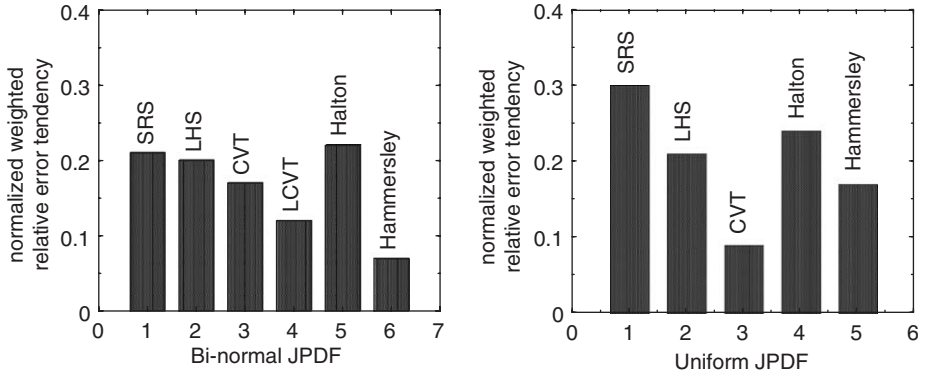

Fig. 10. Normalized weighted measure of sampling method relative error tendency in calculated mean, standard deviation, and $r_T = 0.2$ and 0.5 exceedence probabilities for bi-normal and uniform joint probability sampling densities in 2-D test problems.

methods according to our normalized figure of merit. The shorter the bar, the better the particular method ranks on balance across all the statistical quantities calculated. We see that Hammersley sampling ranked overall best on this series of non-uniform JPDF test problems, then LCVT, CVT, LHS, SRS, and finally Halton.

A second bar chart corresponding to the investigation in [19] is plotted in Fig. 10. The investigation was similar to the one in this paper, but compared calculated statistics based on a uniform JPDF, and did not include the LCVT sampling method. Since there is no mapping here from uniform sets to non-uniform JPDFs, only volumetric uniformity matters here and discrepancy properties are immaterial. Since pure CVT is more volumetrically uniform than LCVT, and for that matter, more volumetrically uniform than all the other sampling methods we have tested, CVT would be expected to generally rank best. This is indeed the case, as shown in the figure.

4. Discussion and conclusion

According to our weighted figure of merit, Hammersley sampling strongly ranked overall best on the set of bi-normal JPDF test problems in this paper, then LCVT, CVT, LHS, SRS, and finally Halton. Furthermore, in [12], Hammersley was found to be significantly more efficient than SRS and LHS for resolving mean and standard deviation of response over a large set of test problems. For resolving response probabilities, Hammersley and modified-Halton were found in [13] to perform roughly the same as LHS on balance over several test problems.

Hence, Hammersley is consistently the best performer or among the top performers in these empirical studies. However, Hammersley does not allow incremental (one at a time) addition of samples to the parameter space, which decreases its efficiency prospects relative to SRS and modified-Halton for sampling procedures involving error estimation and control (see e.g., [11]). Furthermore, when the number of random inputs grows beyond 10 or so dimensions and/or the sampling density in the hypercube becomes high, Hammersley might suffer from the spurious correlation effects that can plague other sub-random sequence methods. This is shown, e.g., for standard Halton in 16-D (Ref. [5]) and 40-D (Ref. [13]). This is something the authors need to further investigate—the answer may already exist in the literature.

Next best were Latinized CVT and then pure CVT, both beating the LHS, SRS, and Halton methods. However, more than three instantiations of SRS, LHS, CVT, and LCVT point sets are needed to more reliably reflect the true performance tendencies of these methods on our test problems. Also, sample sets of much larger size than 100 would be valuable particularly to get another significant digit of resolution in the calculated exceedence probabilities in the study. Moreover, our results are somewhat tied to the specific figure of merit employed in this study. This figure of merit has the advantage that it allows comparison of merit across different types of statistics calculated and

different problem sets, but other better measures may exist for our purposes. Although the weighted metric does not reveal method performance in the individual categories of response mean, standard deviation, and exceedence probability, these can be found in Table 10 for the problems in this paper.

Of course, empirical studies are only point glimpses of the relative accuracy tendencies of one method over another under a very specific set of conditions. Certainly, much more empirical work needs to be performed to assess the performance of CVT and LCVT versus other sampling methods over a diverse problem space, but even more valuable would be more theoretical work to ascertain which method might be expected to perform best under given conditions (the characteristics of the function involved; the number of input random variables/dimensions; character of the JPDF, etc.).

This being said, we have seen strong early empirical indications of the superiority of CVT for *uniform* JPDF problems. In [19], CVT strongly ranked overall best as expected, then Hammersley, LHS, Halton, and finally SRS. In particular, for statistical integration of functions (which involves uniform sampling over the integration domain), CVT appears to be the natural best choice theoretically, as corroborated by findings in [19]. Also, in point placement for response surfaces, CVT appears very promising relative to other structured and unstructured sampling methods (see [7]). Already, for irregular (non-hypercube) interpolation and integration domains, the uniformity of CVT sampling over the domain gives it a well recognized status in the application of 2-D and 3-D meshless finite-element methods [17]. Hence, when *volumetric* sampling uniformity is desirable, early indications are that CVT performs very well versus other sampling methods.

