ELM Higher Order Indices

John Darges

1 Genetic Oscillator

The genetic oscillator takes in 16 inputs. We use an ELM surrogate, trained on 3000 training points, with 900 neurons. In this case, regularization is not necessary. Here, the relative $l2-$error between the surrogate and true model is approximated as 0.038. In Figure 1, a comparison is given between the analytically computed total Sobol' indices of the ELM surrogate and the total indices approximated by Monte Carlo sampling of the true model (These are taken from Mike's paper).

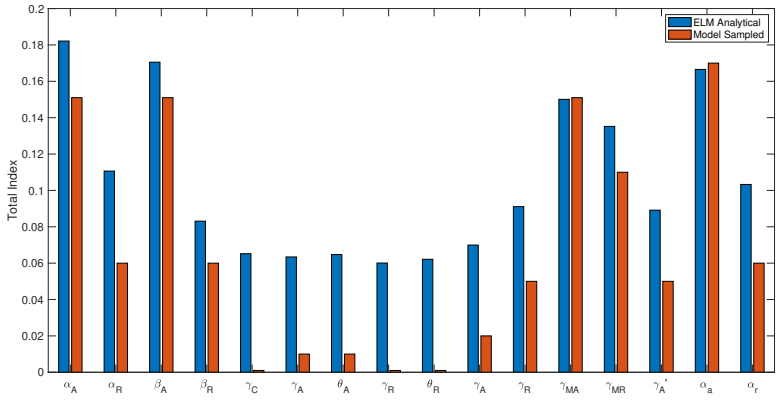

Figure 1: Left: Total indices of ELM surrogate (3000 training, 900 neurons). Right: Total indices computed via sampling model.

Note that we have a large number of indices and that many are of comparable magnitude. Assuming the approximated indices offer an accurate picture, the ELM surrogate overestimates less important indices and underestimates more important indices. The disparity between the ELM indices and the true indices is more dramatic for very unimportant indices. All of the ELM indices take values around 0.07 or greater, even when the true indices are practically negligible in value. However, the ELM indices do follow a similar pattern as with the true indices in terms of ordering.

We can also compare to first order indices of the genetic oscillator. Sampled indices were obtained via a separate computation using 1×10^5 samples.

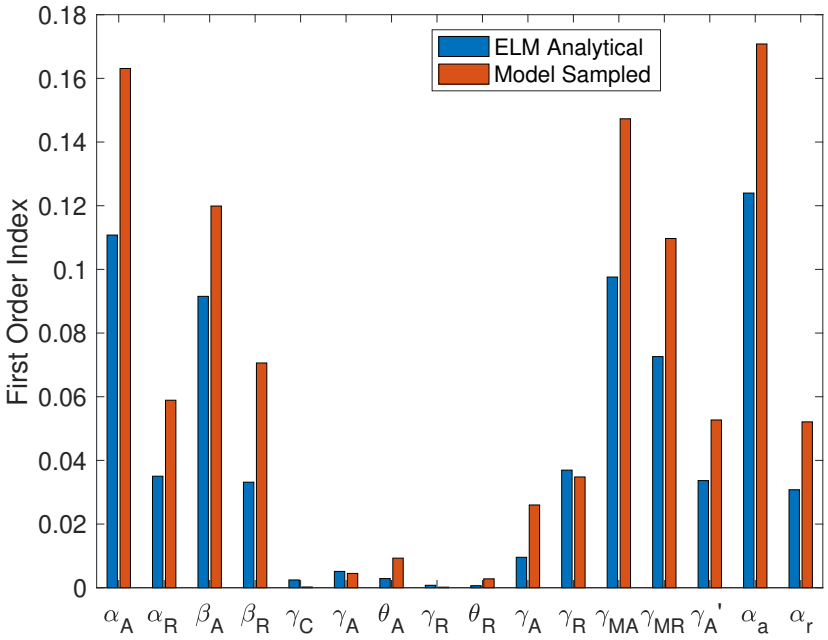

Figure 2: Left: First order indices of ELM surrogate (3000 training, 900 neurons). Right: First order indices computed via sampling model.

2 8-dim G-function

Let us compare the the case of the genetic oscillator to that the G-function. The ELM surrogate uses 3000 training points and 900 neurons. We require regularization in the least squares problem and use 1×10^{-4} as the Tikhonov parameter, as suggested by the L-curve method. The total indices of the G-function have much more distinct values, with the the most important indices taking much larger values than the unimportant indices.

