AIAA SciTech
4-8 January 2016, San Diego, California, USA
18th AIAA Non-Deterministic Approaches Conference

A Set of Test Problems and Results in Assessing Method Performance for Calculating Low Probabilities of Failure[1]

Vicente Romero[2], Laura Swiler, Mohamed Ebeida, Scott Mitchell
Sandia National Laboratories,[3] *Albuquerque, NM*

Abstract

Estimation of the probability of failure to meet critical safety, performance, or constraint requirements or goals is important in engineering design and safety analyses, and in other risk analysis and management pursuits in business, finance, economics, environmental management, etc. This paper presents an interim set of test problems and results in evaluating the cost and accuracy performance of some current methods for calculating failure probabilities of magnitudes 10^{-2} to 10^{-6} in various low to moderate dimensional (2D to 9D) test problems mostly taken from engineering applications.

I. Introduction

Estimation of the probability of failure to meet critical safety, performance, or constraint requirements or goals is important in engineering design and safety analyses, and in other risk analysis and management pursuits in business, finance, economics, environmental management, etc.

The uncertainty space in probability of failure (POF) problems can be sampled adaptively or non-adaptively to provide estimates of failure probability. Adaptive methods use feedback from the response samples in an attempt to guide further sampling to efficiently narrow-in on a POF estimate for a given problem. Non-adaptive methods autonomously sample the uncertainty space without response feedback. An advantage of non-adaptive methods is that the final sample sets are not adapted to a single POF problem, and therefore not biased toward good performance on that problem but potentially poor performance on related POF problems for other response thresholds or other output quantities of the model that may be of interest. For example, POF values may be desired for various potential (uncertain) failure threshold levels and for several response quantities from a physics model such as pressure, temperature, etc. and at multiple points in time and/or space. Thus, non-adaptive methods may be more cost effective when multiple related POF problems are involved. But non-adaptive methods are usually substantially less efficient (higher number of model evaluations/samples) in attaining similar accuracies as adaptive methods when a single POF quantity is involved. By including both adaptive and non-adaptive methods in the following study involving single POF quantities, we can assess cost-accuracy penalties that non-adaptive methods incur relative to adaptive methods when interested in only a single POF quantity.

The POF methods considered all have an element of stochasticity in their performance because they involve random sampling of various sorts started from initial seeds for random number generation (RNG). We therefore evaluate cost, accuracy, and stochastic variability of the methods' performance for several different starting seeds.

Section II introduces the POF estimation methods initially assessed in this study. Section III describes the initial test problems that the methods are evaluated on, and presents results from the performance study. Additional POF estimation methods and test problems will be included in future efforts. Section IV concludes with a brief summary and comparison of the methods' performance.

---

[1] This paper is a work of the United States Government and is not subject to copyright protection in the U.S.
[2] AIAA Senior Member, corresponding author: vjromer@sandia.gov
[3] Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

II. Probability-of-Failure Estimation Methods Evaluated

The methods initially analyzed in this study are: Efficient Global Reliability Analysis (EGRA), Latin Hypercube Sampling (LHS), and Gaussian Process surrogate models (GPs) fit to Latin Hypercube samples. These methods are described in greater detail below. EGRA is an adaptive method, while LHS and GPs are non-adaptive. DAKOTA [Adams et al.] implementations of these methods were used in the study.

A. Efficient Global Reliability Analysis (EGRA)

A foundational idea in reliability analysis is the identification of a specific response contour over the application space, called the limit state function. The limit state function separates the failure region from the non-failure region. A variety of local reliability method approaches exist such as the Mean-Value method and the First Order Reliability method (FORM) [Haldar & Mahadevan, 2000]. These approaches use local information and local optimization methods. Local reliability methods can be very efficient (scaling especially well for high-dimensional UQ problems) on suitable POF problems where model response is not highly nonlinear over the uncertainty space. But local methods can perform poorly (non-robust, inaccurate, inefficient) when model response is highly nonlinear.

The global reliability analysis method EGRA ([Bichon et al., 2008]) was developed to overcome some of the limitations of local reliability methods. EGRA estimates system response using a GP surrogate model (see section II.C below) based on a relatively limited number of simulations, and calculates the performance characteristic of interest by sampling the GP surrogate model instead of sampling with the expensive simulation model. Starting from a GP built initially from a very small number of LHS (see section II.C) random samples, EGRA adaptively chooses where to generate subsequent samples in an attempt to increase the emulator accuracy of the GP in the vicinity of the failure boundary. The resulting GP is then sampled using multimodal adaptive importance sampling [Srinivasan, 2008; Denny, 2001; Richard & Zhang, 2007] to calculate the probability of failure. By locating multiple points on or near the failure boundary, complex and highly nonlinear failure boundaries can be modeled, allowing a more accurate POF estimate. Because EGRA concentrates samples in the vicinity of the failure boundary where accuracy is important, it is relatively efficient in number of samples required for a given accuracy.

B. Monte Carlo sampling of GPs built on Latin Hypercube Sampling

This method involves generating a specified number of Latin Hypercube samples [Conover, 1975] of the joint probability density of the input PDFs, then evaluating the model at the generated points in the uncertainty space. A GP surrogate is constructed from the sample points and response values. Gaussian process emulators (also called kriging models) are popular because they interpolate the data from which they are built; they provide a spatially varying estimate of the uncertainty of the emulation error between sampled values; and they do not require a specific type of sampling design. The Dakota GP formulation is fairly standard, with an exponential correlation function where the correlation lengths are calculated by maximum likelihood estimation. To obtain the POF estimate, the GP is evaluated by Monte Carlo (MC) sampling it with a large number of samples until the confidence interval (see section II.C) on the POF estimate is sufficiently small to get an acceptably precise estimate.

