51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference<BR> 18th
12 - 15 April 2010, Orlando, Florida

AIAA 2010-2850

AIAA-2010-2850

Importance Sampling: Promises and Limitations

Laura P. Swiler[1]
Sandia National Laboratories[2], Albuquerque NM 87185

Nicholas J. West[3]
Stanford University, Stanford CA 94305

Importance sampling is an unbiased sampling method used to sample random variables from different densities than originally defined. These importance sampling densities are constructed to pick "important" values of input random variables to improve the estimation of a statistical response of interest, such as a mean or probability of failure. Conceptually, importance sampling is very attractive: for example one wants to generate more samples in a failure region when estimating failure probabilities. In practice, however, importance sampling can be challenging to implement efficiently, especially in a general framework that will allow solutions for many classes of problems. We are interested in the promises and limitations of importance sampling as applied to computationally expensive finite element simulations which are treated as "black-box" codes. In this paper, we present a customized importance sampler that is meant to be used after an initial set of Latin Hypercube samples has been taken, to help refine a failure probability estimate. The importance sampling densities are constructed based on kernel density estimators. We examine importance sampling with respect to two main questions: is importance sampling efficient and accurate for situations where we can only afford small numbers of samples? And does importance sampling require the use of surrogate methods to generate a sufficient number of samples so that the importance sampling process does increase the accuracy of the failure probability estimate? We present various case studies to address these questions.

I. Introduction

Importance sampling is an unbiased sampling method used to sample random variables from different densities than originally defined. These importance sampling densities are constructed to pick "important" values of input random variables to improve the estimation of a statistical response of interest, such as a mean or probability of failure of a response from a simulation model. The use of input sampling densities will result in biased estimators if they are applied directly to the simulation results. However, the simulation results are weighted to correct for the use of the importance sampling densities, and this ensures that the importance sampling estimators are unbiased.

Conceptually, importance sampling is very attractive: for example one wants to generate more samples in a failure region when estimating failure probabilities. In practice, however, importance sampling can be challenging to implement efficiently, especially in a general framework that will allow solutions for many classes of problems. We are interested in the promises and limitations of importance sampling as applied to computationally expensive finite element simulations which are treated as "black-box" codes. One of the authors is a developer of the DAKOTA software framework [Eldred et al., 2006]. DAKOTA is a "wrapper" around simulation codes and is used to perform a variety of optimization and uncertainty quantification studies. One of the most commonly used uncertainty quantification methods is Latin Hybercube sampling (LHS), a stratified sampling technique which places samples in equi-probability bins throughout the space [Swiler and Wyss, 2004]. The DAKOTA team frequently gets requests

[1] Principal Member of Technical Staff, Optimization and Uncertainty Estimation Dept., P.O. Box 5800, Sandia National Laboratories, MS 1318.

[2] Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

[3] Doctoral Candidate, Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305.

from users to be able to perform additional samples “customized” to a region of interest after an initial LHS sampling is performed. That is, we get requests such as: “I have run an LHS sample of size 50 and I want to take 50 more samples that are more tailored to the failure region (or area of interest). How do I do this?” These questions provide the basic motivation for this paper. A major concern is that even if we are able to provide an importance sampling density that will preferentially sample in the failure region, the resulting failure probability estimate may not be significantly improved due to the small number of samples that are able to be performed.

In this paper, we present a customized importance sampler that is meant to be used after an initial set of Latin Hypercube samples has been taken, to help refine a failure probability estimate. The importance sampling densities are constructed based on kernel density estimators. We examine importance sampling with respect to two main questions: is importance sampling efficient and accurate for situations where we can only afford small numbers of samples? And does importance sampling require the use of surrogate methods to generate a sufficient number of samples so that the importance sampling process does increase the accuracy of the failure probability estimate? We present various case studies to address these questions. The paper is organized as follows: Section 2 provides background on importance sampling, including a discussion of current approaches; Section 3 discusses the importance sampling density based on kernel density estimators that we are using; Section 4 presents some results on various failure estimation problems; Section 5 discusses the use of surrogates in the process; and Section 6 provides a summary.

II. Importance Sampling Background

Accurate computation of high-dimensional integrals is common to many engineering and scientific applications. Monte Carlo methods have been commonly used for many years to approximation the expectation of functions of random variables by using the sample mean. That is, when calculating the expectation:

$$E(r(X)) = \int r(x) f_X(x) dx \qquad (1)$$

where r(X) is a response function: r: $\Re^d \rightarrow \Re^1$ and X is a multidimensional random variable having probability density function , the Monte Carlo estimator of E(r(X)) is:

$$\hat{\mathrm{E}}_{\mathrm{n}}(r(X)) = \frac{1}{n}\sum_{i=1}^{n} r(x_i) \qquad (2)$$

Note that many quantities of interest can be cast as expectations, including probabilities of failure. For calculation of failure probabilities, the probability that r(X) takes on a value in set A of interest (defined as the failure region) can be written as:

$$\mathrm{Prob}(r(X) \in \mathrm{A}) = \mathrm{E}[\mathrm{I}_{\{\mathrm{A}\}}(r(X))] \qquad (3)$$

where $\mathrm{I}_{\{\mathrm{A}\}} = 1$ when r(X)∈A and 0 when r(X)∉A.

