

# Some Test Problems and Results in Assessing Methods for Calculating Low Probabilities of Failure

*Vicente Romero, Laura Swiler, Mohamed Ebeida,  
Scott Mitchell, Matthew Glickman*

**Sandia National Laboratories\***  
**Albuquerque, NM**

**18<sup>th</sup> AIAA Non-Deterministic Approaches Conference**  
**AIAA SciTech 2016, Jan. 4-8, San Diego, CA**

\* Sandia 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.

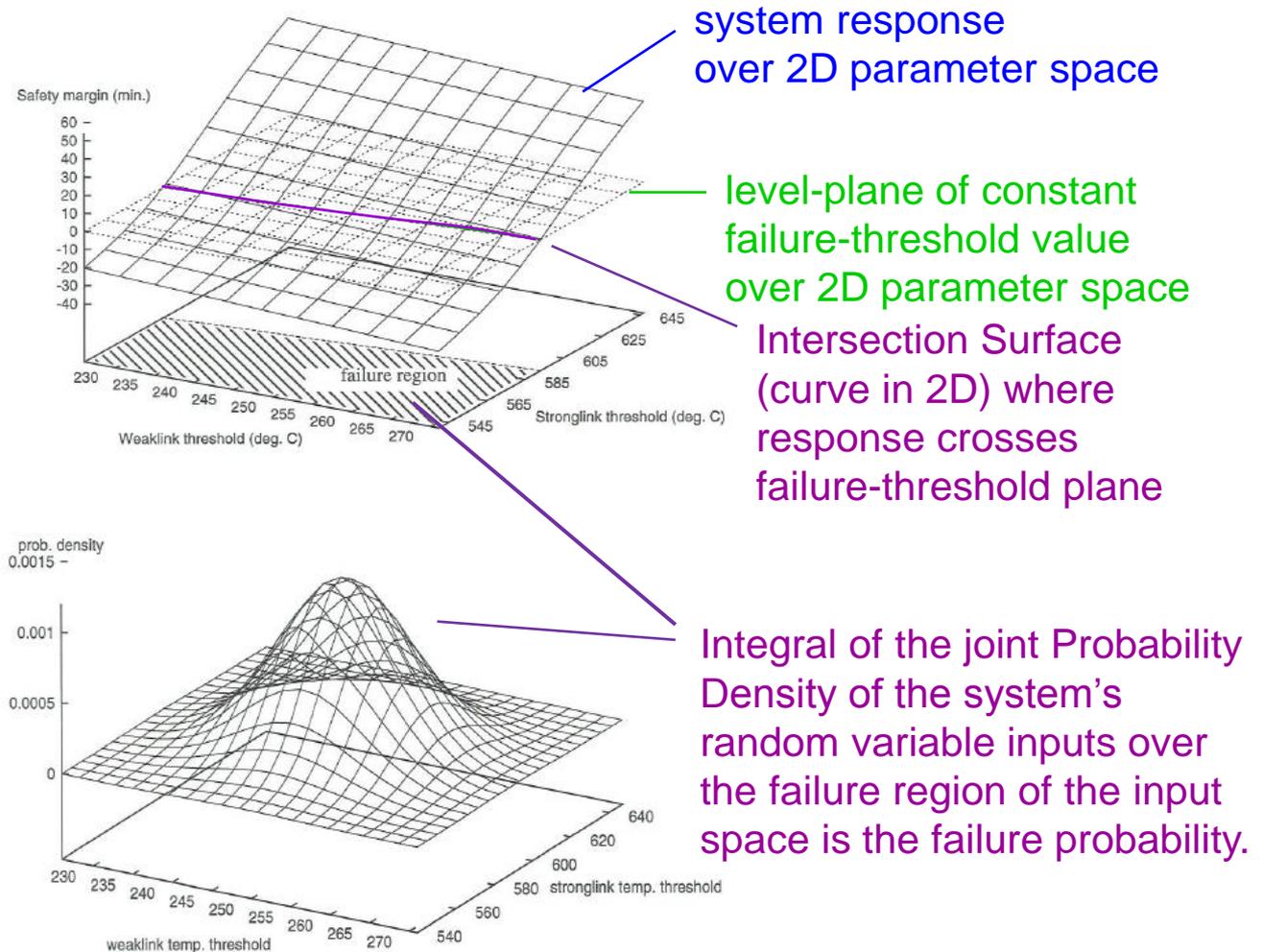
# Introduction and Motivation

---

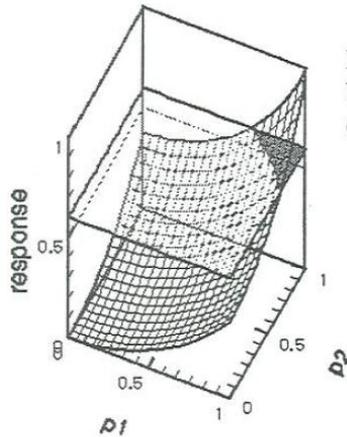


- **Estimation of the probability of failure (POF) to meet critical safety or performance constraints, requirements, or goals is important in:**
  - Engineering design and safety analysis
  - Business, finance, economics
  - Environmental management and regulation
  - etc.
- **We examine the performance of several established and new methods for estimating low probabilities of failure, a difficult and computationally expensive task to do accurately on general problems.**
  - The focus here is representative 2D - 9D engineering problems,  $10^{-2}$  to  $10^{-4}$  failure probabilities, and  $\leq 1000$  model runs
- **Evaluation criteria:**
  - performance in terms of cost, accuracy, robustness
  - ease of implementation and use for engineering practice

# Example 2D Failure Probability Problem with Constant Failure-Level over response space

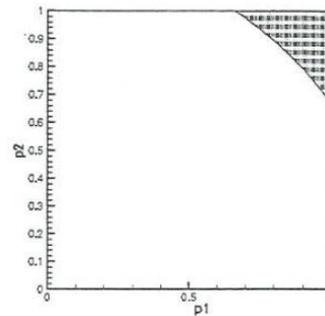


# Example 2D Failure Probability Problems

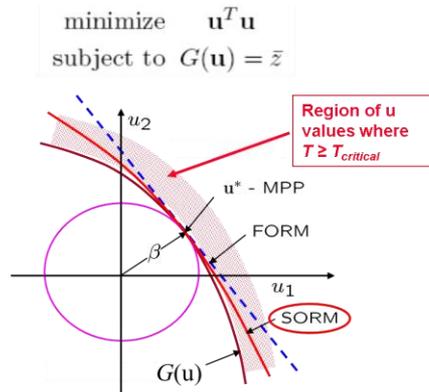


Exact Function cut by threshold plane of response = 0.6

Exact failure region (shaded)

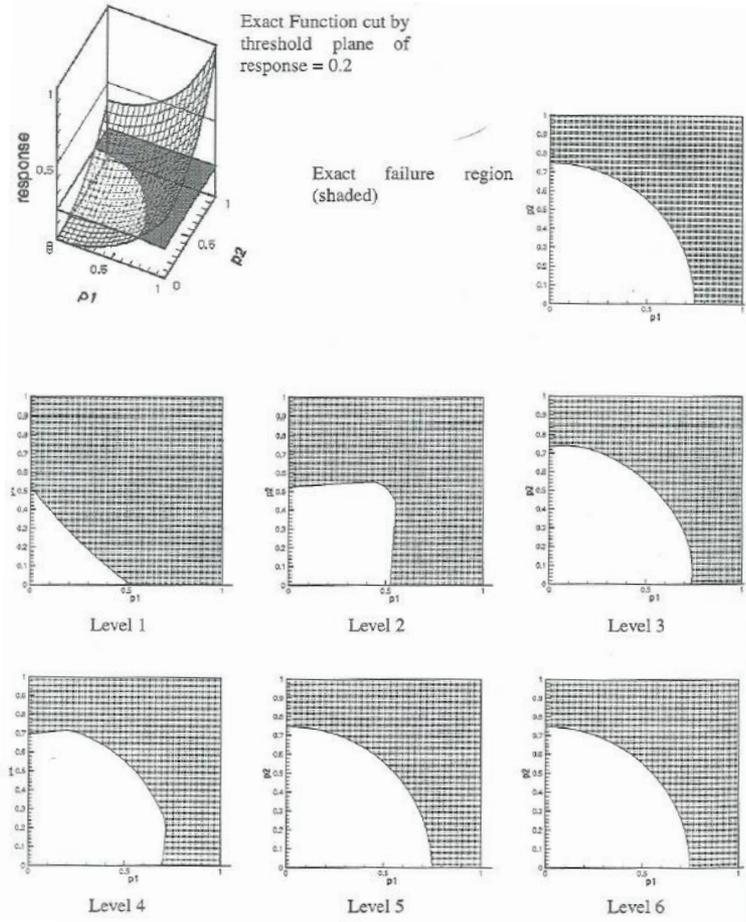


- **Classical Reliability Methods work for this type of problem**
- **relatively inexpensive in high dimensions**



- **Don't need to build global response function over the space**
- **Use optimization to locate a representative point on failure surface and perturb samples about this point to model surface as linear or quadratic**

# Example 2D Failure Probability Problems



## 2D Lattice-Sampling with Finite-Element interpolation patches

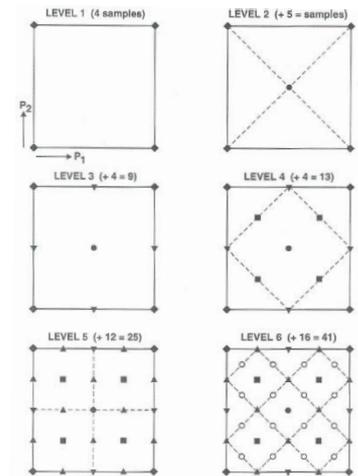
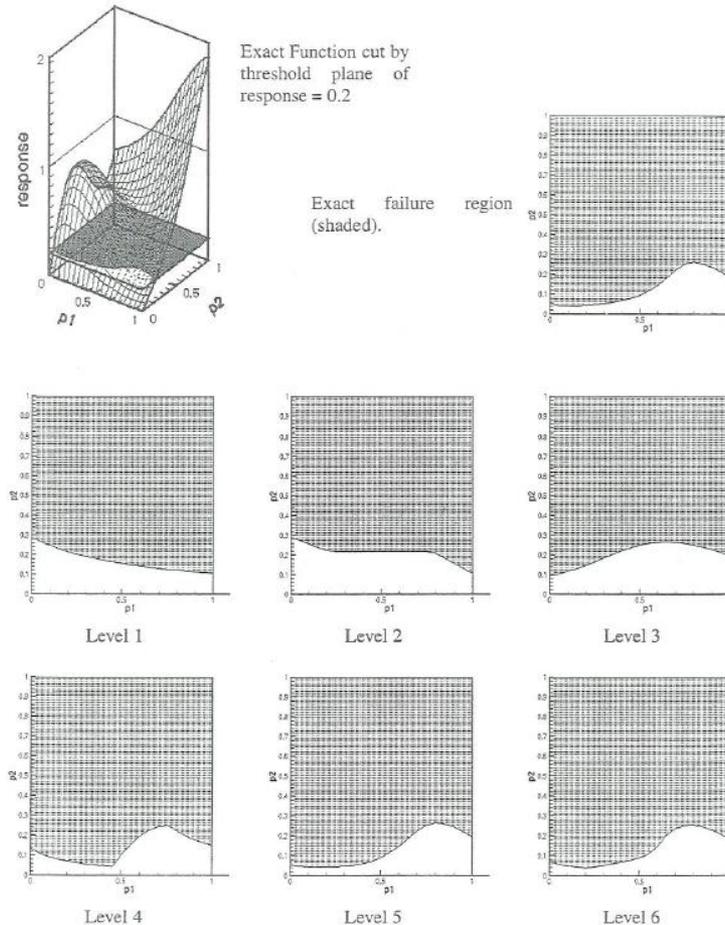


Figure 2.1 2-D Lattice Sampling Levels and associated discretization of the parameter space.