C. Latin Hypercube Sampling with Confidence Intervals on Results

Consider a system that in one or more portions of the input parameter "uncertainty" space exceeds a critical threshold level T above which the system is considered to fail. If the random-variable uncertainty space is randomly sampled via MC, then the number of system response values that exceed the threshold T, divided by the total number of MC samples, provides an estimate P^* of the true failure probability P of the system. If enough samples are taken, then the estimate P^* can be said with some percent likelihood or "confidence" to lie within a corresponding "confidence interval" of the true result P. From [Devore, 1982], when the number N of total MC samples meets the condition

$$N \bullet P \geq 5 \qquad (1)$$

then the following formula for 95% confidence intervals (CI) applies:

$$| P - P^* | \leq 1.96[P(1-P)/N]^{1/2}. \qquad (2)$$

When failure probability is being estimated by MC and the estimate P^* is used on the right side of equation (2) instead of the exact (presumably unknown) probability P, then only *approximate* 95% CI are obtained. They will not strictly hold with the advertised 95% reliability or confidence, but with still perform reasonably close to the advertised odds (see e.g. [Romero, 2000]). Additionally it was found that eqn. (2) gives conservative confidence intervals when LHS is used instead of standard MC. Hence, the estimated 95% CI are ventured to provide reasonable error estimates on MC and LHS results. We employ such CI estimates on the following test problems to assess the reliability of the CI estimates.

III. Test Problems and Method Performance on Them

In the following we present and evaluate cost, accuracy, and stochastic variability of the methods' POF performance for several different starting seeds for random number generation. The DAKOTA[1] implementations of POF-darts is presently restricted to problems with uniform PDFs for the input uncertainties. Therefore the following problems involve only uniform input PDFs.

2D Herbie Function

The D-dimensional Herbie test function [Lee et al., 2011] is

$$y_{Herbie}(\vec{x}) = \sum_{d=1}^{D}[\exp(-(x_d - 1)^2) + \exp(-0.8(x_d + 1)^2) - 0.05\sin(8(x_d + 0.1))]. \qquad (3)$$

Following [Dalby & Swiler, 2014] we consider the response or output PDF of the D=2 version of this function for uniform PDF inputs defined over the [-2, 2] square as shown in Figure 1.

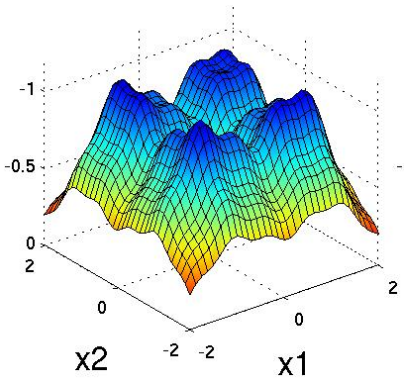

Figure 1. 2D Herbie test function over [-2, 2] square (*plotted upside-down for visualization purposes*).

Figure 2 shows two 2D Herbie test problems A and B with indicated regions where response is lower than the threshold levels specified in the figure. The 2D joint-uniform PDF integrated over the failure regions equals the failure probability in each case. For problem A there are five disjoint failure regions. For problem B there is only one very small failure region.

test problem A, threshold level = -1.065 test problem B, threshold level = -1.12656

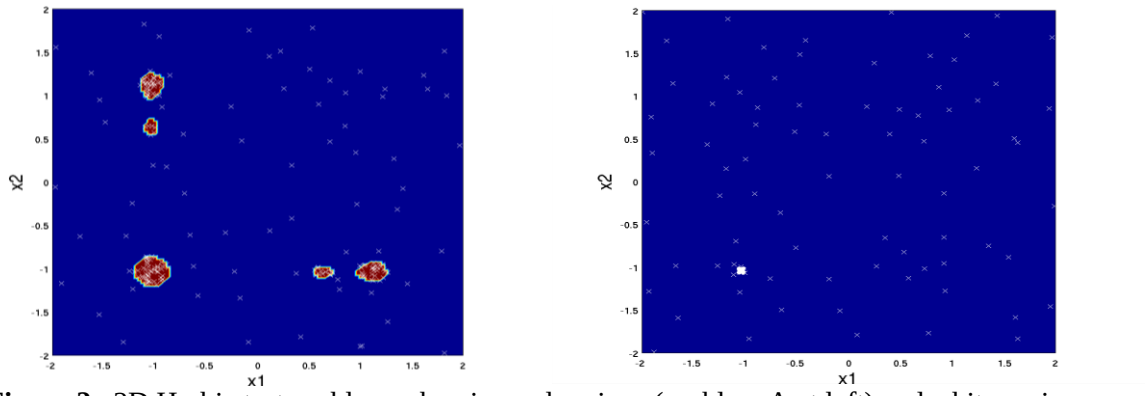

Figure 2. 2D Herbie test problems showing red regions (problem A at left) and white region (problem B at right) where response is lower than the specified threshold levels above the figures.

Table 1 lists failure probability estimates for problem A and all methods tried. The results in the table are plotted in Figure 3. With 95% reliability or confidence, the exact failure probability is deemed to lie within the stated 95% confidence intervals (CIs) about the mean estimate listed above the table. These CIs are calculated by the t-distribution method ([Iman, 1981; Helton et al., 1999]) per the note below the table. Note that these are CIs on the mean estimate (e.g. the mean probability of failure given three LHS replicates), in contrast to the Eqn. 2 CIs on the probability of failure estimates themselves. For problem A the CIs about the mean estimate are contained within the thickness of the black line in the plot. The CI results in the table for the LHS method are calculated with equation (2) and that row's calculated failure probability and number of samples.