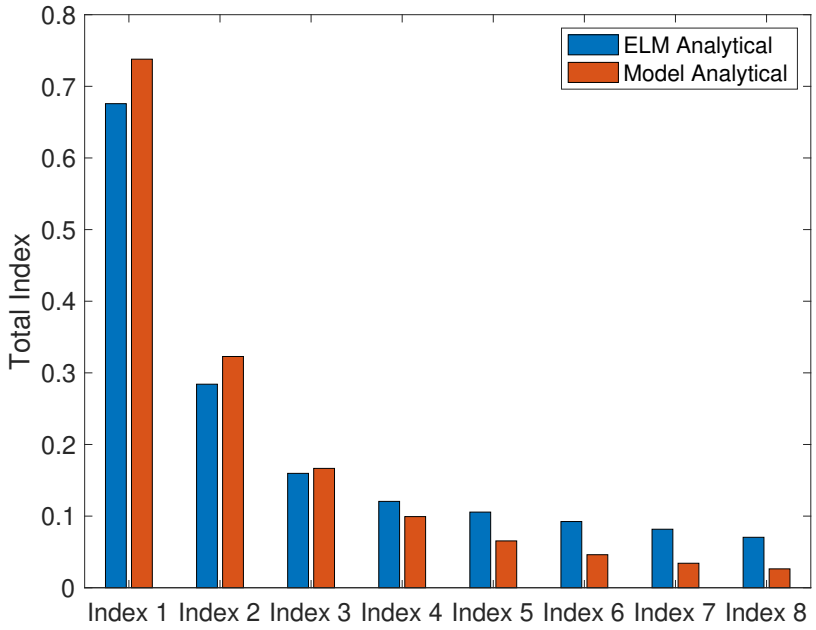

Figure 3: Left: Total indices of ELM surrogate. Right: Total indices of model.

Still, in this case, the ELM indices take values around 0.07 or greater. But, in the case of the G-function, even the least important true indices of the G-function take values that are non-negligible (when compared to the genetic oscillator, which has indices that are essentially zero). So, the disparity between the ELM indices and the true indices appears less pronounced. This effect appears exclusive to the total indices. Below are the regular first order indices of the same ELM surrogate compared to the true regular indices of the model.

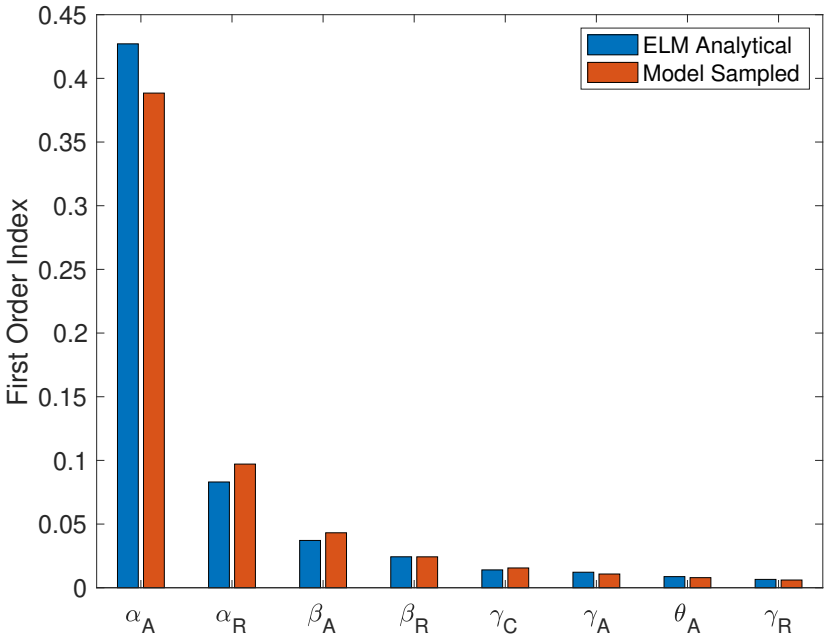

Figure 4: Left: Regular indices of ELM surrogate. Right: Regular indices of model.

The pattern of overestimating unimportant indices and underestimating important indices does not seem to appear when looking at the first order indices. This could suggest that the ELM surrogate inherently overestimates the total interactions of less important variables.