The purpose of importance sampling is to sample the random variables from a different distribution than the original distribution of interest and use those samples to calculate an estimate $\hat{\mathrm{E}}_{\mathrm{n}}(r(X))$, with the goal of reducing the variance in the estimate. To do this, the Monte Carlo estimate must be weighted appropriately. If $h_X(x)$ is the new distribution[4] from which the random variables will be sampled (note h must have the same support as f), the new estimate is derived as follows:

$$\mathbf{E}(r(X)) = \int r(x)\frac{f(x)}{h(x)}h(x)dx = \hat{\mathbf{E}}_h\left[r(X)\frac{f(X)}{h(X)}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{r(x_i)f(x_i)}{h(x_i)}, x_i \sim h(X) \qquad (4)$$

Where $h(x)$ is called the importance sampling density, and the product of the integrand and the original density is called the importance function, and $f(x)/h(x)$ is called the weight function. There are certain choices that one should

[4] From here on, we drop the notation $f_X(x)$ and $h_X(x)$ when referring to density functions and simply use $f(x)$ and $h(x)$.

make when picking h(x). The variance of the importance sampling estimator $\hat{\mathrm{E}}_h$ is minimized when *h(x)* ∝ |*r(x)f(x)*|.

Much of the work in importance sampling has been finding minimum variance estimators for specific cases. The fundamental issue in implementing importance sampling is the choice of the new distribution which encourages the important regions of the input variables. If you have a good distribution, the payoff is a much lower simulation cost, but if you choose a poor importance sampling density, the run times could actually be longer than standard Monte Carlo simulation. There are various measures used to calculate the goodness of the importance sampling scheme. One is the ratio of the variance obtained by a pure Monte Carlo vs. the variance of the importance sampling result: $\sigma_{MC}^2 \big/ \sigma_{IS}^2$; another measure is the ratio of the number of samples required by each scheme, given the same output variance (this is called simulation gain): N_{MC}/N_{IS}. In general, a good importance sampling function should be as follows:

1. *h(x)* > 0 whenever *r(x)f(x)* ≠0.
2. *h(x)* should be close to being proportional to |*r(x)f(x)*|.
3. It should be relatively easy to generate samples from *h(x)* and also to calculate the density *h(x)* for particular values of x.

Some standard approaches to determining *h(x)* including scaling, where the original random variable X is scaled by α or linearly shifted (e.g. *h(x)* = *f(x-c)*) to put more probability density into a particular region (for failure probability estimation, for example). Another classical approach is to assume that *h(x)* belongs to a parametric distribution family. Then, the problem is determining that values of the parameters governing that distribution (for example, determining the mean and variance for *h(x)* if *h* is assumed to be normal). Often these parameters are obtained by optimizing the variance of importance sampling estimator $\hat{\mathrm{E}}_h$, but this implies that one can calculate an analytic expression or approximation for this estimator. Oh and Berger (1993) used an approach where they used a mixture distribution (a set of weighted individual distributions), where the individual distributions were multivariate-t distributions. They then had to determine the weights, location, and scale parameters of the *t*-distributions, which they did by numerical minimization of the estimate of the squared variation coefficient of the weight function.

For black-box simulations that have multiple uncertain inputs which may come from a wide variety of random input distribution types, we cannot generally assume that the importance sampling density will be normal or have a parametric form. Thus, we examined nonparametric methods. The use of nonparametric approaches in importance sampling is fairly recent. The first papers were mid-1990s, and include Bayesian approaches (e.g. Givens and Raftery) and kernel density estimators [Zhang]. The most common nonparametric approach is to use kernel density estimators for the importance sampling density. Zhang (1996) has a nice explanation of the tradeoffs of the increased convergence but higher computation of going to the nonparametric methods: "The most difficult part in parametric importance sampling is choosing a suitable distribution family to start with. There is no general recipe, and the issue remains largely a matter of art in the literature. Most parametric distributions fail to include *g* (the optimal importance sampling density) as a member. Consequently, parametric strategies yield estimates of the expectation that typically have mean squared error of the order $O(N^{-1})$. For a d-dimensional integral, we show in this article that the MSE of a nonparametric estimate is of the order $O(N^{-(d+8/d+4)})$ which nearly doubles the parametric rate when d is small. The faster convergence rate is achieved at the cost of increased computation." The computational cost from kernel density approaches comes from the fact that the number of density components is equal to the sample size, and governing parameters (e.g. bandwidth, weights) have to be determined for each component.

Bayesian implementation of importance sampling [Givens and Raftery] adds another layer of complexity, because there are priors on the parameters governing the importance sampling function *h(x).* In Bayesian analysis, the expectation usually has a normalizing factor in the denominator, and *f(x)* is a probability density which may be intractable. When *f(x)* is a posterior distribution from a Bayesian analysis, it may only be known up to a constant of proportionality and is written as:

$$\mathrm{E}(r(X)) = \frac{\int r(x)w(x)h(x)dx}{\int w(x)h(x)dx} = \frac{\hat{\mathrm{E}}_h[r(X)w(X)]}{\hat{\mathrm{E}}_h[w(X)]} = \sum_{i=1}^{n} r(x_i)v(x_i), v(x_i) = \frac{w(x_i)}{\sum_{i=1}^{n} w(x_i)} \qquad (5)$$

where *w(x)= f(x)/h(x).* We do not use Bayesian approaches in this paper, in part because of the additional Monte Carlo Markov Chain computation required to estimate the parameters governing the importance sampling density.

Many technical fields seem to have discovered importance sampling and have sets of papers/theory particular to that field. For example, there are applications in the statistical physics field, and an entirely different set of applications in the econometrics and mathematical finance field which do not reference each other although the needs are very similar. The papers we read tended to have simple examples with a one-dimensional random variable. Richard and Zhang state: "Importance samples have to be carefully tailored to the problem under consideration. This has proved to be a significant obstacle to routine applications of importance sampling….None of the existing importance sampling methods appear to be applicable to (very) high-dimensional applications." Comments such as these, along with the previous comment about the selection of an importance sampling density being an art, are somewhat discouraging in terms of developing a general framework which would allow users to subsequently refine failure probability estimates using importance sampling. However, there are many examples of successful uses of importance sampling in failure estimation. For example, Bichon and Mahadevan [Bichon et al. 2008, Dey and Mahadevan] use importance sampling in analytic reliability methods. These methods have a nice analytic form, where the inputs are transformed to standard normal space, so the importance sampling does have a natural parametric form (also standard normal, around points that have been identified as on or near the limit state function). In the analytic reliability methods, importance sampling has been very effective. Our goal is to see if we can develop something similar where failure probabilities are calculated based on sampling.