Classical reliability methods may not work well here even though response is monotonic in the input variables

# Example 2D Failure Probability Problems



## 2D Lattice-Sampling with Finite-Element interpolation patches

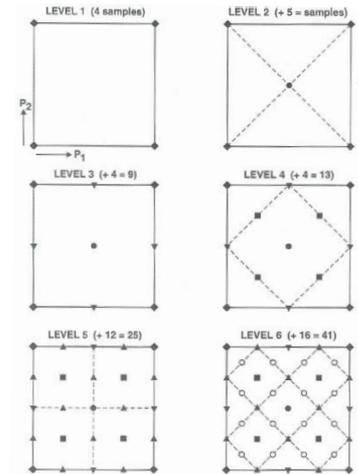
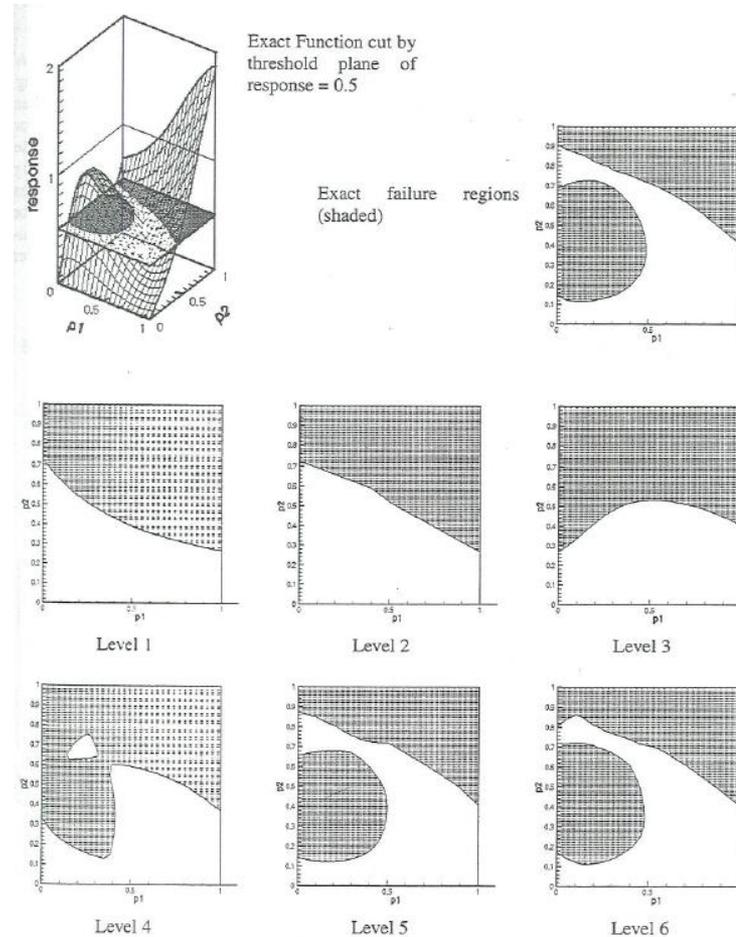


Figure 2.1 2-D Lattice Sampling Levels and associated discretization of the parameter space.

Classical reliability methods won't work well here (non-monotonic response)

# Example 2D Failure Probability Problems



## 2D Lattice-Sampling with Finite-Element interpolation patches

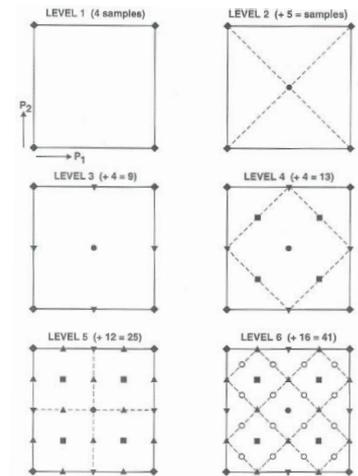


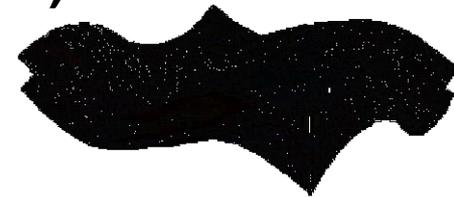
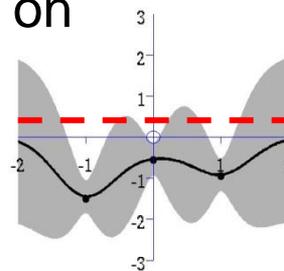
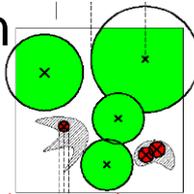
Figure 2.1 2-D Lattice Sampling Levels and associated discretization of the parameter space.

Classical reliability methods won't work here (non-monotonic response)

# Failure Probability Methods being evaluated for these more difficult POF problems

---

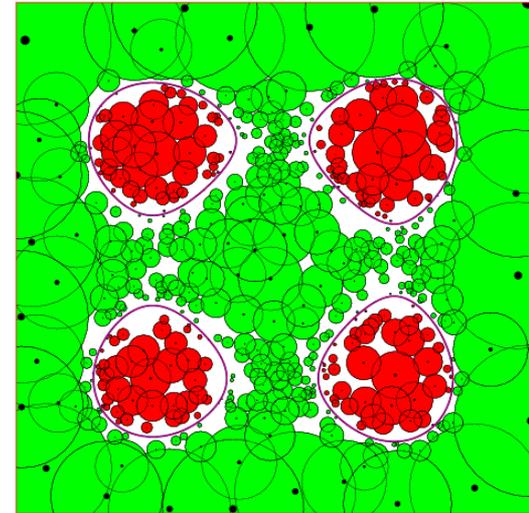
- **Method 1**: **Monte Carlo Sampling** (non-adaptive)
- **Method 2**: **POF Darts** — *new* adaptive Sandia approach based on compu. geom. methods (Mohamed Ebeida, S. Mitchell, L. Swiler)
- **Method 3**: **EGRA** — *Efficient Global Reliability Analysis*, based on gaussian-process response surfaces and adaptive sampling (Barron Bichon, S. Mahadevan et al., Dakota implementation)
- **Method 4**: **Gaussian Processes built on Latin Hypercube Sampling points** (non-adaptive, DAKOTA implementation)
- **These methods all have an element of Stochasticity in their performance**  
→ **must characterize performance variation over multiple trials**



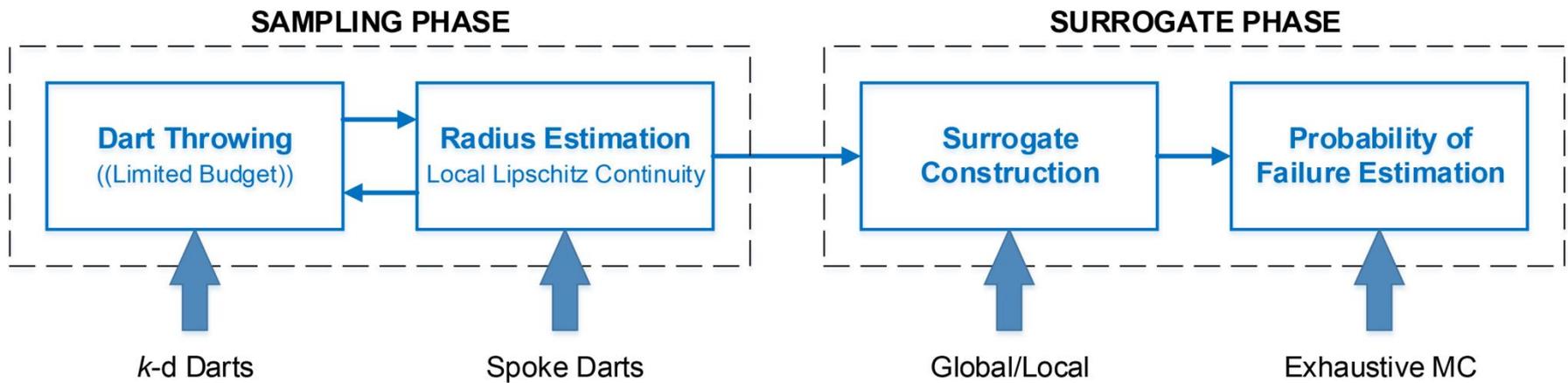
# POF Darts

## Probability-of-Failure Darts

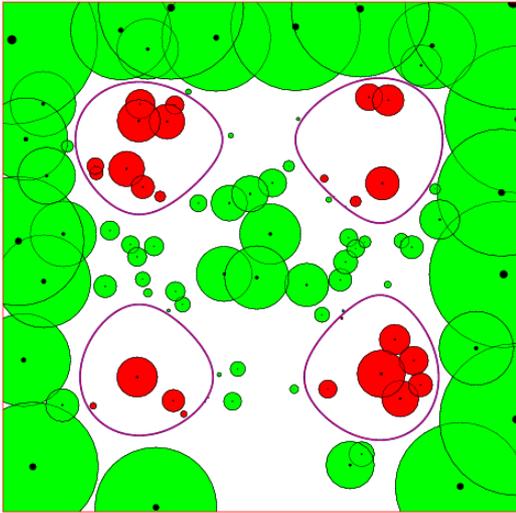
- Based on ideas from computational geometry.
- We employ random disk-packing (e.g. iteratively throwing darts to obtain the centers of the disks) in the uncertain parameter space.
- POF-Darts subdivides the uncertain space into three regions:
  - Failure (covered by red disks)
  - Non-failure (covered by green disks)
  - Unexplored (uncovered)
- We always sample points from the unexplored region.
- The function evaluation at that point determines whether it belongs to failure or non-failure.
- An estimate of the Lipschitz continuity of the function (approximated by the local maximum gradient) is utilized to construct a sphere centered around that point and which lies entirely in the same region as its center.
- As we proceed with this sampling procedure, the unexplored regions shrink and the accuracy of our estimate improves.
- After exhausting our function evaluation budget, we build a surrogate based on the sample points and estimate the probability of failure by exhaustive sampling of that surrogate.



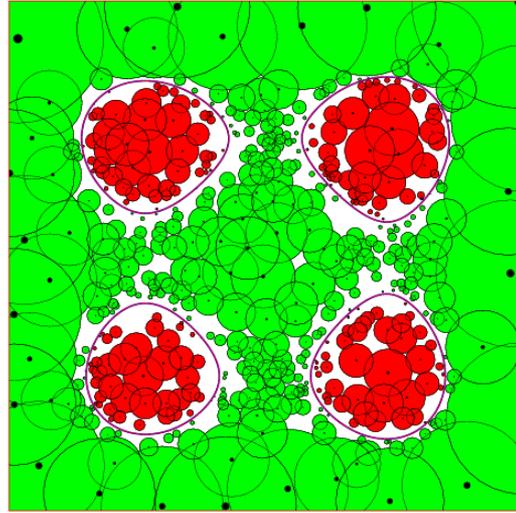
# POF-Darts



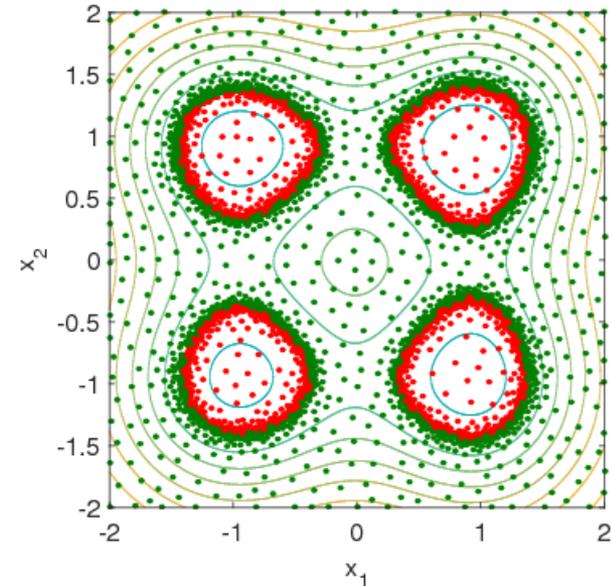
# POF-Darts



**100 Samples and  
associated disks**



**500 Samples and  
associated disks**



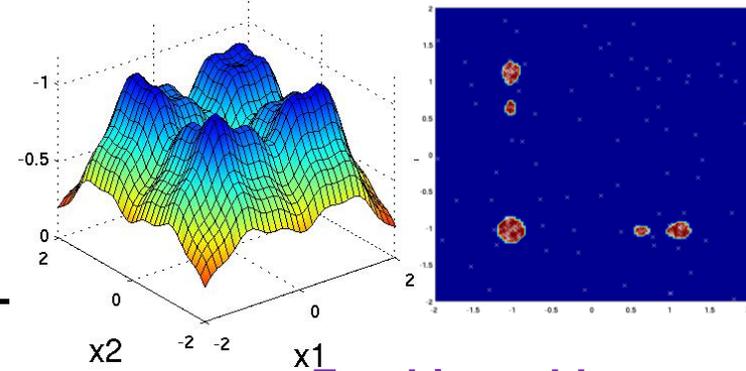
**5000 POF Sample Points**

- These left and center graphic show the Herbie test problem with four failure regions. The exact failure isocontours are in blue and the estimated ones are in red. These overlay in these plots, indicating accurate estimation of probability of failure.
- The right graphic shows only the points, at 5000 samples points, demonstrating that the samples tend to focus around the boundary of the failure region.