3 4-dim G-function

Finally, we can compare with the 4-dimensional G-function. The true indices of this model all possess a degree of importance, so we should expect the effect of overestimation of less important indices to be less pronounced. This surrogate takes a relative error of 0.1522.

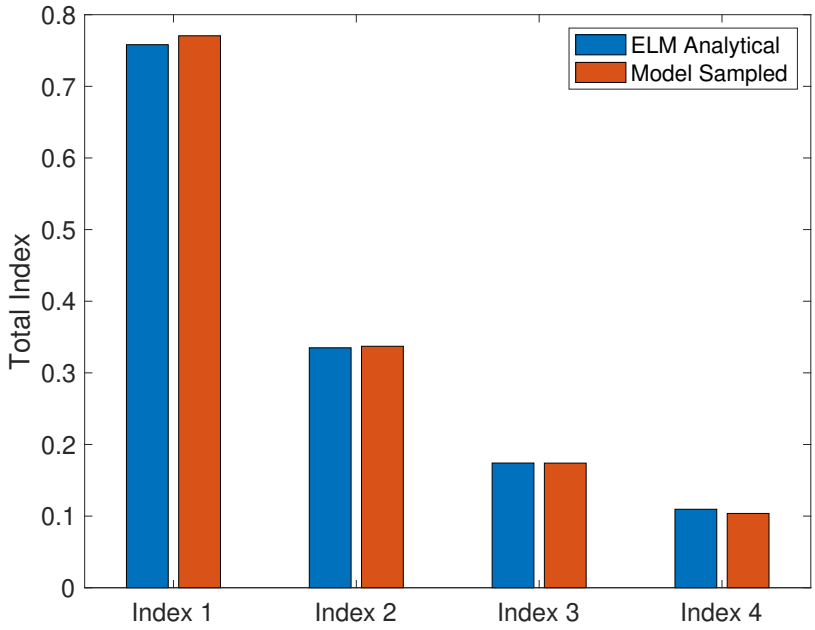

Figure 5: Left: Total indices of ELM surrogate. Right: Total indices of model.

4 Discussion

It is not necessarily surprising that ELM surrogates for higher dimensional models might overestimate unimportant indices. The question is whether this is caused by higher dimensional models requiring more training points or whether this is a problem inherent to ELM, that unimportant variables naturally get overemphasized in the total indices surrogate. Let's look at the formula for the ELM surrogate

$$\hat{f}(x) = \sum_{i=1}^{N} \beta_i e^{\left(b_i + \sum_{k=1}^{d} W_{k,i} x_k\right)}$$

where we consider each summand in the formula as a basis function with basis coefficients given by β. With the weight matrix W being sampled from a normal distribution, all the weights will be nonzero, so every input variable is represented, to varying degree, in every basis function. It could be that many minuscule interactions accumulate over the summation of all the basis functions. In the case of polynomial chaos, the basis polynomials represent all possible interactions between different variables up to a certain degree. It might make sense to try to impose an idea of modes in ELM, to set some of the weight entries to zero so that not all variables are represented in every basis function.

A way to do this is to modify the sampling distribution. We can create a probability distribution where each weight has probability p to be 0 and probability $1 - p$ to be selected from the standard normal distribution. An ELM surrogate created in this fashion, with $p = 0.5$, has relative error 0.014.