III. Importance sampling density via kernel density estimators

Kernel Density Estimation (KDE) [Rosenblatt, Parzen] is a technique used to estimate the density of a random variable *X* given *n* independent samples X_1, ..., X_n of it. If one considers the discrete distribution obtained from the sampled

$$F_n^{\{discrete\}}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < x) \qquad (6)$$

the KDE can be viewed as a smoothed version of this estimator. Let *K(.)* be a probability density function (*K*(.)≥0) and $\int K(x)dx = 1$ then the KDE is

$$f_n^{\{KDE\}}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) \qquad (7)$$

Here, *h* is known as the *bandwidth parameter* which controls the influence of each sample in providing a density estimate at a near-by point. Small *h* corresponds to a small region of influence; a large *h* to a large one. In the case of estimating a hitting probability we only want points in the target set (which may be defined by an indicator function $I(r(X_i) < c)$) to contribute to the approximated density so we use

$$f_n^{\{KDE\}}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} I(r(X_i) < c)K(\frac{x - X_i}{h}) \qquad (8)$$

following [Zhang,1996]. Integrating over all *x* we see that the above estimate gives us the initial Monte Carlo sample estimated probability.

Selection of the bandwidth parameter is the subject of a vast literature. Popular methods include cross validation (see [Silverman and Turlach]) and asymptotic analysis (see [Silverman and Wand]). The analysis used to derive the optimal bandwidth in the asymptotic case requires that the target density (in our case an indicator function) be twice differentiable. Although our response is an indicator function, we chose to use the optimal bandwidth obtained by minimizing the asymptotic mean-squared error of the importance sampling density. This approach produced good results. The optimal bandwidth was calculated according to Equation (5) in Zhang:

$$h_{opt} = \{\frac{dR(K_d)}{\sigma_K^4} AJ^{-1}\}^{1/(d+4)} n^{-1/(d+4)} \quad (9)$$

Where the full kernel is a product kernel of the marginal densities $K_d(\boldsymbol{x}) = K(x_1)...K(x_d)$, σ_K^2 is the variance of the marginal kernel function *K*, *R(K_d)* and *A* are given in Equation 10, and we assumed *J*=1. $\hat{I}_{init}$ refers to the initial estimate of the failure probability generated from the initial set of LHS samples, $\boldsymbol{x}_{init}$, which are used to seed the KDE process.

$$\begin{aligned} \sigma_K^2 &= \int x_1^2 K(x_1) dx_1 \\ R(K_d) &= \int K_d(\boldsymbol{x}) d\boldsymbol{x}; \\ A &= \frac{1}{n\hat{I}_{init}} \sum_{i=1}^{n} I(\boldsymbol{x_{init}}) f(\boldsymbol{x_{init}}) \end{aligned} \quad (10)$$

The kernel function, *K*, can, in theory be any probability density function. However, as we will use this density in importance sampling, we need it to have support over the entire domain. In many settings there is a large variety of usable kernels, however, most of these have finite support and are thus not suitable for use in importance sampling (leading to a biased estimator). We therefore pick the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (11)$$

which has infinite support. This has the effect of making the density a mixture of Gaussians, which will lead to an efficient implementation in the Importance Sampling algorithm.

Much of the literature on KDE addresses densities with infinite support (the random variables can take on any value on the real line). However, in our case, our input random variables are typically drawn from a bounded domain. To address this, we truncate each of the kernels at the boundary of the input random variables and scale the kernel by the mass that remains in the domain (forcing it to remain a probability density). In the case of a product kernel, this reduces to truncating each kernel in one dimension, and the product kernel remains a probability density.

Once we have fit the KDE to the observations, it remains to sample this density. We note that a KDE is a mixture of densities (the kernel centered around observation *i*) and thus we can sample from each of the kernels with uniform probability. In general, the kernel may not be easily sampled; in this case we use the Acceptance Rejection Algorithm to sample from the estimated density. At the moment, we use only a uniform density as the proposal distribution; more efficient choices may be found depending on the kernels used. When a (truncated) Gaussian kernel is used, the KDE is a mixture of Gaussians and can thus be efficiently sampled by uniformly sampling each Gaussian. We then estimate the failure probability via

$$\frac{1}{M} \sum_{i=1}^{M} I(r(X_i) < c) \frac{f(X_i)}{f_n^{\{KDE\}}(X_i)} \quad (12)$$

where the X_i are sampled from $f_n^{\{KDE\}}(X_i)$.

IV. Results for failure probability estimation

This section presents results using the kernel density estimator approach described in Section 3 to generate importance sampling densities. The output measures were failure probability estimates. As discussed in Section 2, we want to identify both advantages and limitations of importance sampling used in a particular way: we assume that an initial set of LHS samples is taken, a threshold on the response function is specified which defines a failure region (or region of interest), and we want to use importance sampling with approximately the same number of samples as were originally taken in the LHS sample to improve the failure probability estimate.