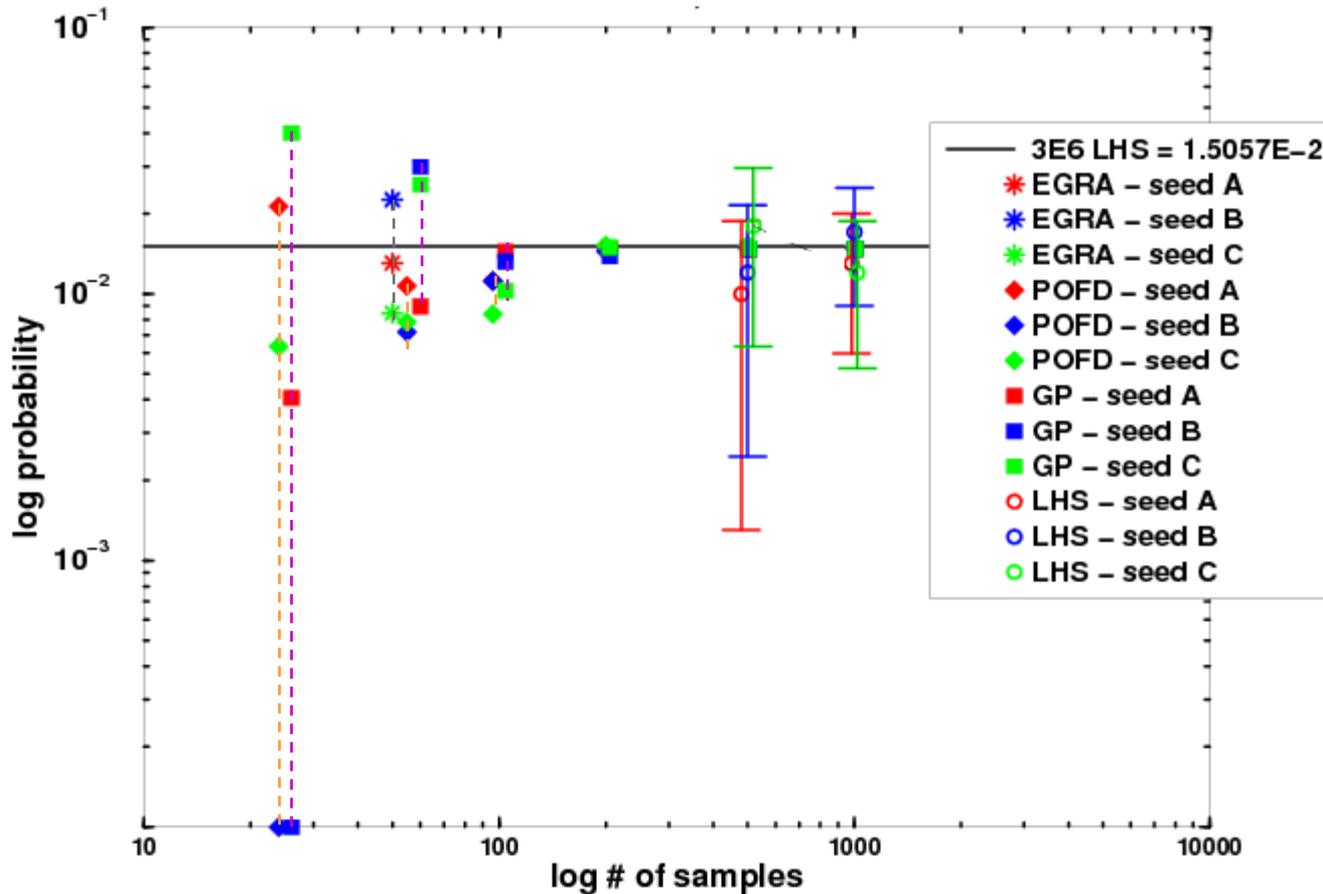
# Test Problem 1

## 2D Herbie test function

- 2 Uniform PDF input uncertainties
- $P_{fail} = 1.506E-2$



For this problem:

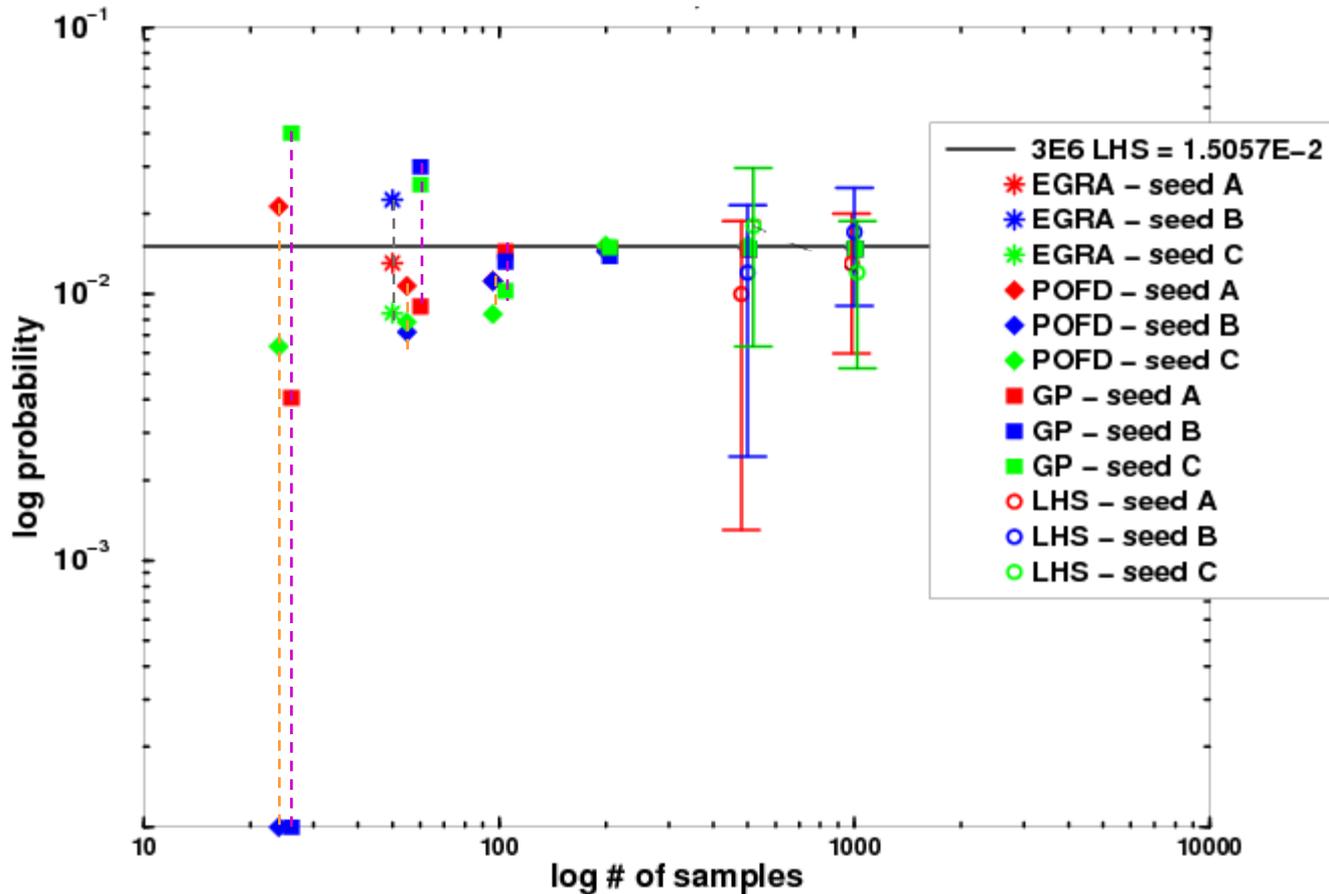
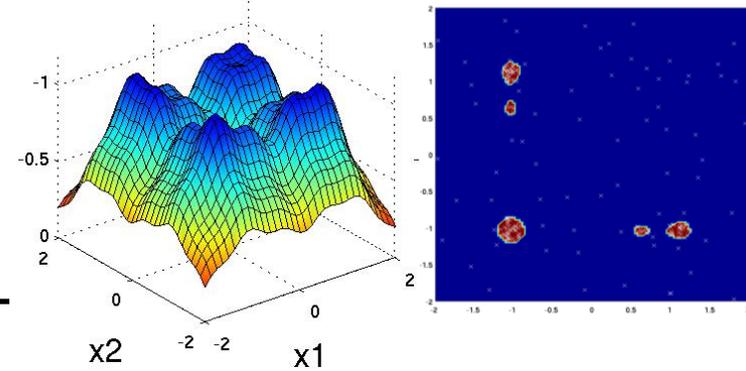


- **POFD-GP** & **LHS-GP** are not reliable with 25 sims. (very high variability with seeds, & large average error).
- **EGRA** “converged” after 55 sims. for all seeds (A, B, C), exhibiting significant seed dependence of individual results, but small average error.
- **LHS-GP** with 55 sims. performed almost as well, having slightly worse variability & avg. error than EGRA.
- **POFD-GP** w/55 sims. performed next best; lowest variability but highest avg. error (though not large).

# Test Problem 1

## 2D Herbie test function

- 2 Uniform PDF input uncertainties
- $P_{fail} = 1.506E-2$



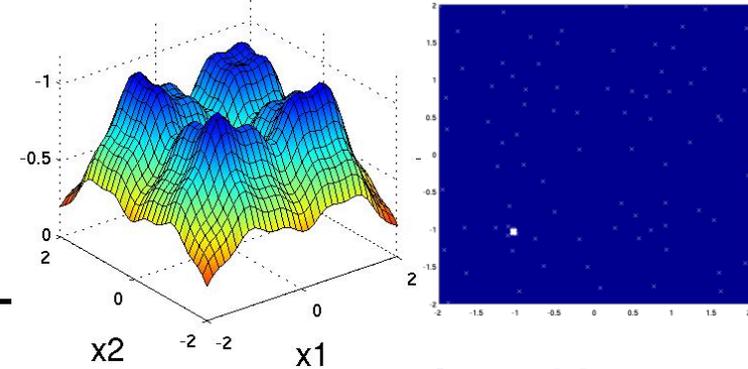
### For this problem:

- **POFD-GP** and **LHS-GP** both improve in variance & avg. error at 100 sims. (with GPs still having somewhat smaller avg. error than POFD). Both improve to negligible variance & avg. error for  $\geq 200$  sims.
- **LHS** shows non-negligible variance & avg. error for point estimates with 500, 1000 samples but confidence intervals are reliable for  $N > 500$  ( $N \cdot p \geq 5$ ).

# Test Problem 2

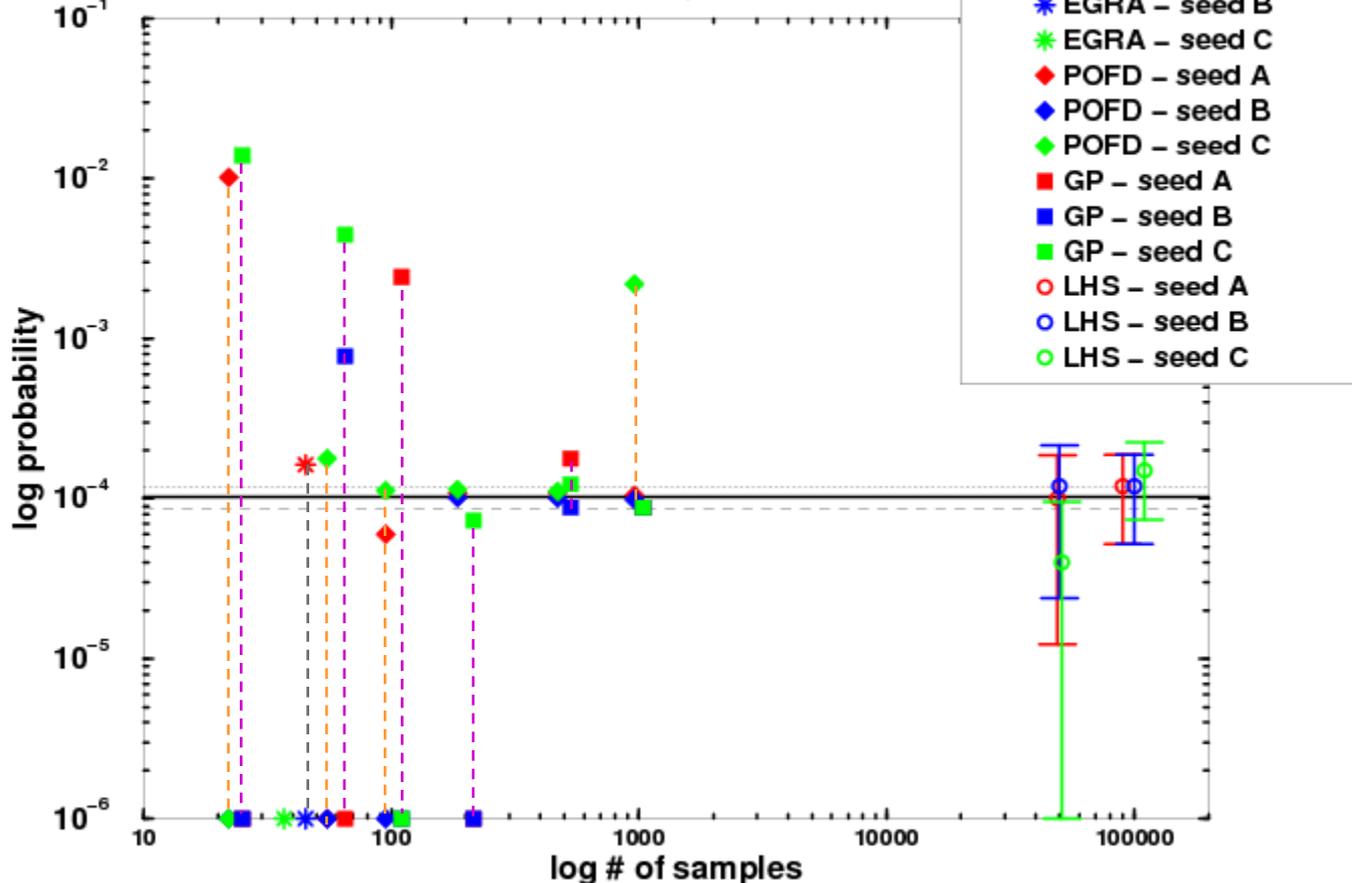
## 2D Herbie test function

- 2 Uniform PDF input uncertainties
- $P_{fail} = 1.023E-4$



### 2D Herbie test function

2 Uniform PDFs,  $P_{fail} \sim 0.0001$



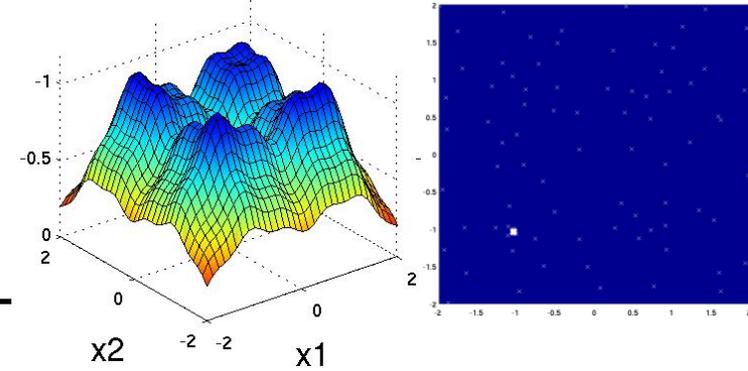
### For this problem:

- **EGRA** “converged” after 55, 55, 37 sims. for seeds A, B, C, giving non-zero failure probability only for seed A w/55 sims. ▶ significant seed dependence & premature convrgnc.
- **POFD-GP** and **GPs** are equally or more unreliable with 55 & 100 sims. (and w/25).
- **POFD-GP** is very accurate & precise with 200, 500 sims. for all seeds, but has an anomaly at 1000 sims. for seed C.
- **LHS-GP** isn’t reliable until  $\geq 500$  sims.

# Test Problem 2

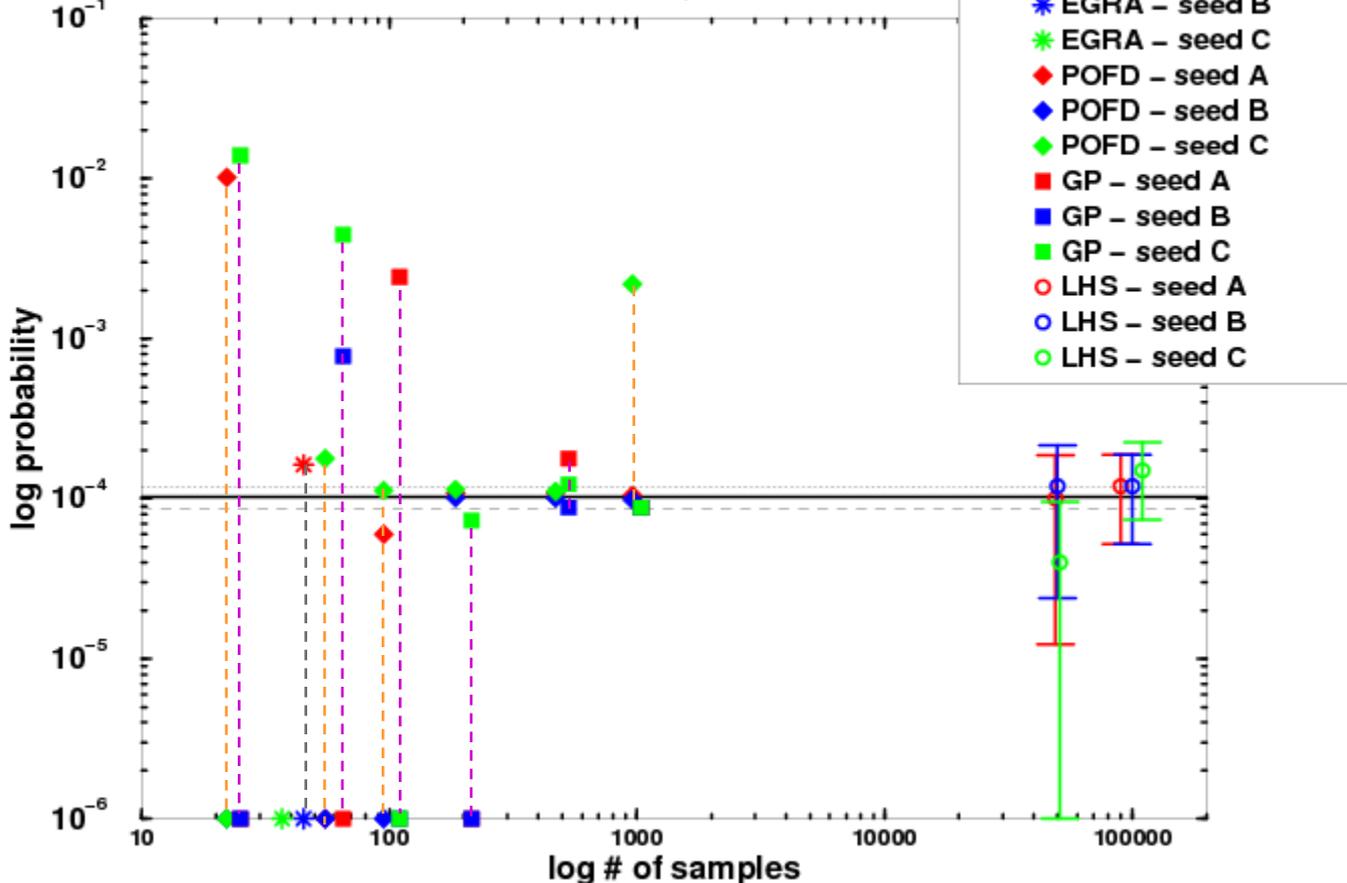
## 2D Herbie test function

- 2 Uniform PDF input uncertainties
- $P_{fail} = 1.023E-4$



### 2D Herbie test function

2 Uniform PDFs,  $P_{fail} \sim 0.0001$



### For this problem:

- **LHS** shows non-negligible variance & avg. error for point estimates with 50K, 100K samples but confidence intervals are reasonably reliable for  $N \geq 50K$  (this equates to  $N \cdot p \geq 5$ ).

# Test Problem 3



## 2D Vibration Absorber Problem

$$y(\beta_1, \beta_2) = \frac{\left| 1 - \left( \frac{1}{\beta_2} \right)^2 \right|}{\sqrt{\left[ 1 - R \left( \frac{1}{\beta_1} \right)^2 - \left( \frac{1}{\beta_1} \right)^2 - \left( \frac{1}{\beta_2} \right)^2 + \left( \frac{1}{\beta_1 \beta_2} \right)^2 \right]^2 + 4\zeta^2 \left[ \frac{1}{\beta_1} - \frac{1}{\beta_1 \beta_2^2} \right]^2}} \quad (7)$$

The random variables of the problem are  $\beta_1$  and  $\beta_2$ , and they follow normal distribution with mean value of 1 and standard deviation of 0.025.  $R$  and  $\zeta$  are deterministic parameters that possess the following values  $R=0.01$ ,  $\zeta=0.01$ . The normalized amplitude of the original system is plotted in Figure 3-b. The value of  $y_{crit}$  in Eq. (6) is adjusted to obtain various values of reliability indices as listed in Table 1.

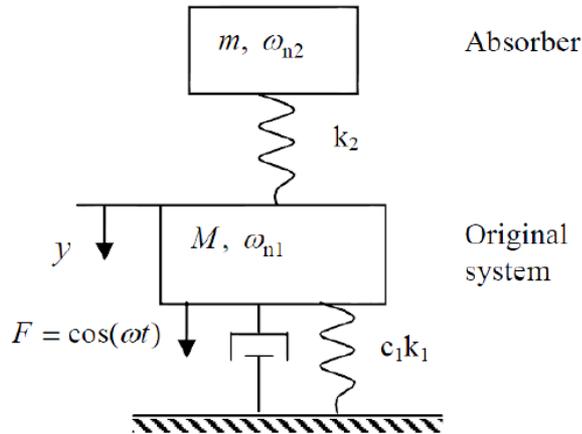


Figure 3-a. Tuned vibration absorber

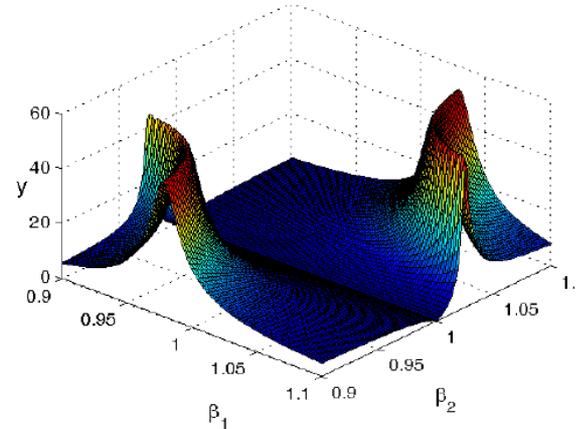
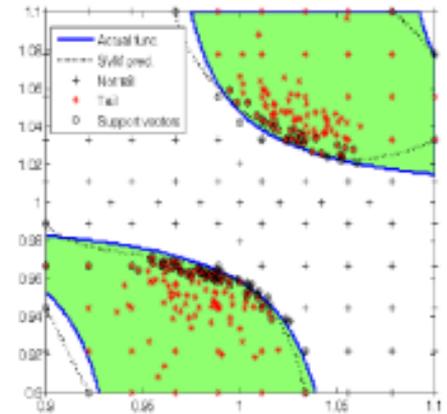


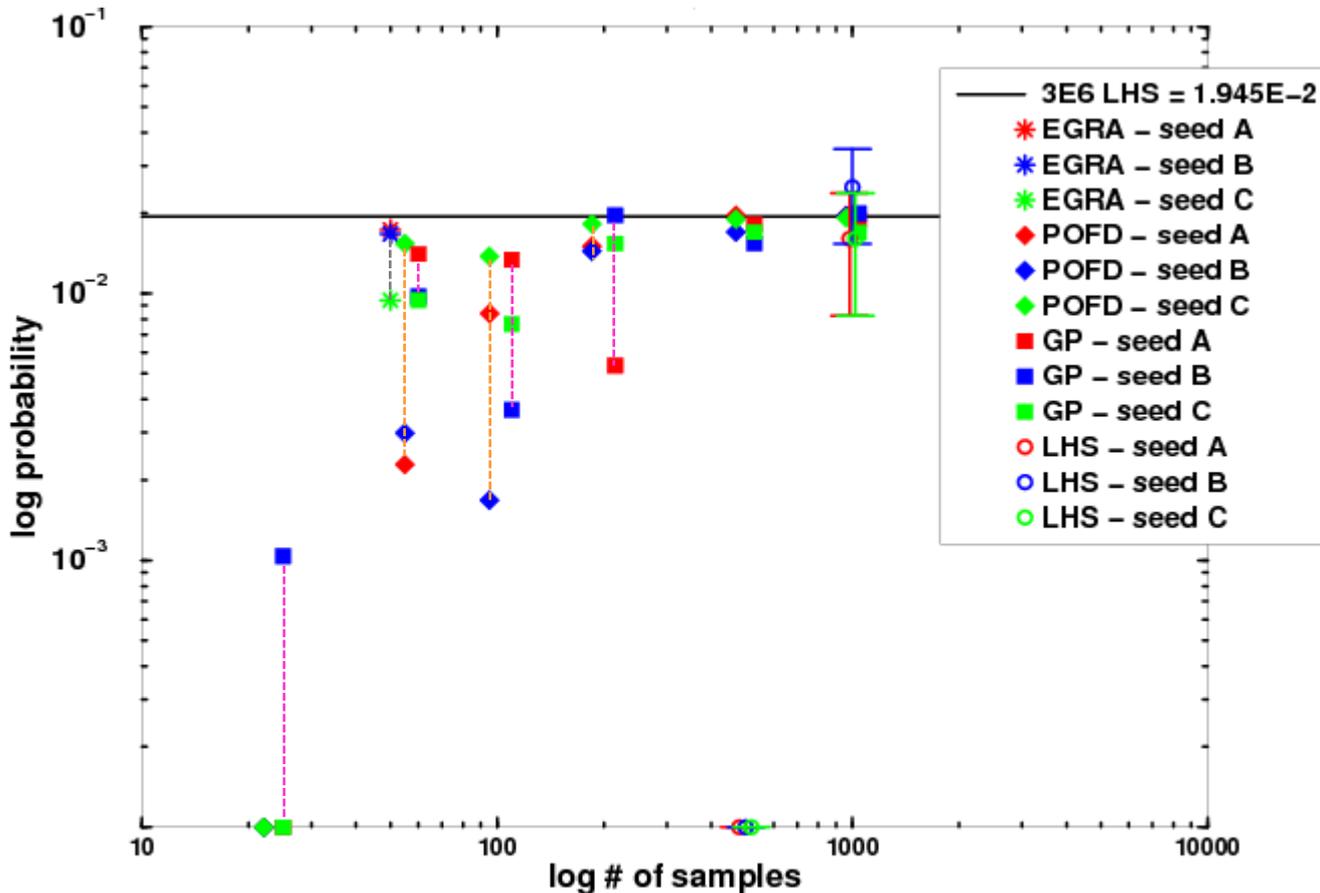
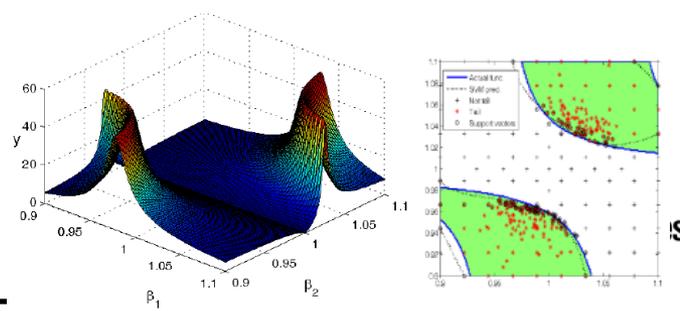
Figure 3-b. The normalized amplitude of the vibration absorber