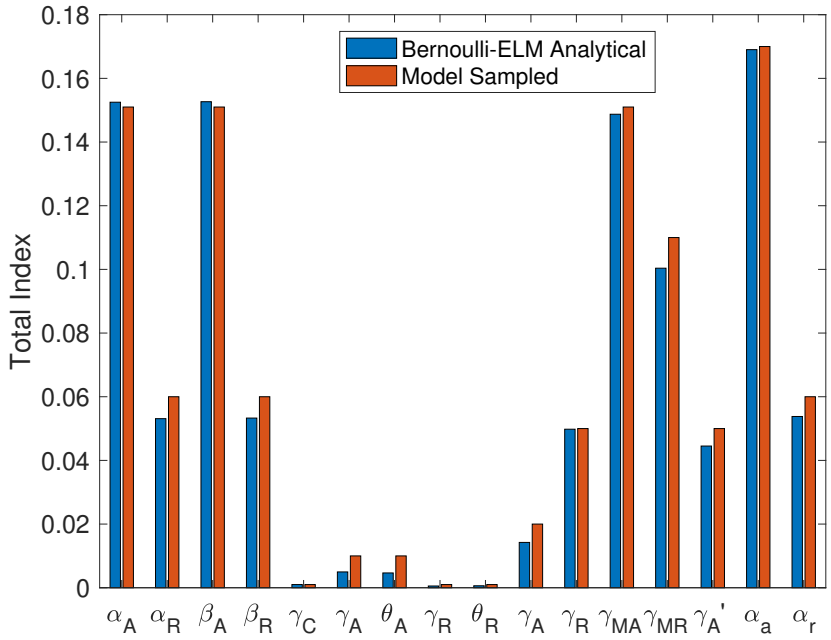

Figure 6: Left: Total indices of Binomial-ELM surrogate. Right: Total indices of model.

5 Examining Higher Order Indices

In the genetic oscillator example, we expect that higher order interactions should not occur. Therefore, higher order regular Sobol' indices should be near zero. We can compare the second order indices (those corresponding to interactions between two variables) of different ELM surrogates, one using standard normal sampling and one using Bernoulli-normal sampling with $p = 0.5$. Below are histograms that display the distribution of values taken by the second order indices.

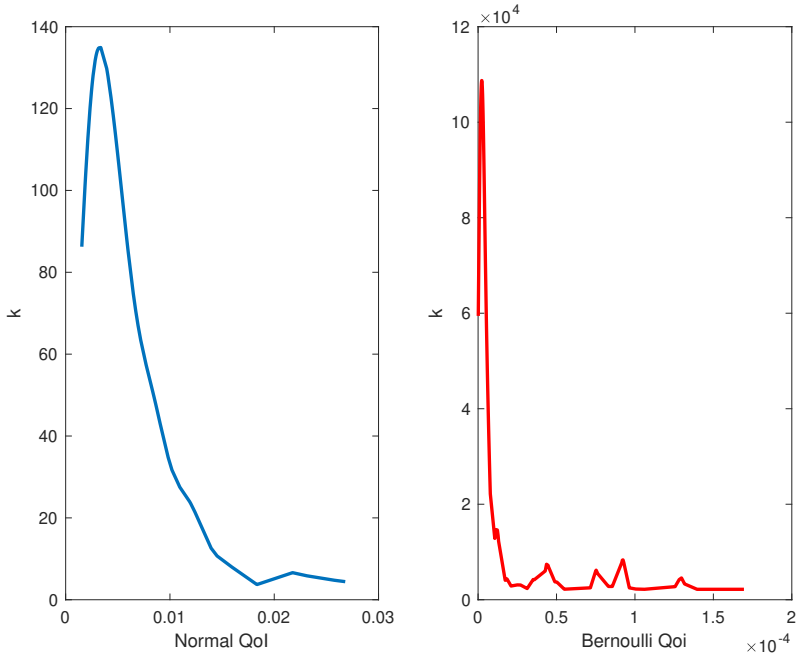

Figure 7: Left: 2nd order regular indices using normal sampling. Right: 2nd order regular indices using Bernoulli-normal sampling.

We can also compare these second order indices with those given when using a sparse-PCE surrogate. The approximation uses an order of accuracy of 5 (36097 nodes).

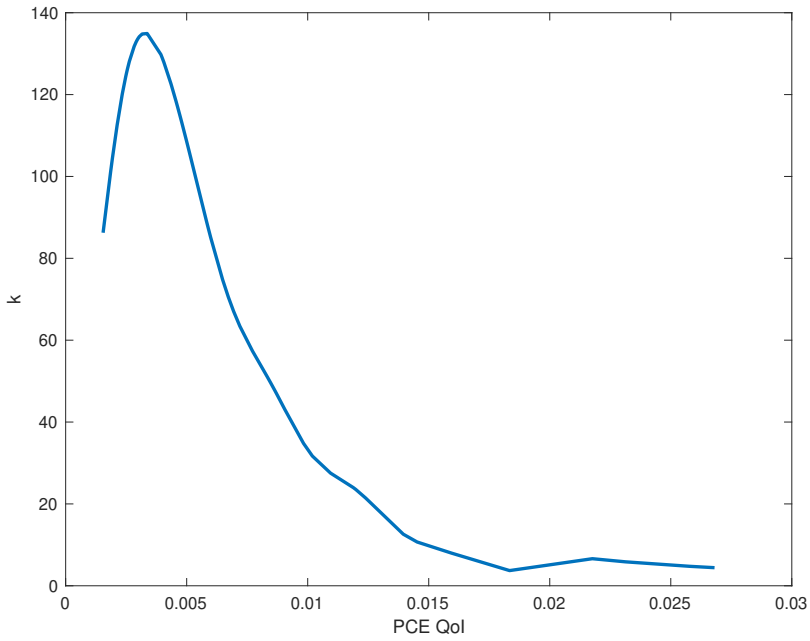

Figure 8: 2nd regular indices when using a PCE surrogate.