To demonstrate with a simple example, our first test involves the Rosenbrock function which has two inputs and is defined as $(x_2 – x_1^2)^2 + (1- x_1)^2$. The bounds on the input variables are: $-2 \leq x_1 \leq 2$ and $-2 \leq x_2 \leq 2$. We assumed that x_1 and x_2 are uniformly distributed between their bounds, and that they are independent. Note that this implies the probability density function of any point $X_i = \{x_{1i},x_{2i}\}$ is 1/16 if both x_1, and x_2 are within their bounds. Thus, $f(X_i) = 1/16$ if $-2 \leq x_1 \leq 2$ and $-2 \leq x_2 \leq 2$ and 0 elsewhere. Figure 1 shows what the Rosenbrock function looks like over this domain. Note that there is a large region near the center of this domain where the function is small and the function increases significantly near the boundaries. We chose the probability that the Rosenbrock function is less than 3 over these bounds to be the failure probability of interest. The area where the Rosenbrock function is less than 3 is a sliver shown in Figure 2 which plots the indicator function (the value 1 means the function is less than 3: note this region is NOT symmetric).

Table 1 shows the results for importance sampling on the Rosenbrock function, where Gaussian kernel density estimation is used to create the importance density. A number of initial LHS points is used to find a set of points satisfying the failure criteria. It is around these points that the density estimators are constructed. The number of importance sampling points listed in the table are the number of points sampled using the importance sampling framework to construct the revised probability of failure. We chose small numbers of (LHS,IS) points to realistically represent combinations that people working with computationally expensive models might be able to afford. We repeat the importance sampling process K times (where K = 100) to generate statistics on the importance sampling failure probability estimates, so all of the results in this table are based on 100 iterations of the importance sampling process. Note that the mean of the failure probability estimates from importance sampling are closer to the “true” estimate obtained by sampling 1 million points, however the standard deviation estimates tend to be large. Also, we report the 5th and 95th percentile estimate of the importance sampling failure probability as well as the percentage of points generated in the importance sampling process that “fail.” Note that in the example in Table 1, the importance sampling process generates approximately three times the number of failure points as from a standard LHS sampling process. The true failure probability is approximately 3.8% but approximately 12-14% of the points generated by the importance sampling density fail.

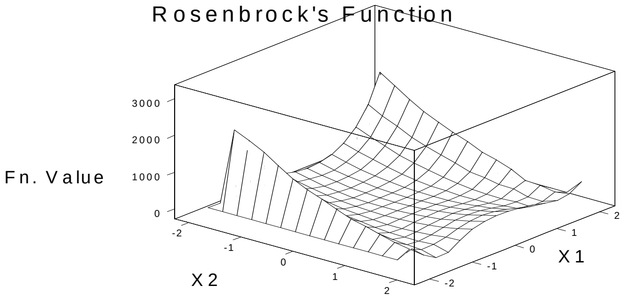

Figure 1. The Rosenbrock function

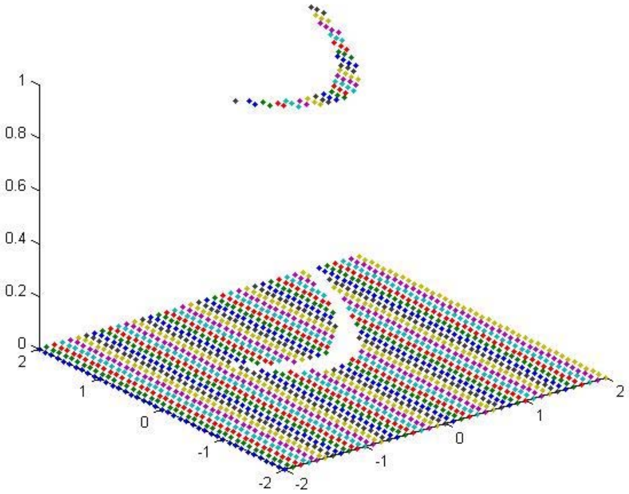

Figure 2. Failure region defined as (Rosenbrock function < 3)

Number of initial LHS samples	Number of Importance samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Mean Failure probability	IS Std dev. Failure Probability	5th percentile IS Failure Probability	95th percentile IS Failure Probability	Mean percentage of IS that "fail"
100	200	0.03900	0.01096	0.03798	0.01139	0.01904	0.05515	0.1283
50	100	0.04320	0.01429	0.03854	0.01665	0.01821	0.06369	0.1159
100	100	0.03600	0.01241	0.03986	0.01567	0.02085	0.06486	0.1330
200	200	0.03785	0.00955	0.03774	0.00823	0.02322	0.05454	0.1412
200	400	0.03702	0.00843	0.03868	0.00768	0.02768	0.05036	0.1432

Table 1. Failure estimation of (Rosenbrock function < 3). True value is 0.0383.

Table 2 shows the results for estimating a smaller failure probability (Rosenbrock function < 1). The true value of this is 0.014919. Note that the kernel density estimation process must have at least one point in the initial set of LHS samples that fails to start the process. As expected, increasing the number of initial LHS points and increasing the number of points taken via importance sampling tends to improve the failure probability estimate.

Number of Initial LHS samples	Number of Importance samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Mean Failure probability	IS Std dev. Failure Probability	5th percentile IS Failure Probability	95th percentile IS Failure Probability	Mean percentage of IS that "fail"
100	200	0.01677	0.00674	0.01494	0.00569	0.00680	0.02497	0.0616
50	100	0.01680	0.00784	0.01288	0.00775	0.00178	0.02443	0.0517
100	100	0.01785	0.00851	0.01506	0.00731	0.00466	0.02849	0.0636
200	200	0.01515	0.00607	0.01472	0.00634	0.00711	0.02236	0.0692
200	400	0.01638	0.00556	0.01377	0.00358	0.00854	0.01913	0.0635

Table 2. Failure estimation of (Rosenbrock function < 1). True value is 0.0149.

Table 3 shows the results for estimating the failure probability (Rosenbrock function < 3) when the input variables are defined as truncated normal distributions, not as uniform distributions. We chose truncated normals to represent the fact that parameters in computational models are often required to be constrained. The 2 input variables were truncated normal with zero mean and standard deviation of two, constrained to be in the interval [-2,2].