Table 1. Estimated failure probabilities for 2D Herbie test problem A

Prob(y < -1.065) ≈ 1.50573E-2 ± 9.610E-5 = [1.4961E-2, 1.5153E-2]*

# of Samples	EGRA			Gaussian Processes			LHS w/conf. intvls.		
	seed A	seed B	seed C	seed A	seed B	seed C	seed A	seed B	seed C
25				4.078 E-3	0.	4.005 E-2			
55	1.301 E-2	2.257 E-2	8.480 E-3	8.976 E-3	2.989 E-2	2.567 E-2			
100				1.438 E-2	1.322 E-2	1.029 E-2			
200				1.479 E-2	1.381 E-2	1.493 E-2			
500				1.471 E-2	1.476 E-2	1.483 E-2	1.0E-2 ±8.7E-3	1.2E-2 ±9.5E-3	1.8E-2 ±1.2E-2
1000				1.476 E-2	1.476 E-2	1.476 E-2	1.3E-2 ±7.0E-3	1.7E-2 ±8.0E-3	1.2E-2 ±6.7E-3

* *95% confidence interval by t-distribution method* ([Iman, 1981; Helton et al.,1999]) *using three independent failure probability estimates based on 10^6 LH samples each*

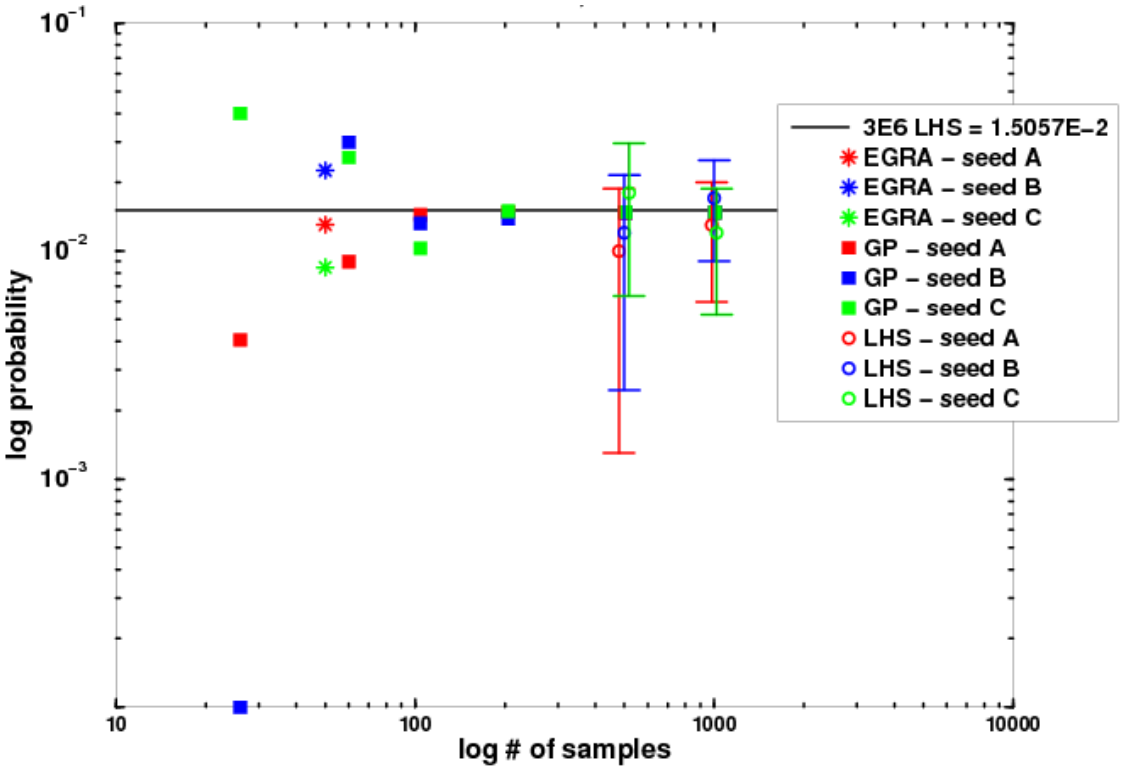

Figure 3. Failure probability estimates for Herbie 2D test problem A. (All results at the bottom of the plot indicate a value of 10^{-4} but are actually values of zero probability that could not be plotted on a log axis.)

Note that the results in Figure 3 and the other results plots below have the abscissa values of some of the results offset so that the GP, EGRA, etc. results do not lie on top of each other. This makes the methods' results distinct from each other so that easier visual comparisons can be made. The actual abscissa values for all methods are given in the tables of results.

From Figure 3 the following observations are made. EGRA "converged" after 55 samples or simulations (a.k.a. "function evaluations", FEVs) with the model for all RNG initial seeds A, B, and C, exhibiting significant seed dependence of individual results but small average absolute error, avg|error|. GPs with 55 FEVs performed almost as well, having slightly worse variability and avg|error| than EGRA. GPs improve in variance and avg|error| at 100 FEVs and has negligible variance and avg|error| for ≥ 200 FEVs. The reasonable performance of GPs with 55 FEVs begs the question of how well they do with fewer samples. With 25 FEVs it is seen that GPs are not reliable; high variability exists with seeds A, B, C, and large avg|error| exists. Finally, LHS shows non-negligible variance and avg|error| for point estimates with N = 500 and 1000 samples, but the confidence intervals are reliable for N ≥ 500 (N•pfail ≥ 5).

Table 2 lists failure probability estimates for problem B in Figure 2. The results in the table are plotted in Figure 4. The mean estimate and 95% confidence intervals regarding the exact failure probability listed above the table are calculated per the note below the table. The CIs in the table for the LHS method are calculated with equation (2) and that row's calculated failure probability and number of samples.