6 Tuning p

This brings forth the question: if sampling weights for ELM using the Bernoulli method, what is the best p to use? We examine this for three different model, training surrogates using different values for p. Each data point is taken as the average of ten models. Each model used separately sampled training sets.

6.1 Some Variable Interactions

The 8 dimension G-function is an ideal case for standard normal sampling because all types of variable interactions occur within the model. Therefore, we expect the best results when our sampling method is closest to the standard normal method (i.e., when $P(x = 0) = 0$).

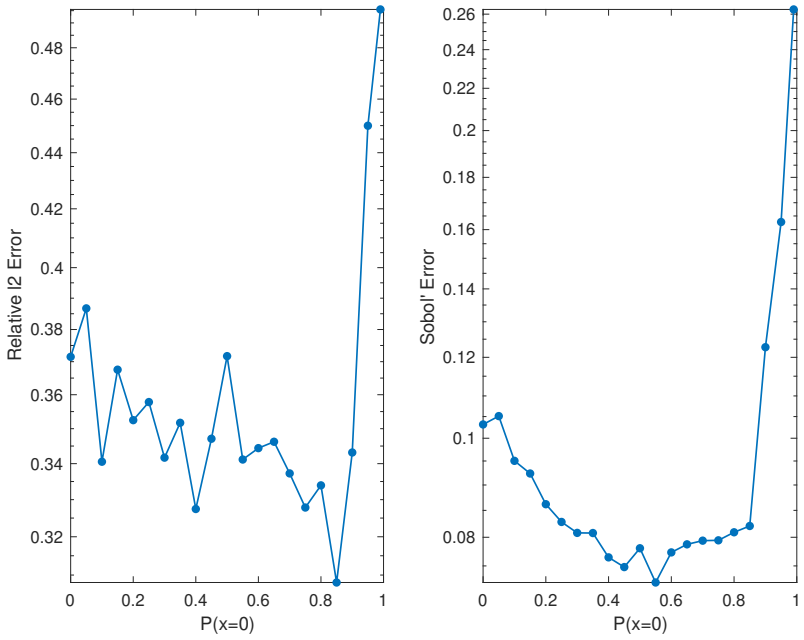

Figure 9: 8 dimensional G-function surrogates using different probabilities, trained with 6000 training points and 1800 neurons.

We can compare this to a hybrid case, which combines the additive function and the G-function

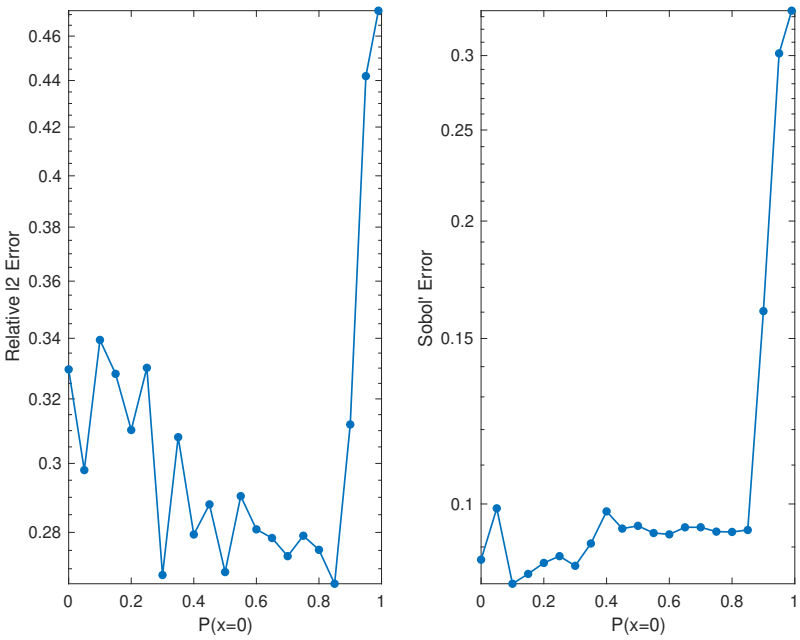

Figure 10: Surrogates for $f(x) = g_4(x_1, x_2, x_3, x_4) + \sum_{i=4}^{8} \sin(x_i)$ using different probabilities, trained with 6000 training points and 1800 neurons.