Table 3 shows that the standard deviation and the 5th/95th importance sampling failure estimates narrow as the number of initial LHS samples and number of importance samples increases. Additionally, the percentage of importance samples that are in the failed region increases as the number of input samples increases.

Number Of Initial LHS samples	Number Of Importance Samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Mean Failure probability	IS Std dev. Failure Probability	5th percentile IS Failure Probability	95th percentile IS Failure Probability	Mean percentage of IS that "fail"
100	200	0.04430	0.01175	0.04611	0.02673	0.02626	0.06523	0.1244
50	100	0.04580	0.01556	0.04678	0.02032	0.02244	0.07007	0.1142
100	100	0.04660	0.01369	0.04599	0.01750	0.02248	0.07060	0.1285
200	200	0.04690	0.01064	0.04662	0.00854	0.03290	0.05830	0.1455
200	400	0.04613	0.00810	0.04500	0.00681	0.03337	0.05423	0.1410

Table 3. Failure estimation of (Rosenbrock function < 3) with truncated normal inputs. True value is 0.0458.

The next example tested was one with a discontinuous failure domain. This function, called F2, is given as:

$$F2 = [0.8r + 0.35\sin(\frac{2\pi r}{\sqrt{2}})][1.5\sin(1.3\theta]$$

$$\theta = \arctan(\frac{p_2}{p_1}), r = \sqrt{p_1^{\ 2} + p_2^{\ 2}}$$

This example is shown in Figure 3 and is documented in [Romero and Bankston, 1998]. It has two input variables, p_1 and p_2, which are taken as uniform distributions on [0,1]. Table 4 shows the failure probability estimate for the function being a threshold value of 0.5 as depicted in Figure 3 as approximately 55%. The importance sampling process generates more samples in the failure region (approximately 68%). This case demonstrates that the kernel density estimation is accurately capturing the discontinuous failure region.

Number of initial LHS samples	Number Of Importance samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Mean Failure probability	IS Std dev. Failure Probability	5th percentile IS Failure Probability	95th percentile IS Failure Probability	Mean percentage of IS that "fail"
100	200	0.55757	0.02969	0.55447	0.02995	0.50432	0.60248	0.6800
50	100	0.54753	0.04141	0.55115	0.04928	0.45469	0.62585	0.6538
100	100	0.55260	0.03678	0.55141	0.03976	0.48320	0.61335	0.6767
200	200	0.55133	0.02630	0.55560	0.02610	0.51041	0.59695	0.6959
200	400	0.55892	0.01779	0.54946	0.02053	0.51490	0.58168	0.6921

Table 4. Failure estimation of (F2>0.5). True value is 0.5547.

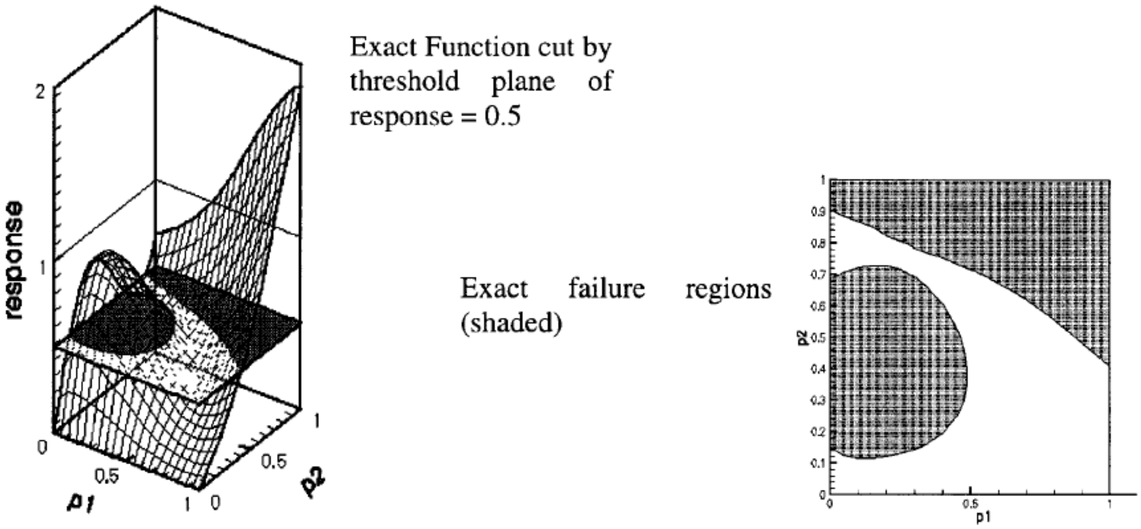

Figure 3. Test function F2, demonstrating failure probability estimation over discontinuous failure domain

The final test problem examined was the Rosenbrock function in 5 dimensions. Again, we assume a bounded domain on the inputs of $[-2,2]^5$, uniformly distributed. We examine the failure probability (Rosenbrock-5D <100). The results are shown in Table 5. There are a few things to note: importance sampling in this problem is obtaining failure probability estimates closer to the true value than LHS sampling, which appears to be biased slightly high. Also, it is important to remember that we are generating 100 iterations of the importance sampling process. Thus, although the mean estimates look good, if a user were performing this importance sampling process just once, they could obtain a very high estimate (e.g. 95th percentile estimate) or a very low estimate (e.g. such as the 5th percentile estimate). Thus, in practice, one might expect to see poorer results both with LHS sampling and importance sampling unless multiple replicates are taken. Finally, in this higher dimensional problem, we do see the percentage of importance sampling points that "fail" increasing as the number of initial LHS points is increased and/or the number of importance samples increases. Additionally, the spread of the failure estimates generated by importance sampling narrows as the number of samples increases. Finally, the percentage of importance sampling points that fail is approximately 8 times the original failure probability for this test problem.