Table 2. Estimated failure probabilities for 2D Herbie test problem B

Prob(y < -1.12565) ≈ 1.0225E-4 ± 1.4890E-5 = [8.7360E-5, 1.1714E-4]*

# of Samples	EGRA			Gaussian Processes			LHS w/conf. intvls.		
	seed A	seed B	seed C	seed A	seed B	seed C	seed A	seed B	seed C
25				0.	0.	1.3920 E-2			
37			0.						
55	1.6329 E-4	0.		0.	7.80 E-4	4.4560 E-3			
100				2.4360 E-3	0.	0.			
200				0.	0.	7.300 E-5			
500				1.790 E-4	8.80 E-5	1.230 E-4			
1000				8.80 E-5	8.80 E-5	8.80 E-5			
5x10^4							1.0E-4 ±8.8E-5	1.2E-4 ±9.6E-5	4.0E-5 ±5.5E-5
10^5							1.2E-4 ±6.8E-5	1.2E-4 ±6.8E-5	1.5E-4 ±7.6E-5

** 95% confidence interval by t-distribution method [Iman, 1981; Helton et al.,1999] using four independent failure probability estimates based on 10^6 LH samples each*

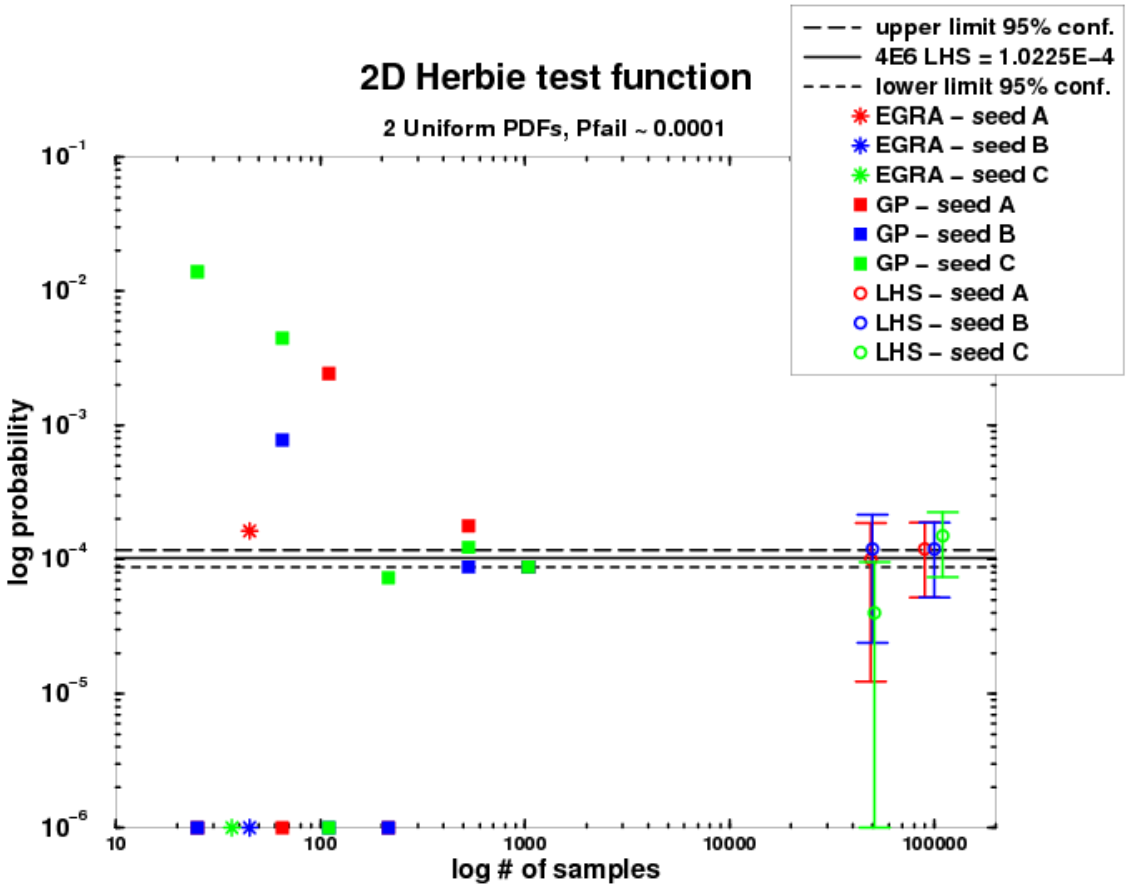

Figure 4. Failure probability estimates for Herbie 2D test problem B. (All results at the bottom of the plot indicate a value of 10^{-6} but are actually values of zero probability that could not be plotted on a log axis.)

The following observations are made. EGRA "converged" after 55, 55, and 37 FEVs for seeds A, B, and C, giving non-zero failure probability only for seed A with 55 FEVs. (All results at the bottom of the plot indicate a value of 10^{-6} but are actually values of zero that could not be plotted on the log axis.) Thus, significant seed dependence and premature convergence occurred with EGRA for this problem. The GP method is not reliable until $\geq$ 500 FEVs. LHS shows non-negligible variance and avg|error| for point estimates with $5x10^4$ and 10^5 samples, but confidence intervals are reasonably reliable for $N \geq 5x10^4$ ($N \bullet pfail \geq 5$).

2D Vibration Problem

The 2D vibration absorber problem taken from [Ramu, 2007; Acar & Ramu, 2014] has the following response function where uniform PDF inputs β1 and β2 are defined over the uncertainty ranges $0.9 \leq \beta 1, \beta 2 \leq 1.1$.

$$y(\beta_1,\beta_2) = \frac{|1-(1/\beta_2)^2|}{\sqrt{\left[1-\left(\frac{1}{\beta_2}\right)^2-\left(\frac{1}{\beta_1}\right)^2[R+1-(1/\beta_2)^2]\right]^2+4\gamma^2\left(\frac{1}{\beta_1}\right)^2\left[1-\left(\frac{1}{\beta_2}\right)^2\right]^2}} \quad (4)$$

Constants $R = \gamma = 0.01$ are the other parameter inputs to eqn. (4). Two disjoint partial "cactus-leaf" like failure regions exist in the 2D parameter space where response exceeds a certain threshold value. See e.g. [Acar & Ramu, 2014] for visualization of the irregularly shaped failure regions.