# Test Problem 3

## 2D Vibration Amplitude problem

- 2 Uniform PDF input uncertainties
- $P_{fail} = 1.945E-2$



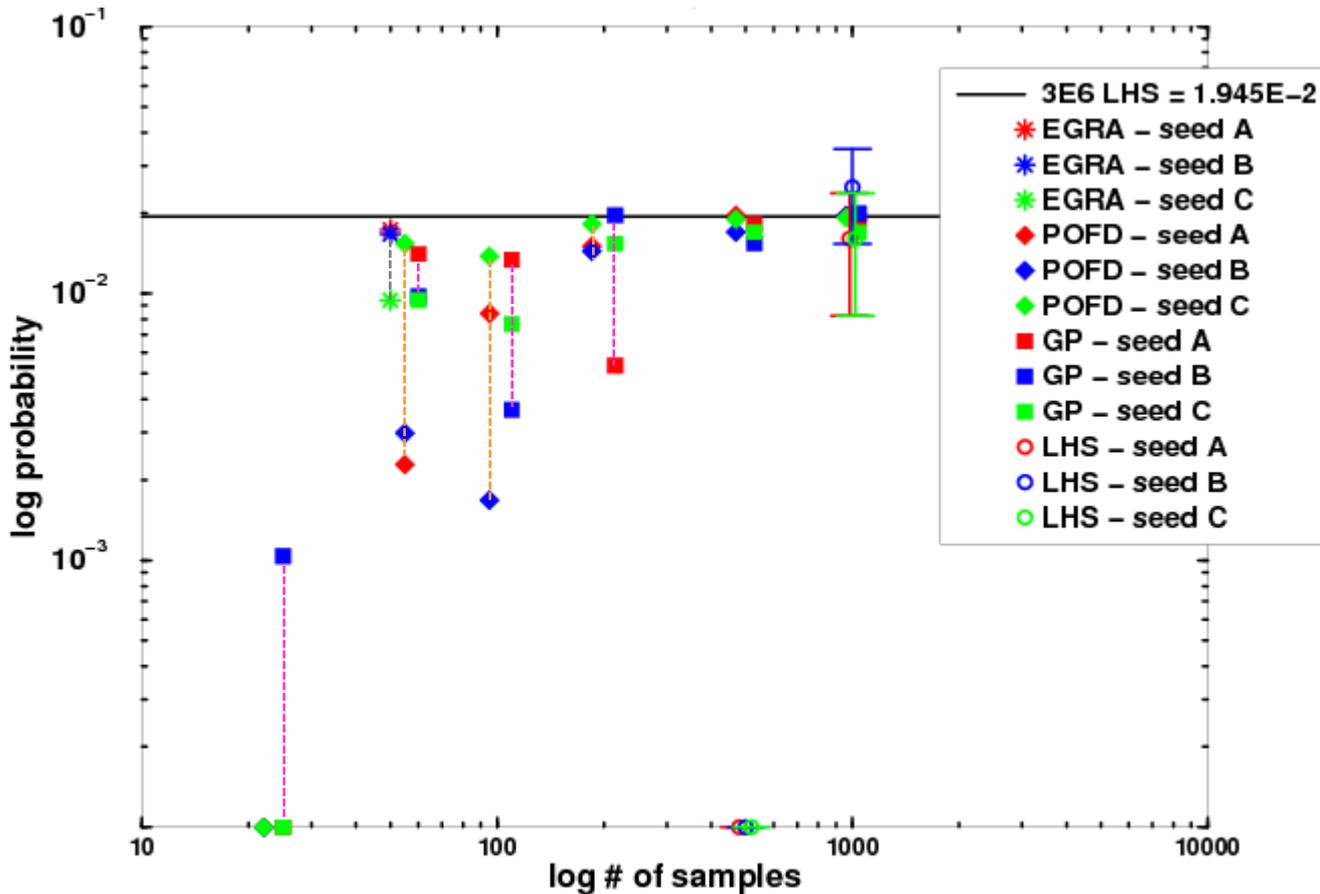
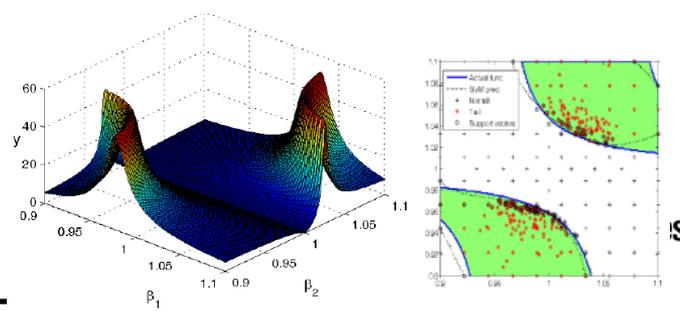
### For this problem:

- **EGRA** “converged” after 55 sims. with all seeds A, B, C, giving reasonable precision and accuracy—better than GPs and POFD at 55 sims.
- **LHS-GPs** are more accurate on average & more precise than POFD at 55 and 100 sims.
- **POFD-GP** improves to be very accurate and precise at 200, 500, 1000 sims., better than GPs.
- **POFD-GP** and **LHS-GPs** are not reliable with 25 sims. (large avg. error)

# Test Problem 3

## 2D Vibration Amplitude problem

- 2 Uniform PDF input uncertainties
- $P_{fail} = 1.945E-2$



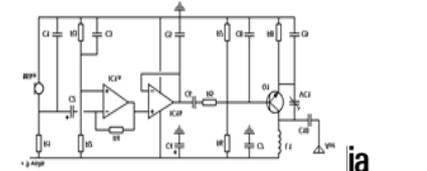
For this problem:

- **LHS** is very inaccurate with 500 sims. (0.0 probability values and 0.0 confidence intervals for all seeds A, B, C), but for 1000 sims. gives reasonable accuracy and precision of point estimates and reliable conf. intervals. ( $N \cdot p \geq 20$ )

# Test Problem 4

## 5D Circuit problem

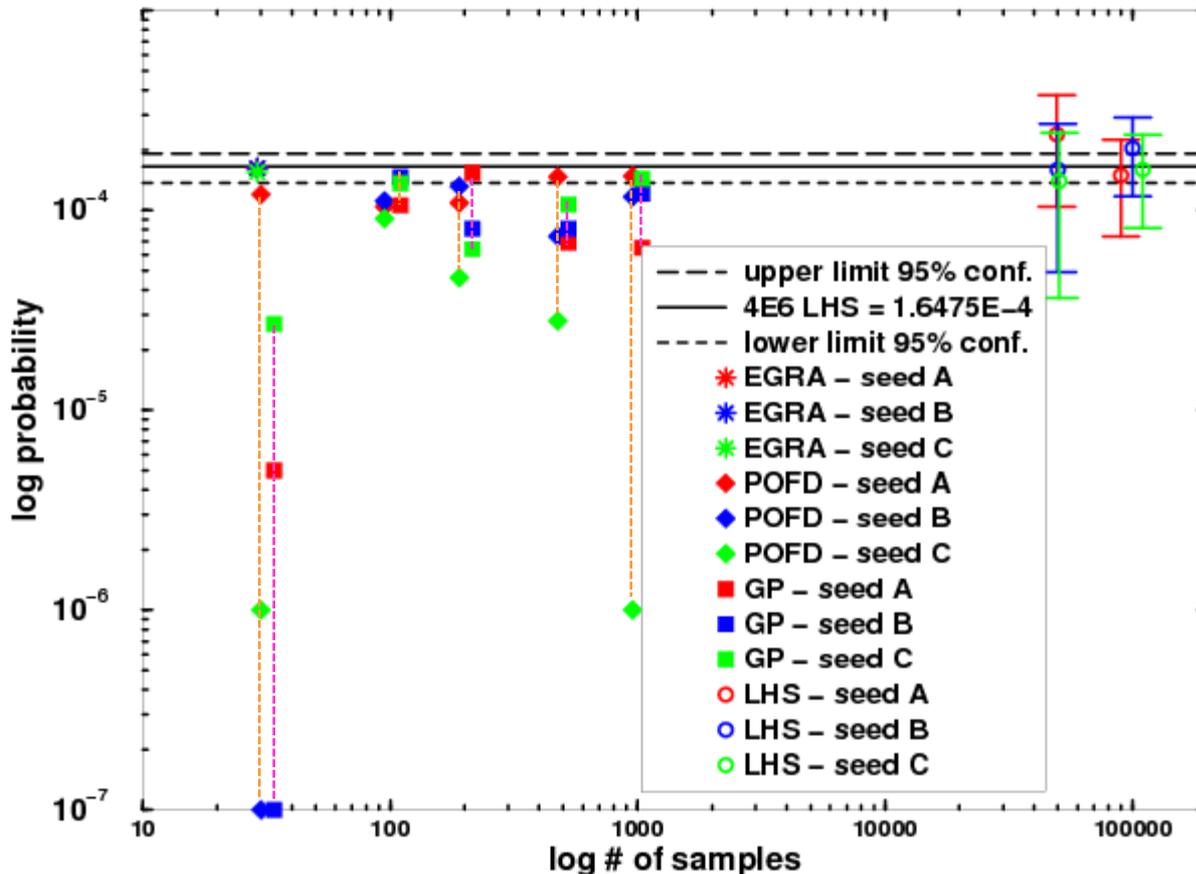
- 5 Uniform PDF input uncertainties
- $P_{fail} \approx 1E-4$



cartoon figure of a generic circuit

For this problem:

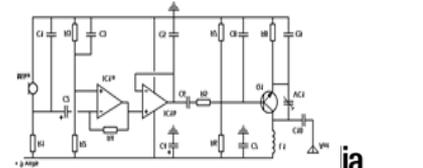
- **EGRA** converged w/ 31 sims., is most *accurate and precise* over seeds A, B, C.
- **POFD-GP** w/ 31 sims. gives almost as good results as EGRA for seeds A, B and C (plot at left is outdated). POFD-GP retains high accuracy-cost effectiveness through 1000 samples, see accuracy cost slide later.