Number of initial LHS samples	Number of Importance samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Mean Failure probability	IS Std dev. Failure Probability	5th percentile IS Failure Probability	95th percentile IS Failure Probability	Mean percentage of IS that "fail"
100	200	0.01123	0.00542	0.00788	0.00287	0.00349	0.01350	0.0706
50	100	0.01447	0.00740	0.00861	0.00494	0.00220	0.01827	0.0621
100	100	0.01250	0.00566	0.00873	0.00539	0.00195	0.01542	0.0738
200	200	0.00963	0.00403	0.00861	0.00353	0.00367	0.01434	0.0868
200	400	0.00977	0.00340	0.00882	0.00350	0.00501	0.01158	0.0882

Table 5. Failure estimation of (Rosenbrock 5D < 100). True value is 0.0090.

V. Use of surrogates in the importance sampling process

As demonstrated in Section IV, the use of small numbers of samples in importance sampling helped generate more points in the failure region, but did not always improve the failure probability estimates significantly. This section examines the use of surrogate methods (also called meta-models or response surface models) in the importance sampling process. The approach we take is simple: as before, we generate an initial number of LHS samples over the input domain according to the distributions on the input parameters. Based on the LHS samples, we construct a surrogate. The importance sampling density is again based on kernel density estimators. The only difference is that instead of evaluating input points from the importance sampling density with the "true" function, we evaluate the input points using the surrogate. Thus, the failure probability estimation is based on the surrogate response function, $\hat{r}(x)$ instead of the true response function, $r(x)$:

$$failure\_estimate_{is\_surrogate} = \frac{1}{M}\sum_{i=1}^{M} I(\hat{r}(X_i) < c)\frac{f(X_i)}{f_n^{\{KDE\}}(X_i)} \qquad (13)$$

The construction of meta-models for computational simulations is a large field of study; for good overviews see [Simpson et al., Viana et al.]. For the purposes of this study, we focus on two classes of surrogates which span the spectrum: quadratic polynomials (which are very cheap to evaluate via standard regression methods but tend to be inaccurate) and Gaussian process models (which are expensive to construct but much more accurate). More details about quadratic regression can be found in [Kutner et al.], and more details about Gaussian process models can be found in [Rasmussen and Williams]. We use a typical approach in both situations: both are implemented in Matlab. For the quadratic polynomial, we use the Matlab command 'regstats', and for the Gaussian process model, we use the library GPML using an exponential correlation function.

We first considered a quadratic polynomial surrogate. Table 6 presents some importance sampling results, based on the initial version of the Rosenbrock problem presented in Section IV where the goal is to obtain an estimate of the probability that the Rosenbrock function is less than 3, over uniform bounds on the input variables as follows: $-2 \leq x_1 \leq 2$ and $-2 \leq x_2 \leq 2$. The important thing to notice is that these quadratic polynomial surrogates, while not awful (average R^2 value around .77), are not very accurate and thus the resulting importance sampling estimates of failure probability are inaccurate (e.g. around 20% where the true value of this failure probability is 0.0383. If we look more carefully at the results of the quadratic surrogate as shown in Figure 4, we see that while the polynomial follows the general trend of the Rosenbrock function, it "under-predicts" significantly in many cases, and often predicts a negative value for this function, which is not accurate (the Rosenbrock function is always positive). The predictions of negative values for Rosenbrock contribute to the overall failure probability estimation, since the indicator function $I(\hat{r}(X_i) < c)$ will be 1 when $\hat{r}(X_i) < 3$. In this case, more points are "flagged" as failure points and contribute to the failure probability estimation as given in Equation 13, thus resulting in a failure estimate that is very high. The main lesson from examining quadratic polynomial surrogates is that even if the surrogate gets the general trend of the function approximately correct, this does not satisfy the requirements necessary to obtain accurate failure probability estimates via importance sampling.

Number of initial LHS samples	Number of Importance samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Mean Failure probability	IS Std dev. Failure Probability	Mean percentage of IS that "fail" according to surrogate	Mean R-squared value
100	200	0.03790	0.00942	0.23711	0.05731	0.4889	0.774
50	100	0.04080	0.01650	0.24725	0.08307	0.4594	0.792
100	100	0.03775	0.01162	0.23631	0.07030	0.5008	0.775
200	200	0.03895	0.01033	0.23833	0.07215	0.5130	0.761
200	400	0.03795	0.00746	0.23754	0.05025	0.5121	0.764

Table 6. Failure estimation of (Rosenbrock function < 3) based on quadratic polynomial surrogate. True value is 0.0383.