Table 3 lists failure probability estimates for a threshold level $T = 48$. The results in the table are plotted in Figure 5. The CIs about the mean estimate are contained within the thickness of the black line in the plot. The mean estimate of the exact failure probability and the 95% confidence intervals on the mean estimate and the CIs in the table for the LHS results are calculated via the procedures explained earlier.

Table 3. Estimated failure probabilities for 2D Vibration problem

Prob(y > 48) ≈ 1.9450E-2 ± 5.0995E-5 = [1.9397E-2, 1.9499E-2]*

# of Samples	EGRA			Gaussian Processes			LHS w/conf. intvls.		
	seed A	seed B	seed C	seed A	seed B	seed C	seed A	seed B	seed C
25				0.	1.0370 E-3	0.			
55	1.7304 E-2	1.6701 E-2	9.4082 E-3	1.4067 E-2	9.7260 E-3	9.4240 E-3			
100				1.3362 E-2	3.6640 E-3	7.6640 E-3			
200				5.3680 E-3	1.9557 E-2	1.5347 E-2			
500				1.8099 E-2	1.5354 E-2	1.6963 E-2	0.	0.	0.
1000				1.8870 E-2	1.9936 E-2	1.6892 E-2	1.60E-2 ±7.8E-3	2.50E-2 ±9.7E-3	1.60E-2 ±7.8E-3

* *95% confidence interval by t-distribution method using three independent fail probability estimates based on 10^6 LH samples each*

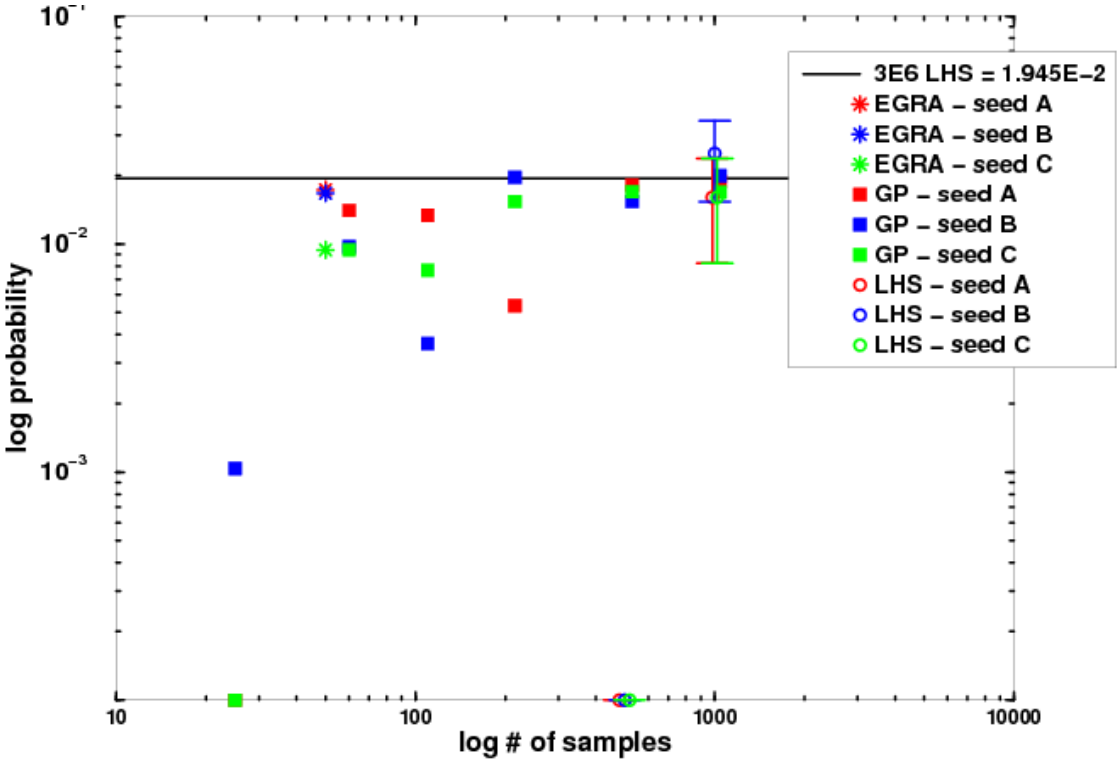

Figure 5. Failure probability estimates for 2D Vibration Amplitude problem. (All results at the bottom of the plot indicate a value of 10^{-4} but are actually values of zero that could not be plotted on a log axis.)

EGRA "converged" after 55 simulations with all seeds A, B, C, giving reasonable precision and accuracy—usually better than GPs. Because GPs gave results with 55 simulations that are less than an order of magnitude off the indicated true failure probability, it was decided to see how they'd do with 25 FEVs. GPs were not reliable with 25 simulations; some results are 0.0 probabilities. LHS is very inaccurate with 500 simulations (0.0 probabilities and 0.0 confidence intervals for all seeds A, B, C), but with 1000 FEVs LHS gives reasonable accuracy and precision of point estimates and reliable confidence intervals. 1000 FEVs equates to approximately N•pfail ≥ 20, so this particular combination of function, threshold level, and joint PDF did not meet the N•pfail ≥ 5 ventured criterion equation (1) for reasonable LH confidence intervals.