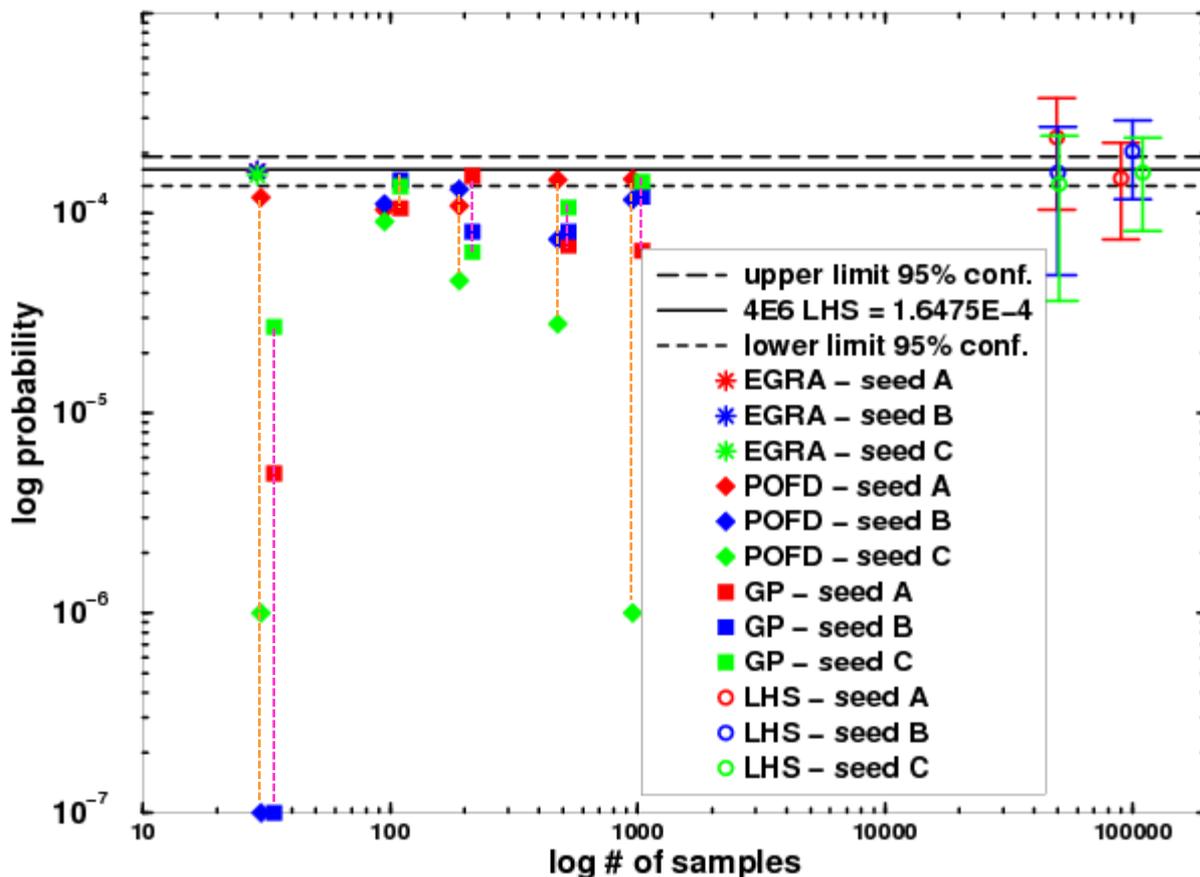
# Test Problem 4

## 5D Circuit problem

- 5 Uniform PDF input uncertainties
- $P_{fail} \approx 1E-4$



cartoon figure of a generic circuit



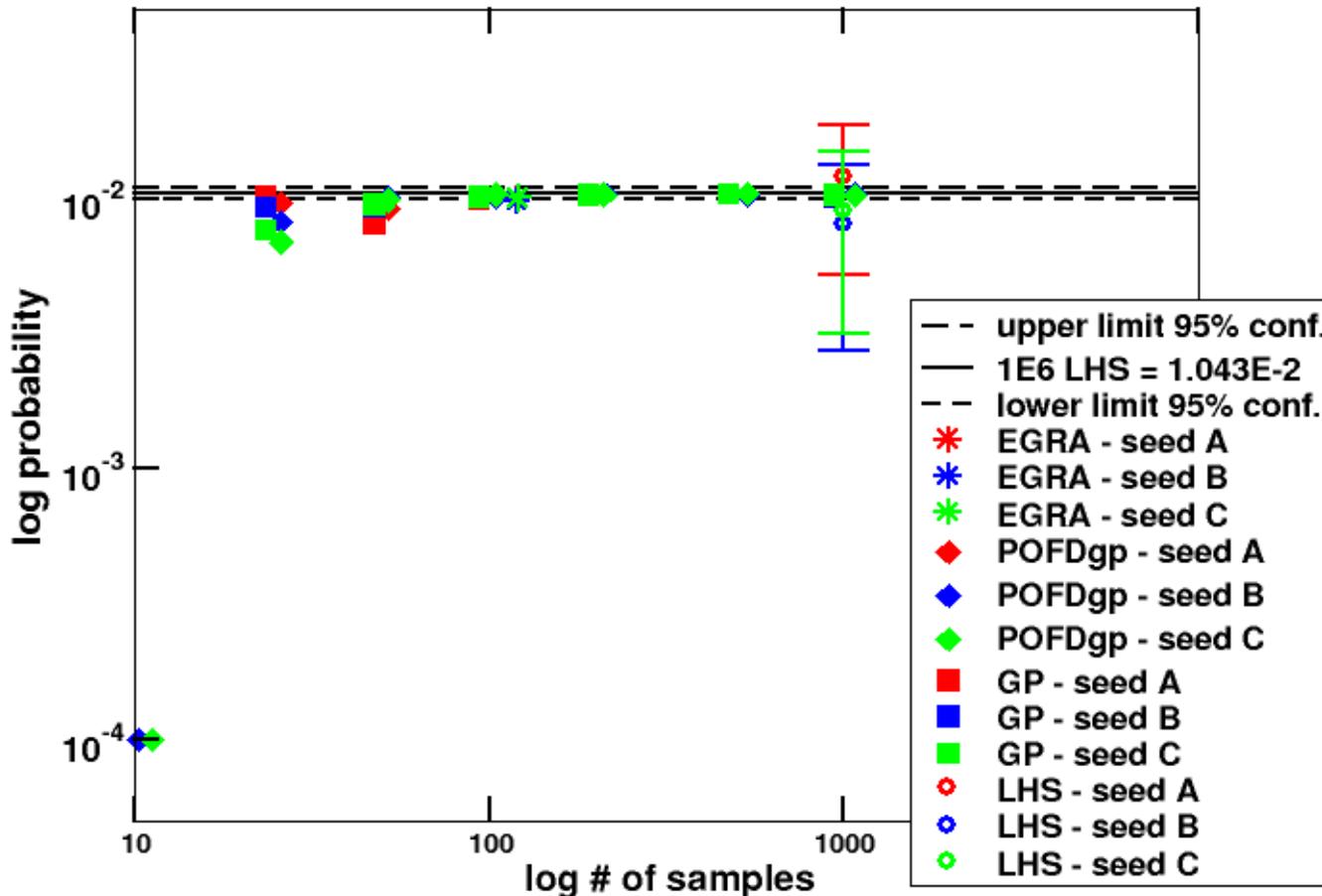
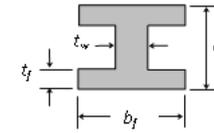
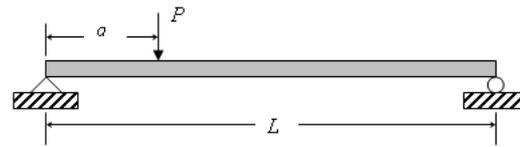
For this problem:

- **LHS-GPs** w/ 31 sims. perform less well than POFD or EGRA, having low accuracy and repeatability. But good precision and accuracy are obtained at 100 sims. Average accuracy declines with more sims.
- **LHS** gives accurate point estimates and reliable, fairly small confidence intervals for all seeds A, B, C with  $5 \times 10^4$  and  $10^5$  sims. This equates to  $N \cdot p \geq 5$ .

# Test Problem 5

## 8D I-Beam problem

- 5 Uniform PDF input uncertainties
- $P_{fail} \approx 1E-2$



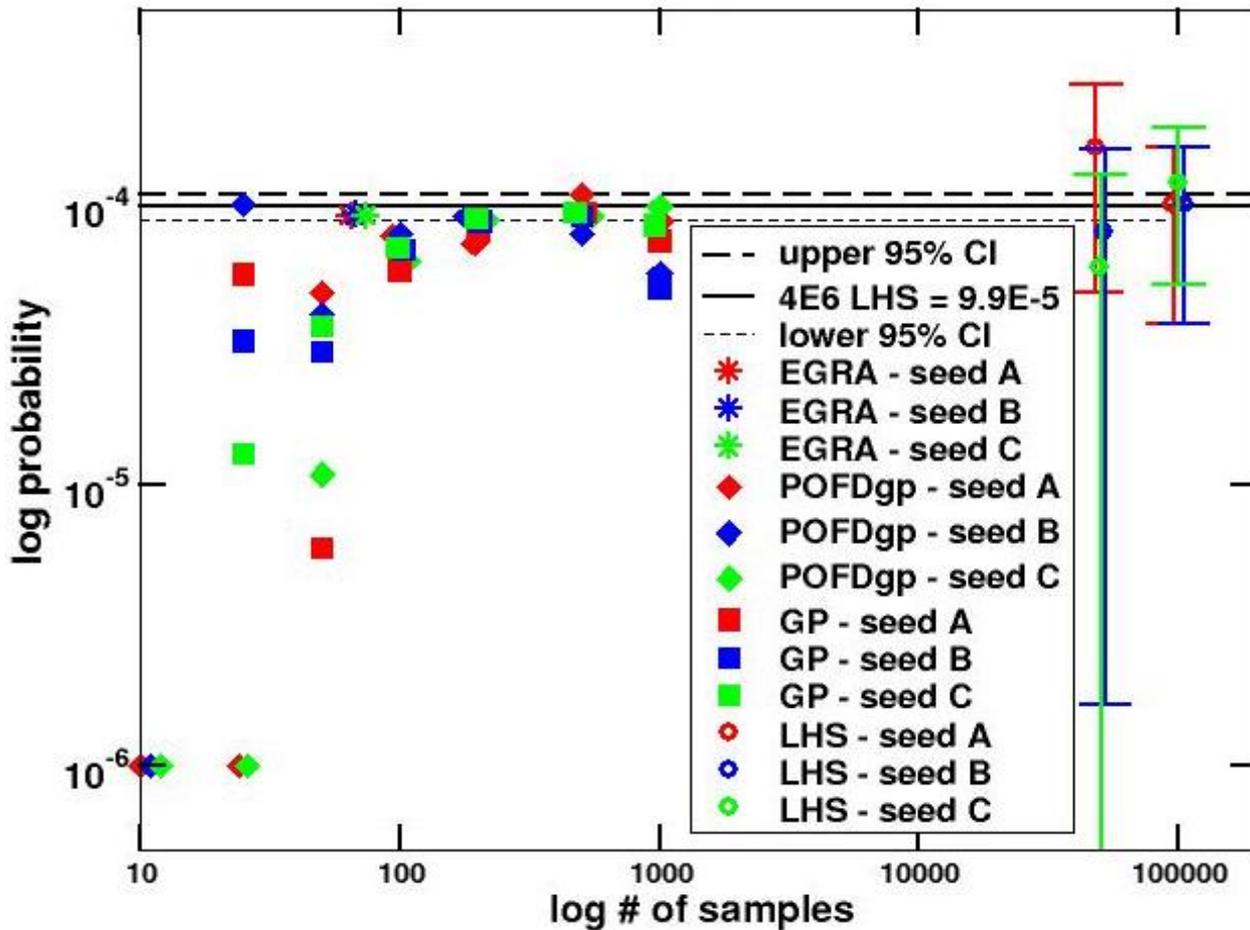
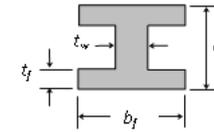
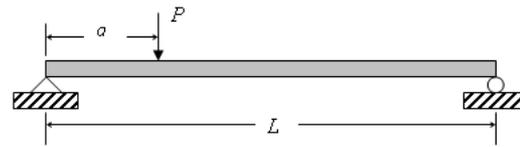
### For this problem:

- **EGRA** converged w/ 3-seed average of 105 sims., is fairly accurate but LHS-GP and POFD-GP w/100 sims. are both more accurate.
- **POFD-GP** and **LHS-GP** at 25 and 50 samples have better accuracy cost performance than EGRA. **LHS-GP** retains better performance out to 500 samples. **POFD-GP** retains better performance out to 1000 samples.
- **LHS** gives accurate point estimates and fairly small confidence intervals for seeds A, B, C with  $10^3$  sims. This equates to  $N \cdot p \geq 10$ .