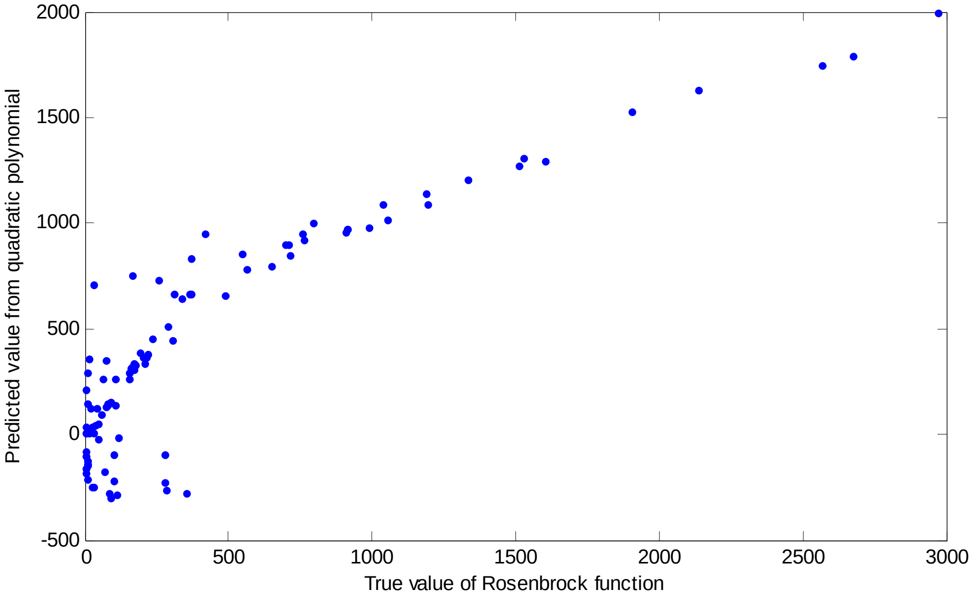

Figure 4. Quadratic Polynomial Surrogate Model for Rosenbrock Function

The other surrogate that we examined is expensive but much more accurate: Gaussian process surrogate models. The Gaussian process surrogate models tended to result in estimates that were similar to those obtained by performing the importance sampling directly on the function. Table 7 shows the failure estimates and Figure 5 shows the goodness of the predicted values using the GPs. Note that the Gaussian process predictions equal the true value of the Rosenbrock function, as indicated by the straight line in Figure 5, in contrast with Figure 4. In these cases, the failure estimation is better. HOWEVER, we were not able to average these results over one hundred iterations as we did for the previous analysis, because we had frequent problems with the ill-conditioning of the covariance matrix and estimation of the hyperparameters governing the covariance model. Thus, Table 7 shows results only for 10 iterations of the importance sampling process, where each iteration is based on a Gaussian process surrogate model. In general, the Rosenbrock function is very hard for GP emulators because of the wide range of the output values over a small domain. We do see some behavior that we expect, such as the results from the GP surrogates built from larger numbers of initial points (e.g. 200 vs. 50) gave more stable importance sampling estimates.

Number of initial LHS samples	Number of Importance samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Failure probability	IS Std dev. Failure Probability	percentage of IS that "fail" according to surrogate
100	200	0.04100	0.00955	0.03652	0.01038	0.1215
50	100	0.03799	0.01543	0.04053	0.01507	0.1240
200	400	0.03650	0.00506	0.03746	0.00281	0.1435

Table 7. Failure estimation of (Rosenbrock function < 3) based on Gaussian process surrogate. True value is 0.0383.

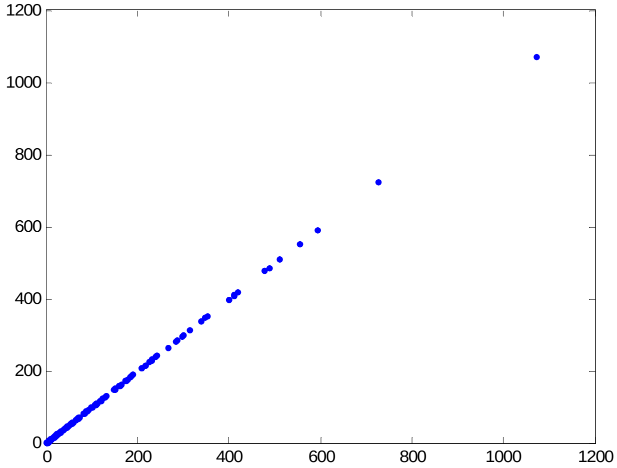

Figure 5. Gaussian Process Surrogate Model for Rosenbrock Function

To address the ill-conditioning of the GP, we took the log of the output and based the importance sampling estimates on prediction of the Rosenbrock function (by exponentiating the GP prediction). This gave better conditioned covariance estimates but overall gave poorer results than the GP in Figure 5, as indicated by less accurate GP predictions as well as less accurate importance sampling estimates shown in Figure 6 and Table 8. Note that, as expected, the importance sampling estimates are better than the very inaccurate quadratic polynomial but not as good as the GP based on the actual (not log transformed) data.

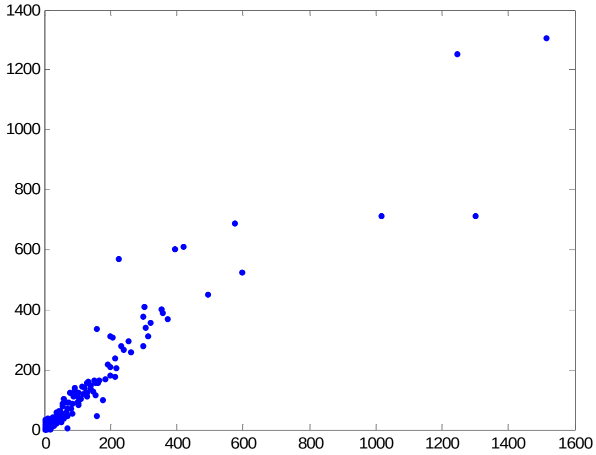

Figure 5. Gaussian Process Surrogate Model for Rosenbrock Function, performed in log transform space

Number of initial LHS samples	Number of Importance samples	LHS Mean Failure probability	LHS Std dev. Failure Probability	IS Failure probability	IS Std dev. Failure Probability	percentage of IS that "fail" according to surrogate
100	200	0.03957	0.01012	0.02628	0.01872	0.1045
50	100	0.03960	0.01495	0.02377	0.01999	0.0968
200	400	0.03758	0.00780	0.03420	0.01223	0.1461

Table 8. Failure estimation of (Rosenbrock function < 3) based on Gaussian process surrogate using a log transform. True value is 0.0383.

This has been a very limited investigation of surrogate model use in importance sampling. However, this examination has demonstrated that the surrogate must be extremely accurate to generate accurate failure probability estimates using importance sampling.