5D Electronic Circuit

A numerical model of a proprietary circuit with uniform PDFs for five inputs to the model is assessed next. Details of the circuit or model cannot be presented, but the performance of the POF methods on contrived failure probability problems can be presented. A threshold value that makes one of the output responses indicate a circuit failure of probability ~0.0001 was determined by iteration. Table 4 lists failure probability estimates. The results in the table are plotted in Figure 6. The mean estimate of the exact failure probability and the 95% confidence intervals on the mean estimate and the CIs in the table for the LHS results are calculated via the procedures explained earlier.

EGRA converged with 31 function evaluations and is most accurate and precise over seeds A, B, C. GPs with 31 simulations do not perform well, having low accuracy and repeatability. But good precision and accuracy are obtained at 100 FEVs. Precision and average accuracy then decline with more simulations. LHS gives accurate point estimates and reliable, fairly small confidence intervals for all seeds A, B, C with orders of magnitude more simulations, $5x10^4$ and 10^5 (N•pfail ≥ 5).

Table 4. Estimated failure probabilities for 5D Circuit problem

Pfail ≈ 1.6475E-4 ± 2.81E-5 = [1.370E-4, 1.930E-4]*

# of Samples	EGRA			Gaussian Processes			LHS w/conf. intvls.		
	seed A	seed B	seed C	seed A	seed B	seed C	seed A	seed B	seed C
31	1.6212 E-4	1.6286 E-4	1.5526 E-4	5.0 E-6	0.	2.70 E-5			
100				1.060 E-4	1.460 E-4	1.360 E-4			
200				1.540 E-4	8.10 E-5	6.40 E-5			
500				6.90 E-5	8.10 E-5	1.070 E-4			
1000				6.50 E-5	1.210 E-4	1.440 E-4			
5x10^4							2.40E-4 ±1.4E-4	1.60E-4 ±1.1E-4	1.40E-4 ±1.0E-4
10^5							1.50E-4 ±7.6E-5	2.05E-4 ±8.8E-5	1.60E-4 ±7.8E-5

* *95% confidence interval by t-distribution method using four independent fail probability estimates based on 10^6 LH samples each*

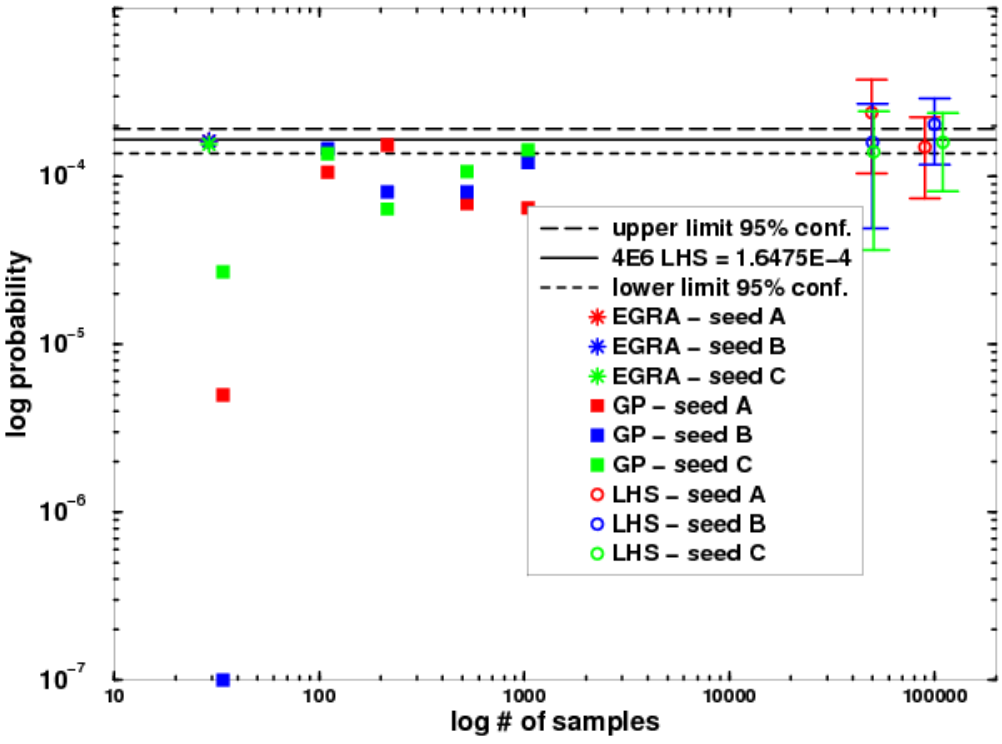

Figure 6. Failure probability estimates for 5D Circuit problem. (All results at the bottom of the plot indicate a value of 10^{-7} but are actually values of zero that could not be plotted on a log axis.)

9D Steel Column Problem

This problem involves determining the probability that the stress on a steel column will not meet a specified margin of safety relative to its yield stress *Fs*. The problem is loosely modeled after the one in [Kuschel & Rackwitz, 1997] and [Bichon, 2010]. The margin g is calculated from the following equations.

$$g = F_S - P\left(\frac{1}{2BT} + \frac{D_o}{BTH}\frac{E_b}{E_b - P}\right) \quad (5)$$

$$P = P_1 + P_2 + P_3 \quad (6)$$

$$E_b = \frac{\pi^2 EBTH^2}{2L^2}. \quad (7)$$

The current problem uses L = 7.5 m for the length of the column (deterministic) and nine input random variables (uniform PDFs) with upper and lower extents listed in Table 5.