6.2 Few or No Variable Interactions

In the case of the genetic oscillator, we expect that an optimal p will be much closer to 1 since there appear to be few variable interactions.

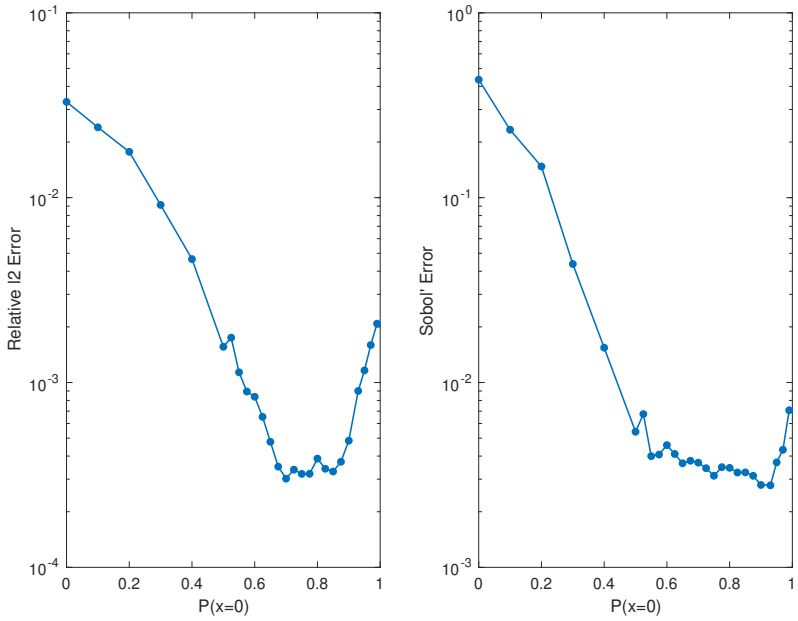

Figure 11: Genetic oscillator surrogates using different probabilities, trained with 3000 training points and 1000 neurons.

We look at the case of an additive function, $f(x) = \sum_{i=1}^{16} \sin(x_i)$, where all variables contribute equally. The optimal p is very close to 1.

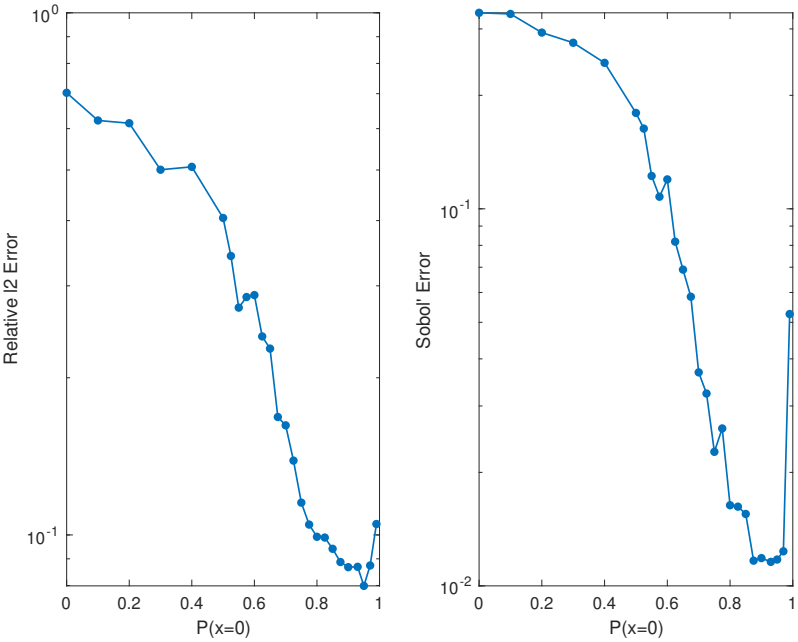

Figure 12: Surrogates for $f(x) = \sum_{i=1}^{16} \sin(x_i)$ using different probabilities, trained with 1000 training points and 330 neurons.