The non-incremental nature of CVT sampling, however, does decrease its potential efficiency advantage relative to SRS and modified-Halton for sampling procedures involving error estimation and control (see [11]). To reiterate, much more empirical and theoretical work remains to be done to broadly assess and characterize the efficacy of CVT and LCVT versus other sampling methods for various sampling tasks.

Acknowledgements

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the US Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000. Gunzburger was supported in part by the Computer Science Research Institute at Sandia National Laboratories, under contract 18407.

References

[1] Burkardt J, Gunzburger M, Saka Y. Latinized, improved LHS, and CVT point sets in hypercubes. IEEE Trans Inf Theory, submitted.

[2] Conover WM. On a better method for selecting input variables. In: Helton JC, Davis FJ, editors, 1975 unpublished Los Alamos National Laboratory manuscript, reproduced as Appendix A of "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems". Sandia National Laboratories report SAND2001-0417, printed November 2002.

[3] McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;21(2).

[4] Helton JC, Davis FJ. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab Eng Syst Saf 2003;81(1).

[5] Manteufel RD. Distributed hypercube sampling algorithm. In: Third AIAA non-deterministic approaches forum paper AIAA-2001-1673, 42nd structures, structural dynamics, and materials conference, April 16–19, 2001, Seattle, WA.

[6] Beachkofski BK, Grandhi RV. Improved distributed hypercube sampling. In: Fourth AIAA non-deterministic approaches forum paper AIAA-2002-1274, 43rd structures, structural dynamics, and materials conference, April 22–25, 2002, Denver, CO.

[7] Romero VJ, Burkardt, JS, Gunzburger MD, Peterson JS, Krishnamurthy T. Initial application and evaluation of a promising new sampling method for response surface generation: centroidal Voronoi tessellation. In: Fifth AIAA non-deterministic approaches forum paper AIAA-2003-2008, 44th structures, structural dynamics, and materials conference, April 7–10, 2003, Norfolk, VA.

[8] Ye KQ, Li W, Sudjianto A. Algorithmic construction of optimal symmetric latin hypercube designs. J Stat Plann Inference 2000;90:145–59.

[9] Hanson, K. Quasi-Monte Carlo: halftoning in high dimensions? In: Bouman CA, Stevenson RL, editors. Proceedings of SPIE 5016, computational imaging, 2003.p.161–72.

[10] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in Fortran: the art of scientific computing. 2nd ed. Cambridge: Cambridge University Press; 1992.

[11] Romero V, Slepoy R, Swiler L, Giunta A, Krishnamurthy T. Error estimation approaches for progressive response surfaces-more results. In: Proceedings paper # 114 of the 2006 international modal analysis conference (IMAC XXIV), January 30–February 2, 2006, St. Louis, MO.

[12] Kalagnanam JR, Diwekar UM. An efficient sampling technique for off-line quality control. Technometrics 1997;39(3).

[13] Robinson D, Atcitty D. Comparison of quasi- and pseudo-Monte Carlo sampling for reliability and uncertainty analysis. In: First AIAA non-deterministic approaches forum paper AIAA-1999-1589, 40th structures, structural dynamics, and materials conference, April 12–15, St. Louis, MO.

[14] Ju L, Du Q, Gunzburger M. Probabilistic algorithms for centroidal Voronoi tessellations and their parallel implementation. Parallel Comput 2002;28:1477–500.

[15] Okabe A, Boots B, Sugihara K, Chui S. Spatial tessellations: concepts and applications of Voronoi diagrams. 2nd ed. Chichester: Wiley; 2000.

[16] Du Q, Faber V, Gunzburger M. Centroidal Voronoi tessellations: applications and algorithms. SIAM Rev 1999;41:637–76.

[17] Burkardt J, Gunzburger M, Nugyen H, Saka Y. Constrained centroidal Voronoi tessellations for meshing. SIAM J Sci Comput, submitted.

[18] Romero VJ, Bankston SD. Progressive-lattice-sampling methodology and application to benchmark probability quantification problems, SAND98-0567. Albuquerque, NM: Sandia National Laboratories; 1998.

[19] Romero VJ, Burkardt JS, Gunzburger MD, Peterson JS. Initial evaluation of centroidal Voronoi tessellation method for statistical sampling and function integration. Sandia National Laboratories library archive SAND2003-3672C, an extended version (with comparison of exceedence probabilities) of the paper in the

proceedings of the fourth international symposium on uncertainty modeling and analysis (ISUMA'03), September 21–24, 2003, University of Maryland, College Park, MA.

[20] Iman RL, Shortencarier MJ. A FORTRAN77 program and user's guide for the generation of latin hypercube and random samples to use with computer models. SAND83-2365 (RG). Albuquerque, NM: Sandia National Laboratories; 1984.

[21] Burkardt JV, Gunzburger MD, Peterson JS. Research sampling software, http://www.csit.fsu.edu/~burkardt/f_src/cvt_dataset/cvt_dataset.html

[22] Iman RL, Conover WJ. A distribution-free approach to inducing rank correlation among input variables. Commun Stat 1982;B11(3):311–34.

[23] Romero VJ. Effect of initial seed and number of samples on simple-random and latin-hypercube Monte Carlo probabilities—confidence interval considerations. In: Paper PMC2000-176, Eighth ASCE specialty conference on probabilistic mechanics and structural reliability, University of Notre Dame, IN, July 24–26, 2000.

[24] Romero VJ, Mellen AL. 2002, preliminary research findings not yet published. Contact: vjromer@sandia.gov