Table 5. Uniform PDF Inputs for Steel Column problem

variable	*description*	*PDF lower extent*	*PDF upper extent*
Fs	yield stress, MPa	260	575
P_1	dead weight load, kN	250	650
P_2	variable load, kN	150	870
P_3	variable load, kN	150	870
B	flange breadth, mm	185	215
T	flange thickness, mm	11.5	27.5
H	profile height, mm	75	125
D_0	initial deflection, mm	-20	80
E	elastic modulus, MPa	1	41

A positive safety margin exists when the value of g is positive. We specify a desired safety margin of g=260 MPa and determine the probability prob($g \leq$ 260 MPa) that this margin is not met. The failure probability for this problem with the specified inputs is ~0.001. Table 6 lists failure probability estimates by the various methods. The results in the table are plotted in Figure 7. The mean estimate of the exact failure probability and the 95% confidence intervals on the mean estimate and the CIs in the table for the LHS results are calculated via the procedures explained earlier.

Table 6. Estimated failure probabilities for 9D Steel Column problem

Prob(g < 260) ≈ 1.035E-3 ± 2.4843E-5 = [1.0102E-3, 1.0598E-3]*

# of Samples	EGRA			Gaussian Processes			LHS w/conf. intvls.		
	seed A	seed B	seed C	seed A	seed B	seed C	seed A	seed B	seed C
10									
25				1.202 E-3	1.050 E-3	1.037 E-3			
50				1.052 E-3	1.086 E-3	1.072 E-3			
100				1.033 E-3	1.036 E-3	1.035 E-3			
108			1.0327 E-3						
114		1.0359 E-3							
142	1.0106 E-3								
200				1.040 E-3	1.044 E-3	1.041 E-3			
500				1.040 E-3	1.041 E-3	1.039 E-3			
10^4							1.20E-3 ±6.8E-4	1.10E-3 ±6.5E-4	1.20E-3 ±6.8E-4

* *95% confidence interval by t-distribution method using three independent fail probability estimates based on 10^6 LH samples each*

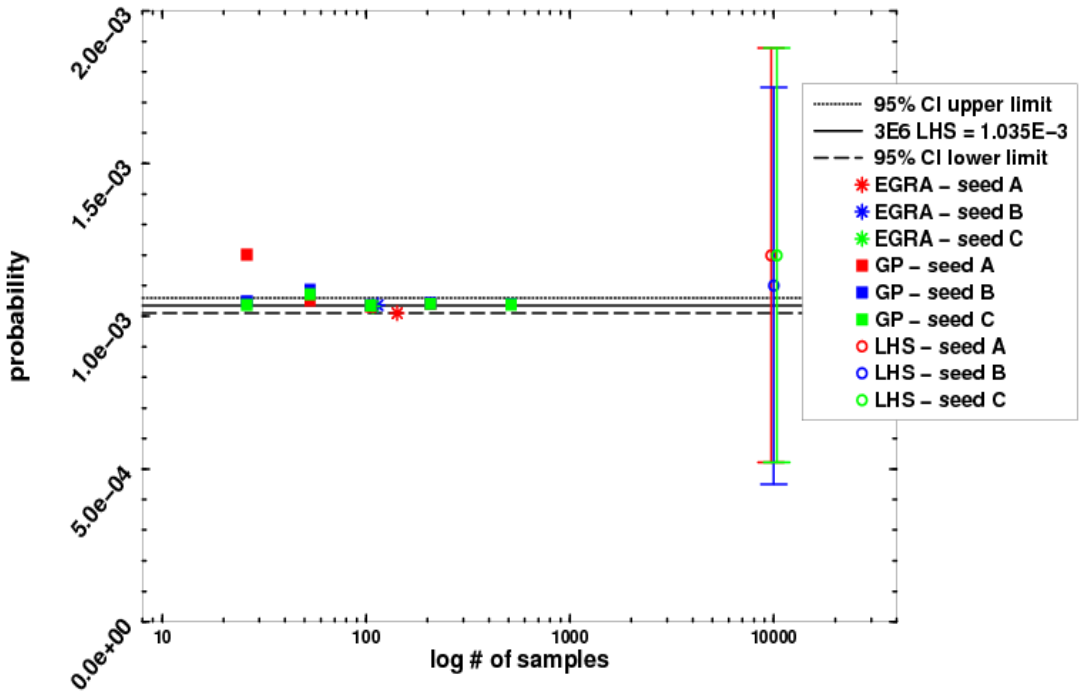

Figure 7. Failure probability estimates for 9D Steel Column problem.

GPs achieved reasonable accuracy and precision with as little as 25 samples. Reasonable performance with such few samples indicates that the function is probably only mildly nonlinear over the UQ space. GPs were not tried with 10 samples because the current implementation in DAKOTA does not allow GPs to be used with such few samples for a 9D problem. GPs achieved high accuracy and precision for ≥ 50 sims. EGRA required 142, 114, 108 FEVs to converge for seeds A, B, C respectively, giving high accuracy and precision even with the highly varying number of samples to convergence. LHS required two orders of magnitude more FEVs (10^4) for reliable 95% confidence intervals, so N•pfail ≥ 10 is required in contrast to the ventured rule of thumb equation (1).

IV. Comparative Performance of the Methods

For all methods the prediction errors were to the non-conservative side (prediction of smaller failure probabilities than actual) in over 2/3 of the trials for all test problems and numbers of samples.

The adaptive EGRA method was often the most cost effective method based on a combination of accuracy and cost considerations. But EGRA sometimes exhibited significantly premature convergence which led to undesirably large estimation error. Furthermore, EGRA produced estimates that could vary significantly with the starting seed and sample set used. The other methods, being stochastic estimation methods as well, also exhibited significant seed dependence.

The non-adaptive LHS/GP method was able to determine reasonably accurate failure estimates for most of the test problems with 100 function evaluations or fewer. The non-adaptive LHS/GP method generally appears to incur a significant accuracy penalty vs. the adaptive EGRA method for the same # of samples.