# Test Problem 6

## 8D I-Beam problem

- 5 Uniform PDF input uncertainties
- $P_{fail} \approx 1E-4$



### For this problem:

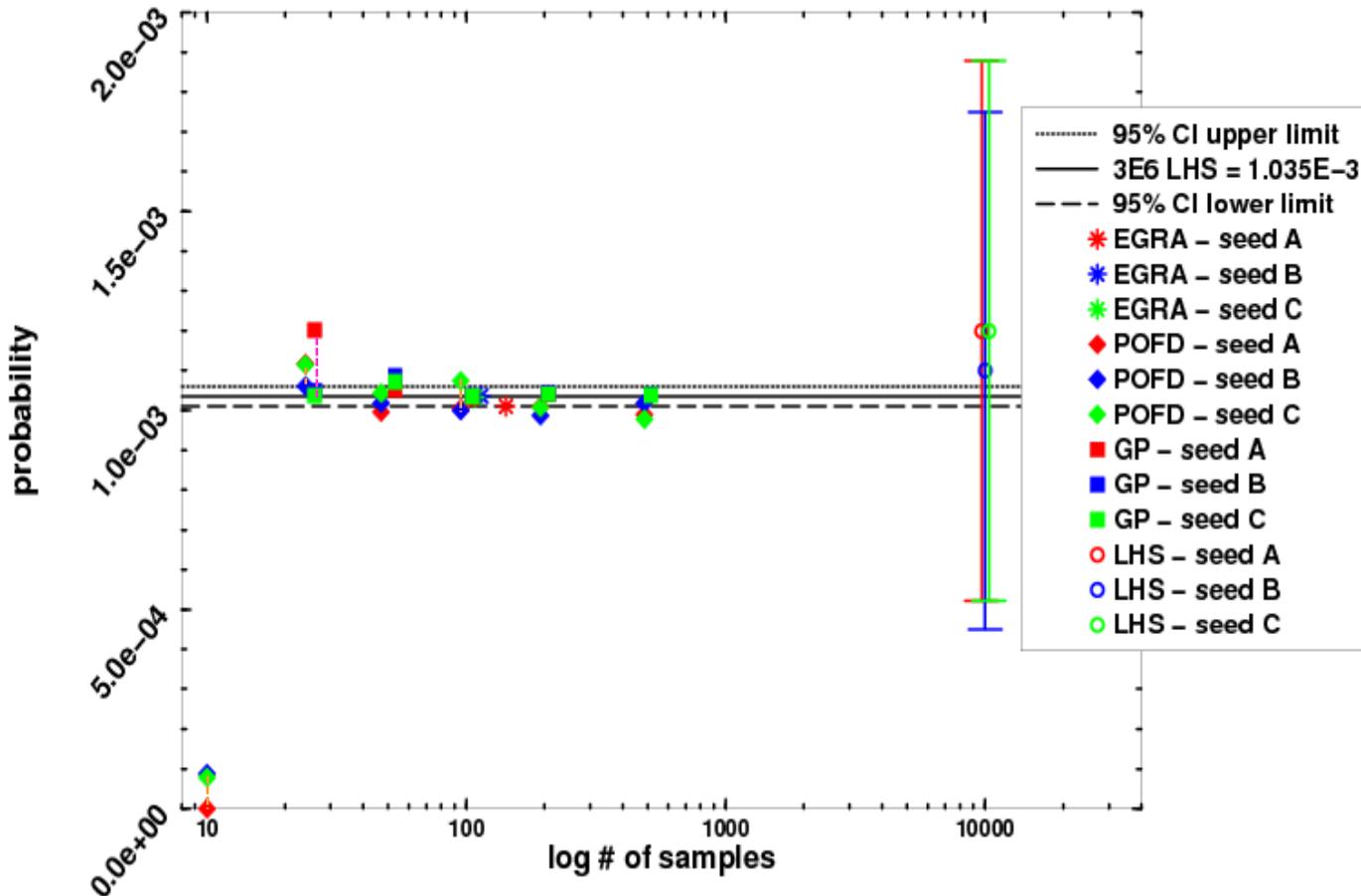
- **EGRA** converged w/ 3-seed average of 70 sims., less than for 1e-2 version of 8D problem. EGRA is signif. more accurate than LHS-GP and POFD-GP w/100 sims.
- **POFD-GP** and **LHS-GP** perform fairly similarly out to 1000 samples and never achieve better accuracy cost performance than EGRA.
- **LHS** does not achieve reasonably accurate point estimates and small confidence intervals for seeds A, B, C until  $10^5$  sims. This equates to  $N \cdot p \geq 10$ .

# Test Problem 7

## 9D Steel Column Failure problem



- 9 Uniform PDF input uncertainties
- $P_{fail} \approx 1E-3$

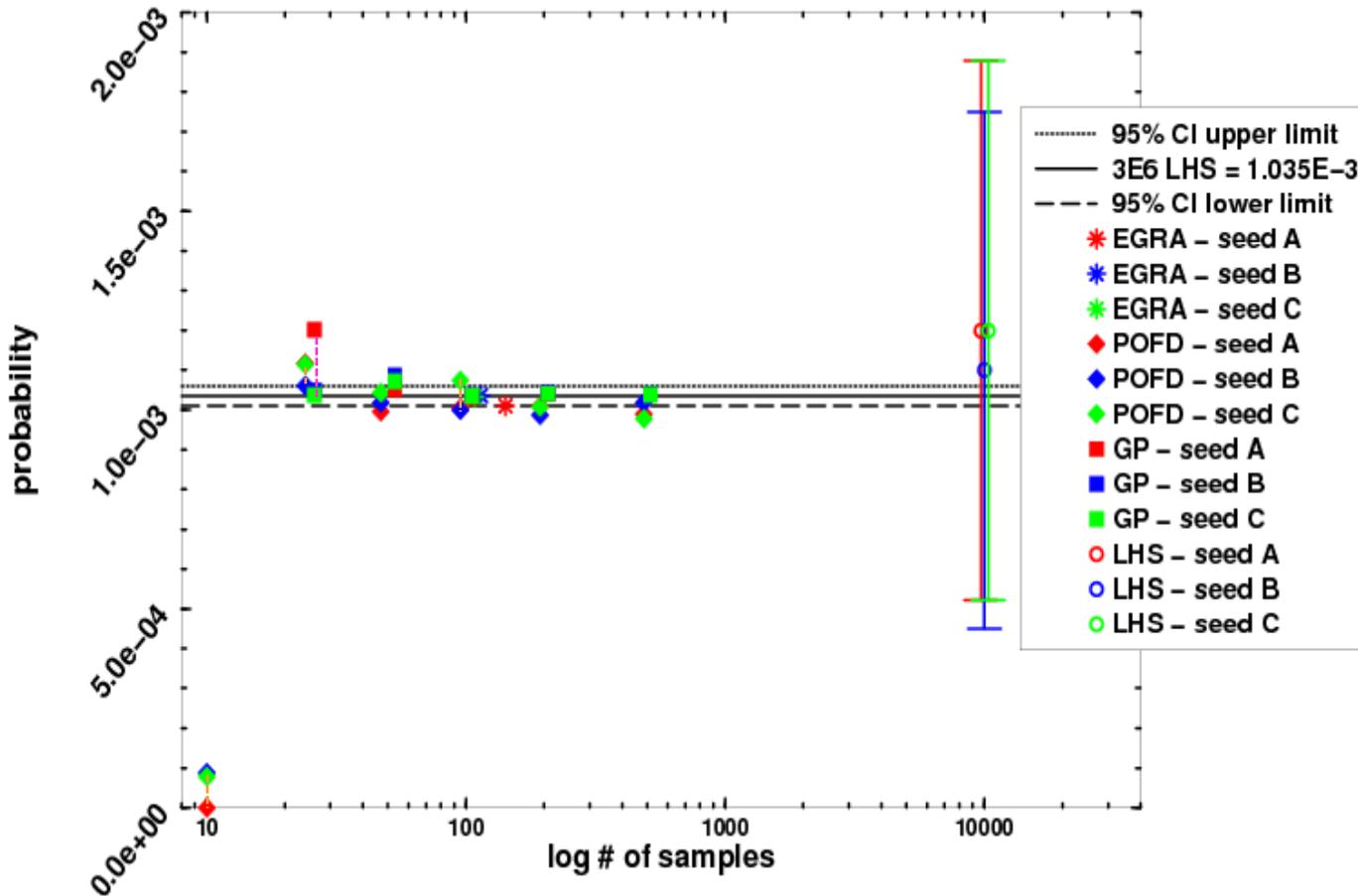


### For this problem:

- **POFD-GP** and **LHS-GP** performed comparably, achieving reasonable accuracy and precision with as little as 25 samples—indicating that the function is probably only mildly nonlinear over the UQ space. Both methods achieved high accuracy and precision for  $\geq 50$  sims.
- **EGRA** req'd. 142, 114, 108 sims. to converge for seeds A, B, C respectively, giving high accuracy and precision even with the highly varying # of samples to convergence.

# 9D Steel Column Failure problem

- 9 Uniform PDF input uncertainties
- $P_{fail} \approx 1E-3$



## For this problem:

- **LHS** requires 2 orders of magnitude more sims. for a reasonable expectation of reliable 95% conf. intvls., per rule of thumb  $N \cdot p \geq 10$ .

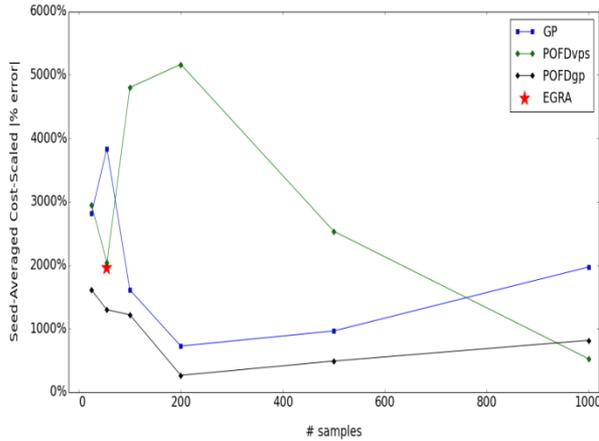
# Method Performance Metrics



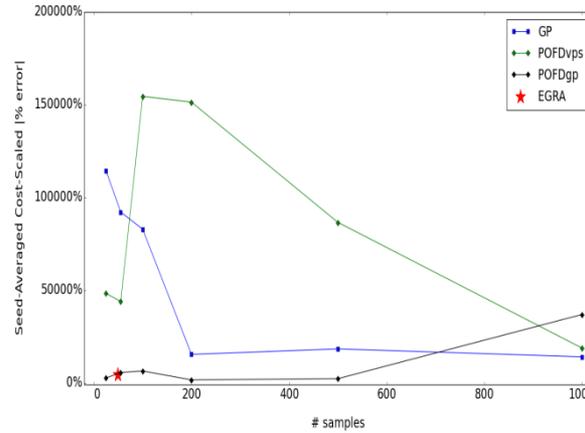
- 
- **Cost-Scaled Average Error over multiple trials**
    - Multiply each result's  $|\%error|$  by # of samples
      - this accuracy-cost measure accounts for # samples
    - If # of samples is doubled and the error drops commensurately by a factor of 2, then the results have the same cost-scaled error score
    - If same error occurs at  $N_1=10$  samples and  $N_2=20$  samples, the 20-sample result is  $\frac{1}{2}$  as cost-efficient, has a 2X cost-scaled error score
    - Allows comparing accuracy-cost performance for slightly different #s of samples  $N_i$ , and combining/averaging performance over multiple sample numbers, e.g. a range  $N_i = 25, 50, 100, 200, 500, 1000$
    - Also average over 3 stochastic realizations for each mthd. at each  $N_i$ .
  - **Cost-Scaled Average Error with 10X penalty on under-prediction**
    - Under-predicting a failure probability by a given error magnitude  $|e|$  is treated as far worse than over-predicting failure probability by the same magnitude
    - For these methods and tests, **penalized scores correlate highly with non-penalized scores**; show only non-penalized rankings here.

# Cost-Scaled Error Results

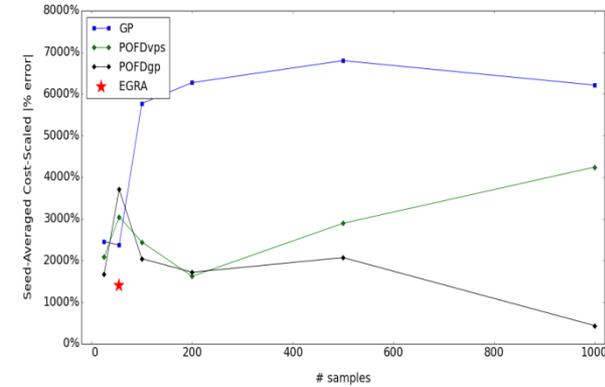
2D noisy Herbie w/PoF =  $\sim 1.51e-2$



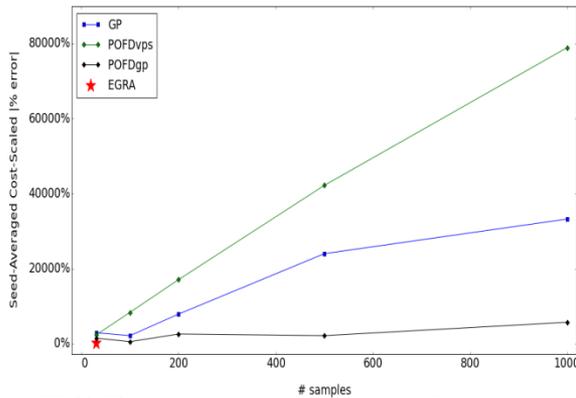
2D noisy Herbie w/PoF =  $\sim 1.02e-4$



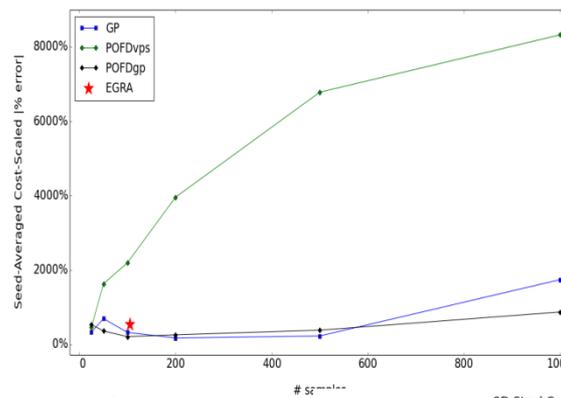
2D Vibration Amplitude problem w/PoF= $1.945E-2$