VI. Conclusions

This paper has presented an importance sampling approach based on kernel density estimators. Although many nonparametric approaches to importance sampling in the literature use some type of density estimation, we tried to develop an approach that is robust, could be implemented for any type of "black-box" simulation problem, and is initially seeded or started based on an initial set of Latin Hypercube samples from the black-box simulation. We tested this approach on a variety of test problems, focusing on situations with small sample sizes to be representative of expensive computational simulations. We found that importance sampling using this approach is reasonably robust and can produce failure probability estimates that are comparable to or more accurate than similar failure estimates produced by small numbers of LHS sample points. However, the accuracy of the failure estimates was not greatly improved by using the importance sampling approach given the limitations on the number of samples that we used. The main benefit we see by using this approach is that the kernel density estimators provide a quick way to generate more samples in the failure region, for situations where the users do not know a priori anything about where the failure region may be located. We found that importance sampling increased the number of samples in the failure region by a factor of 3 to 8 for our test cases. However, the overall failure probabilities obtained were not significantly better than those obtained by pure sampling, in large part because the number of points generated was small. Users need to be cautious about the accuracy of failure estimates generated by importance sampling in small sample situations. Finally, surrogate models should only be used in the importance sampling process in the case of very accurate surrogate models.

Acknowledgments

The authors thank the Advanced Simulation and Computing (ASC) Program of the National Nuclear Security Administration (NNSA) at Sandia National Laboratories for funding this research.

References

Asmussen, S. and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis.* Springer-Verlag, 2007.

Bichon, B.J., Eldred, M.S., Swiler, L.P., Mahadevan, S., and McFarland, J.M., "*Multimodal Reliability Assessment for Complex Engineering Applications using Efficient Global Optimization*," paper AIAA-2007-1946 in Proceedings of the 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference (9th AIAA Non-Deterministic Approaches Conference), Honolulu, HI, April 23-26, 2007.

Denny, M. "*Introduction to importance sampling in rare-event simulations*." European Journal of Physics, 22(2001), pp. 403-311.

Dey, A. and Mahadevan, S., "*Ductile Structural System Reliability Analysis using Adaptive Importance Sampling*", Structural Safety, Vol. 20, 1998, pp. 137-154.

Evans, M. and Swartz, T. "*Bayesian Integration using Multivariate Student Importance Sampling.*" Computing Science and Statistics, 27, p. 456-461.

Eldred, M.S., Brown, S.L., Adams, B.M., Dunlavy, D.M., Gay, D.M., Swiler, L.P., Giunta, A.A., Hart, W.E., Watson, J.-P., Eddy, J.P., Griffin, J.D., Hough, P.D., Kolda, T.G., Martinez-Canales, M.L. and Williams, P.J., "*DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 4.0 Users Manual,*" Sandia Technical Report SAND2006-6337, October 2006. Updated November 2008 (version 4.2). Website: http://www.cs.sandia.gov/DAKOTA.

Givens, G. H. and Raftery, A. E. "*Local Adaptive Importance Sampling for Multi-variate Densities with Strong Nonlinear Relationships.*" Journal of the American Statistical Association, 1996, vol. 91, no. 433, pp. 132-141.

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

Gentle, J.E. *Random Number Generation and Monte Carlo Methods.* Springer Statistics and Computing Series, 2004.

Kutner, M. H., Nachtsheim C. J., Neter, J. And W. Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin Series Operations and Decision Sciences, 2004.

Knuth, D. E. *The Art of Computer Programming, Vol 2: Seminumerical Algorithms*. 3rd edition, 1997.

Oh, M.S. and J. O. Berger. "*Integration of Multimodal Functions by Monte Carlo Importance Sampling.*" Journal of the American Statistical Association, Vo. 88, No. 422, pp. 450-456, 1993.

Parzen, E. *On Estimation of a Probability Density Function and More*. The Annals of Mathematical Statistics, Volume 33, Number 3, September 1962, pp. 1065-1076.

Rasmussen, C. E. and C. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.

Richard, J-F. and Zhang, W. "*Efficient High-Dimensional Importance Sampling.*" Journal of Econometrics, Volume 141, Issue 2, December 2007, Pages 1385-1411.

Romero, V.J. and S. D. Bankston. "Efficient Monte Carlo Probability Estimation with Finite Element Response Surfaces built from Progressive Lattice Sampling." Sandia Technical Report, SAND1998-0607C.

Rosenblatt, M. *Remarks on Some Nonparametric Estimates of a Density Function*. The Annals of Mathematical Statistics, Volume 27, 1956, pp. 832-835.

Silverman, B.W. *Density Estimation.* Chapman and Hall, 1986.

Simpson, T. W., Toropov, V., Balabanov, V., and F. A. C. Viana. "*Design and Analysis of Computer Experiments in Multidisciplinary Design Optimization: A Review of How Far We Have Come – or Not.*" 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, British Columbia, Canada, AIAA, AIAA-2008-5802.

Swiler, L.P. and Wyss, G.D., *A User's Guide to Sandia's Latin Hypercube Sampling Software: LHS UNIX Library Standalone Version,* Sandia Technical Report SAND2004-2439, July 2004.

Turlach, B.A. *Bandwidth Selection in Kernel Density Estimation: A Review*. CORE and Institut de Statistique Technical Report.

Viana, F. A. C., Haftka, R.T., Steffen Jr, V., Butkewitsch, S., and Leal, M. F., "*Ensemble of Surrogates : A Framework based on minimization of the mean integrated square error*," 49th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 10th AIAA Non-Deterministic Approaches Conference, Schaumburg, USA, April 7-10, 2008.

Wand, M.P. and Jones, M.C. *Kernel Smoothing*. Chapman and Hall, 1994.

Zhang, P. "*Nonparametric Importance Sampling*." Journal of the American Statistical Association, Vol. 91, no. 435, pp. 1245-1253, 1996.