Given the significant seed dependence of all the methods it appears that for each method a multi-seed strategy may need to be developed that can take the multiple estimates and produce an appropriately reliable (say 90% reliable) error/uncertainty bar about the mean of the estimates. This will significantly decrease the economy of these methods but it appears they would still deliver substantial cost savings compared to reliable MC estimates with similar sized error/uncertainty bars (the LHS method was often one to two orders of magnitude more expensive to obtain accuracies comparable to the other stochastic methods). Empirically, the number of samples N for reliable error/uncertainty bars (confidence intervals) on LHS estimates ranged from 5 ≤ N•pfail ≤ 20.

IV. Closing

More POF methods, test problems, and analysis of relative cost-accuracy performance are presented in associated studies [Ebeida et al., 2015], [Swiler & Romero, 2015], [Romero et al., 2016].

Acknowledgement

This work was supported by the Laboratory Directed Research and Development (LDRD) and Accelerated Strategic Computing Verification and Validation (ASC V&V) programs at Sandia National Laboratories.

References

[1] Adams, B.M., Bohnhoff, W.J., Dalbey, K.R., Eddy, J.P., Eldred, M.S., Gay, D.M., Haskell, K., Hough, P.D., and Swiler, L.P., "DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 5.0 User's Manual," Sandia Technical Report SAND2010-2183, December 2009. Updated May, 2013 (version 5.3.1).

[2] Haldar, A. and S. Mahadevan, *Probability, Reliability, and Statistical Methods in Engineering Design*, John Wiley and Sons, 2000.

[3] Bichon, B., M. Eldred, L. Swiler, S. Mahadevan, J. McFarland, "Efficient Global Reliability Analysis for Nonlinear Implicit Performance Functions," *AIAA Journal* (2008) Vol. 46 no.10 (2459 2468).

[4] Srinivasan, R., *Importance Sampling*, Springer-Verlag, Berlin, 2002.

[5] Denny, M., "Introduction to importance sampling in rare-event simulations," *Eur. J. Phys.*, 22(4):401–411, 2001.

[6] Richard, J.-F. and Zhang, W., "Efficient high-dimensional importance sampling," *J. Econometrics*, 141(2):1385–1411, 2007.

[7] Conover, W. M, (1975) "On a better method for selecting input variables," unpublished Los Alamos National Laboratory manuscript, reproduced as Appendix A of "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems" by J.C. Helton and F.J. Davis, Sandia National Laboratories report SAND2001-0417 printed November 2002.

[8] Devore, J.L. (1982), *Probability & Statistics for Engineering and the Sciences*, Brooks/Cole Publishing Co., Wadsworth, Inc., Belmont, CA., pp. 99 - 104.

[9] Romero, V.J., "Effect of Initial Seed and Number of Samples on Simple-Random and Latin-Hypercube Monte Carlo Probabilities (Confidence Interval Considerations)," Proceedings of the 8th ASCE Joint Specialty Conference on Probabilistic Mechanics and Structural Reliability (PMC2000), University of Notre Dame, July 24-26, 2000.

[10] Lee, H., Gramacy, R., Linkletter, C., Gray, G., "Optimization subject to hidden constraints via statistical emulation," *Pacific Journal of Optimization*, 7:467–478, 2011.

[11] Dalbey, K. and L. P. Swiler. "Gaussian Process Adaptive Importance Sampling," *International Journal for Uncertainty Quantification*, Vol. 4(2). 2014. pp. 133-149. DOI: 10.1615/Int.J.UncertaintyQuantification.2013006330.

[12] Iman, R.L. "Statistical Methods for Including Uncertainties Associated with the Geologic Isolation of Radioactive Waste which allow for a Comparison with Licensing Criteria," *Proceedings of the Symposium on Uncertainties Associated with the Regulation of the Geologic Disposal of High-Level Radioactive Waste*, NUREG/CP-0022; CONF-810372, Eds. D.C. Kocher, March 9-13, 1981, Gatlinburg, Tennessee.

[13] Helton, J.C., D.R. Anderson, H.-N. Jow, M.G. Marietta, G. Basabilvazo (1999), "Performance Assessment in Support of the 1996 Compliance Certification Application for the Waste Isolation Pilot Plant," *Risk Analysis*, Vol. 19, No. 5.

[14] Ramu, P., "Multiple Tail Models Including Inverse Measures for Structural Design Under Uncertainties," 2007 PhD Thesis, U. of Florida, Gainesville, FL.

[15] Acar, E., Ramu, P., "Reliability Estimation Using Guided Tail Modeling with Adaptive Sampling", 16th AIAA Non-Deterministic Approaches Conference, SciTech 2014, Jan. 13-17, National Harbor, MD.

[16] Kuschel, N., and Rackwitz, R., "Two basic problems in reliability-based structural optimization." *Math. Method Operations Research*, Vol. 46, 1997, pp. 309-333.

[17] Bichon, B. J., 2010, "Efficient Surrogate Modeling for Reliability Analysis and Design," Ph.D thesis, Vanderbilt University, Nashville, TN.

[18] Swiler, L.P., and V.J. Romero, (2015) "A Survey of Advanced Probabilistic Uncertainty Propagation and Sensitivity Analysis Methods," Sandia National Laboratories document SAND2015-4494B and chapter accepted for Joint Army/Navy/NASA/Air Force (JANNAF) *Advances in UQ, V&V, and Simulation Credibility in Propulsion and Energetics* to be published April, 2016 by NASA.

[19] Ebeida, M.S., S.A. Mitchell, L.P. Swiler, V.J. Romero, A. Rushdi, (2015) "POF-Darts: Geometric Adaptive Sampling for Probability of Failure," submitted to *SIAM/ASA J. UNCERTAINTY QUANTIFICATION*.

[20] Romero, V.J., L.P. Swiler, M.S. Ebeida, S.A. Mitchell, M. Glickman, "Some Test Problems and Results in Assessing Method Performance for Calculating Low Probabilities of Failure," to be submitted to *AIAA Journal* in 2016.