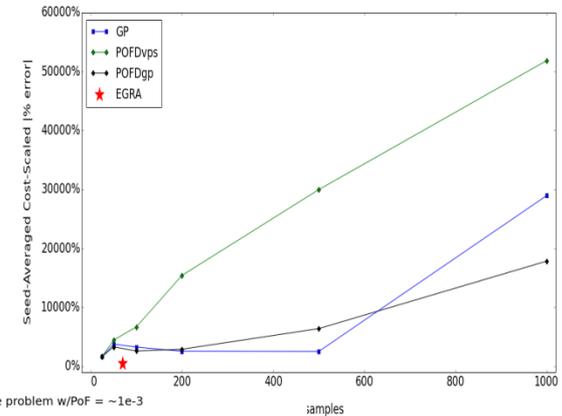
5D Circuit Problem w/PoF= $\sim 1e-4$



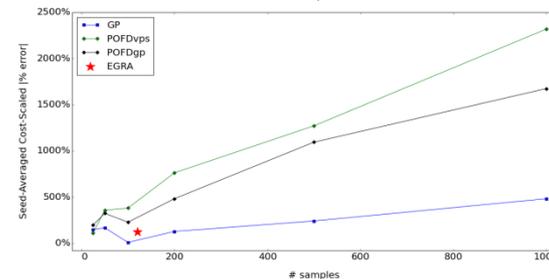
8D I-beam w/PoF =  $\sim 1e-2$



8D I-beam w/PoF =  $\sim 1e-4$



9D Steel Column Failure problem w/PoF =  $\sim 1e-3$

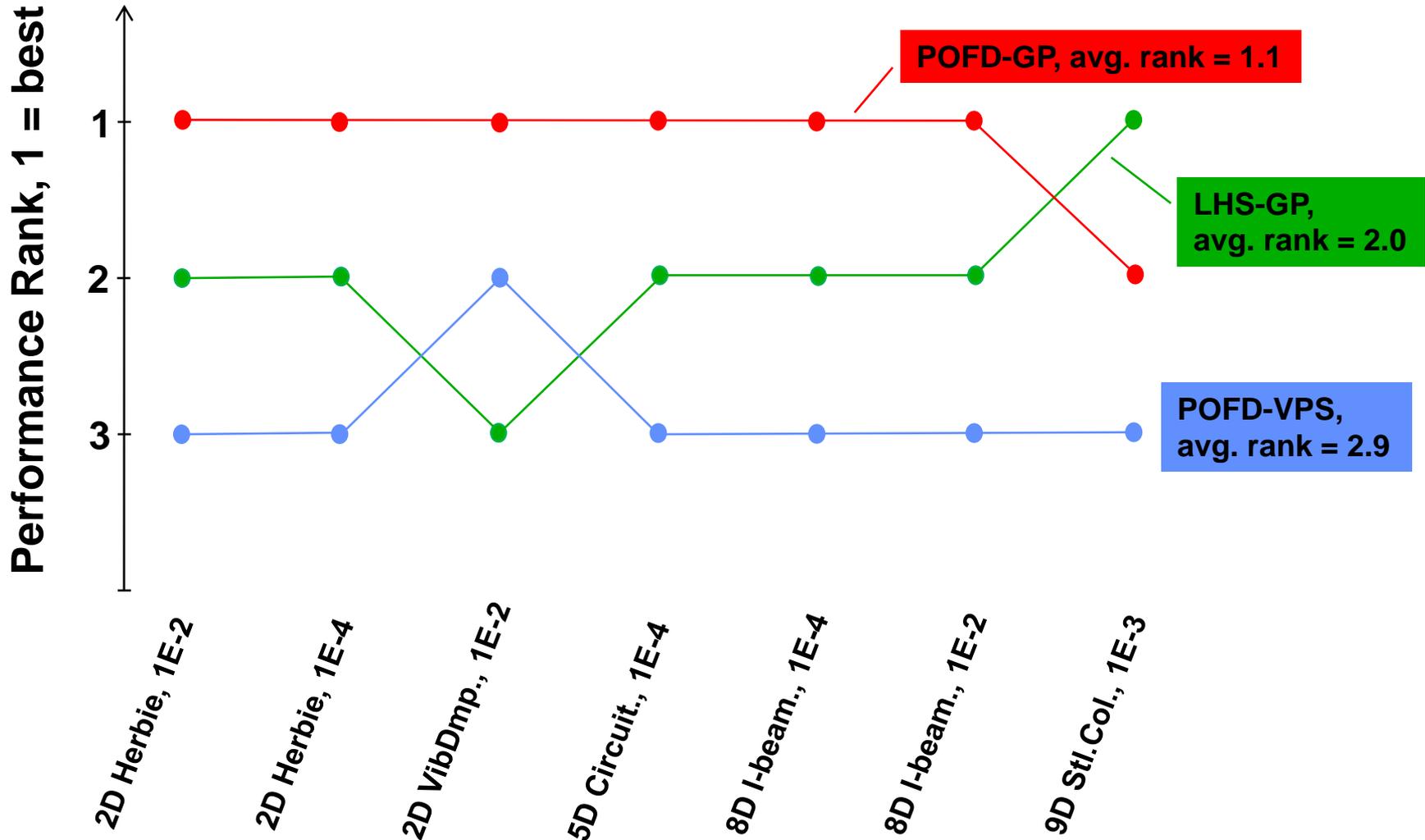


- **EGRA converges with less than 150 samples in all 7 cases and does competitively well in 6 of the 7. But EGRA only has the lowest accuracy cost point in 2 of the 7 problems. POFD-VPS has the least good performance on average.**

# Method Performance Rankings (not incl. EGRA)

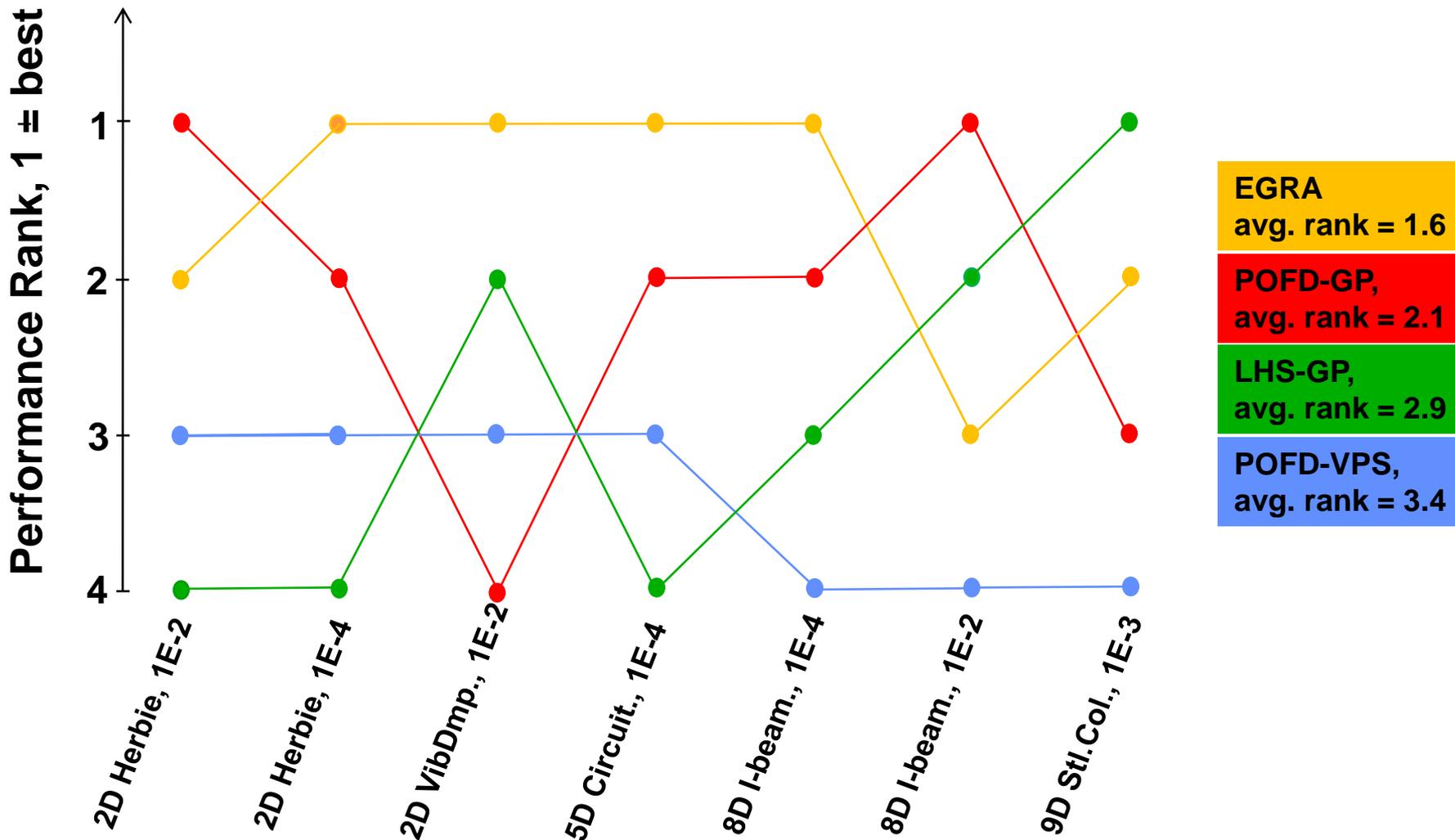
for each problem average over

3 realizations at  $N_i \approx 25, 50, 100, 200, 500, 1000$  samples



# Method Performance Rankings

at ~same # samples EGRA took to converge  
(averaged over 3 realizations from each mthd.)



# Summary Observations, Discussion, and Recommendations

---



- Among the non-EGRA methods, POFDgp usually performed best, followed by non-adaptive LHS-GP, then by adaptive POFD-vps.
- Comparing these methods to EGRA at the ~same #samples EGRA required to converge, EGRA was most accurate in 4 of the 7 problems, POFD-gp in 2/7, and LHS-GP in 1/7. But in 4 of the 7 problems, other methods had better results with less samples than EGRA—3 times for POFD-gp and twice for LHS-GP.
- EGRA converged with ~ 30 – 142 samples (loosely correlated with problem dimension but not with Pfail magnitude!)
- EGRA convergence was often sooner than might be desired; more accuracy could often be obtained with the other methods at the cost of more samples. Often the additional sampling cost was more than justified by the amount of accuracy improvement, showing a better cost-accuracy effectiveness per-sample than EGRA. EGRA has the best accuracy cost in only 2 of the 7 problems.

# Summary Observations, Discussion, and Recommendations

---



- Overall, EGRA and POFD-GP were the best performers here, with neither clearly better than the other.
- This brings to light the promising potential of the new POFD-GP method, which has only been under development for a few years, many less than EGRA.

# Summary Observations , Discussion, and Recommendations

---



- All the methods exhibited significant stochastic variability of cost-accuracy performance and the majority of results from all methods under-predict the true failure probability.
- ➔ Robust Error Estimation needs to be developed for the non-MC methods.
  - Perhaps can use variance from multiple realizations to base error estimates on.
  - This will increase their cost significantly, but they will likely still have significant accuracy-cost advantage vs. Monte Carlo methods.

# Summary Observations, Discussion, and Recommendations

---



- LHS-GP is a non-adaptive method and still did relatively well here.
- This indicates that when multiple failure probabilities are to be estimated to prescribed accuracies when multiple output quantities are involved like pressure and temperature, and/or multiple response threshold levels are to be investigated, then LHS-GP would be more cost effective than running an adaptive method for each of the (multiple) analysis cases.
- Also, EGRA and POF-darts are adaptive, so
  - inherently sequential algorithms; little sampling parallelization possible
- The non-adaptive LHS and LHS-GP results can be re-processed to get other characteristics of response such as mean, standard deviation, and the full PDF of response.
- Only non-adaptive methods can simultaneously yield other such characteristics of response.

# Summary Observations, Discussion, and Recommendations

---



- Failure probability estimates from LHS samples alone (without interpolation w/GP) were found to be non-competitive in terms of accuracy cost, but provided reliable confidence-interval error bands for failure probability magnitudes  $10^{-2}$ ,  $10^{-3}$ ,  $10^{-4}$  and  $5 \leq \#samples \times P_{fail} \leq 20$ .
- Other statistical (non-adaptive) sampling approaches like Halton and Hammersley Quasi-Monte Carlo (QMC) sequences, Centroidal Voronoi Tessellation (CVT), and Orthogonal Arrays (OAs), have been found to often be more efficient than LHS for statistical estimates of mean, standard deviation, and failure probabilities.
  - may be worth a follow-on study on the test problems here and others (using standard confidence-interval error band formulas or CI estimates from replicated sampling)

# Summary Observations, Discussion, and Recommendations

---



## Interpolation Alternatives to GP

- **Several studies in the literature suggest that other interpolation approaches like Radial Basis Functions may perform better than GPs when applied to a set of randomized sample points like LHS.**
  - **May be worth a follow-on study with variants of the present suite of test problems and others.**
  - **But GP has a large advantage of providing local and potentially global error estimation.**

## Statistical (Non-Adaptive) Sampling Alternatives to LHS

- **The literature and past experiences suggest that other non-adaptive sampling methods besides LHS (such as Halton and Hammersley QMC, OAs, and CVT in combination with GP or other interpolators are sometimes more cost effective than LHS.**
  - **May be worth a follow-on study with variants of the present suite of test problems and others for a more definitive quantification.**

# Summary Observations , Discussion, and Recommendations

---



- **Literature reviews suggest that methods like Polynomial Chaos, Stochastic Collocation, Compressed Sensing, may perform better in various situations.**
  - **It would be useful to compare accuracy cost of these other methods on the present suite of test problems and others.**
- **Robust Error Estimation with any of these methods would make them more trustworthy, relevant, and useful like the “old [expensive] standard” Monte Carlo.**