6.3 Significant Variable Interactions

We consider a function where all components of the ANOVA decomposition, except for the highest order component, are trivial. In this case, the function's first order Sobol' indices are 0 and all indices are 1. Due to the nature of this function, higher magnitude weight entries are favored, to the point that simply sampling from a standard normal distribution is insufficient at higher dimensions.

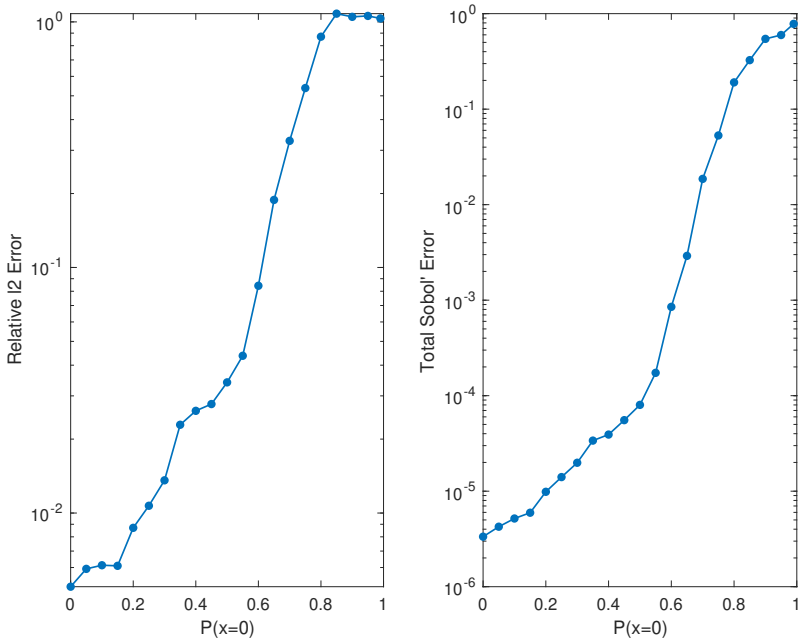

Figure 13: Surrogates for $f(x) = \prod_{i=1}^{4} \sin(x_i)$ using different probabilities, trained with 3000 training points and 1800 neurons.

Note: When sampling our weight matrix W using a Bernoulli distribution, we can compute the expected number of rows in the weight matrix with entries all zero, $N \cdot p^d$. Recall, the number of rows of W is the number of neurons, N, and the number of columns is the input dimension, d. We can then estimate, for a given p, when H is unlikely to have full rank. Despite this, we can still use Tikhonov regularization to avoid this issue.

7 Interactions Experiments

7.1 All Interaction Types

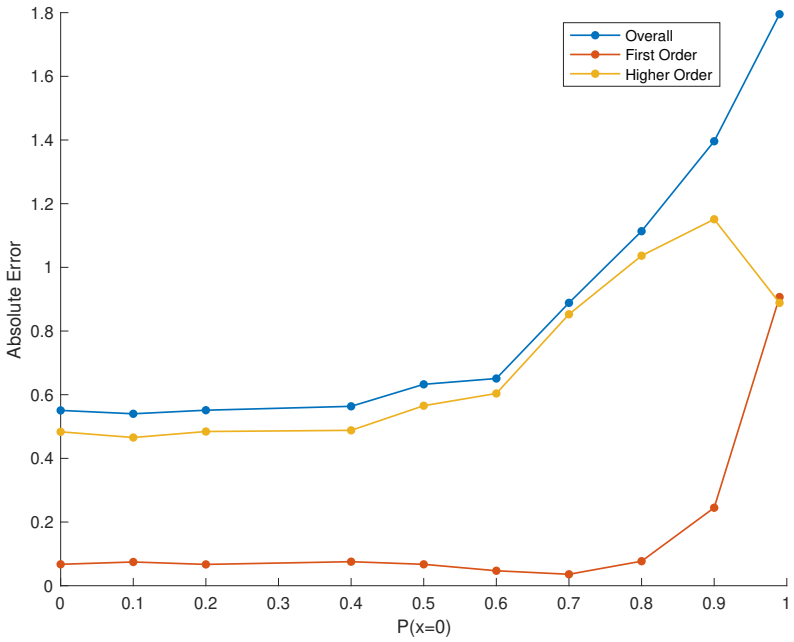

Figure 14: Surrogates for six dimensional polynomial with all types of interactions using different probabilities, trained with 1000 training points and 300 neurons.

7.2 No Higher Order Interactions

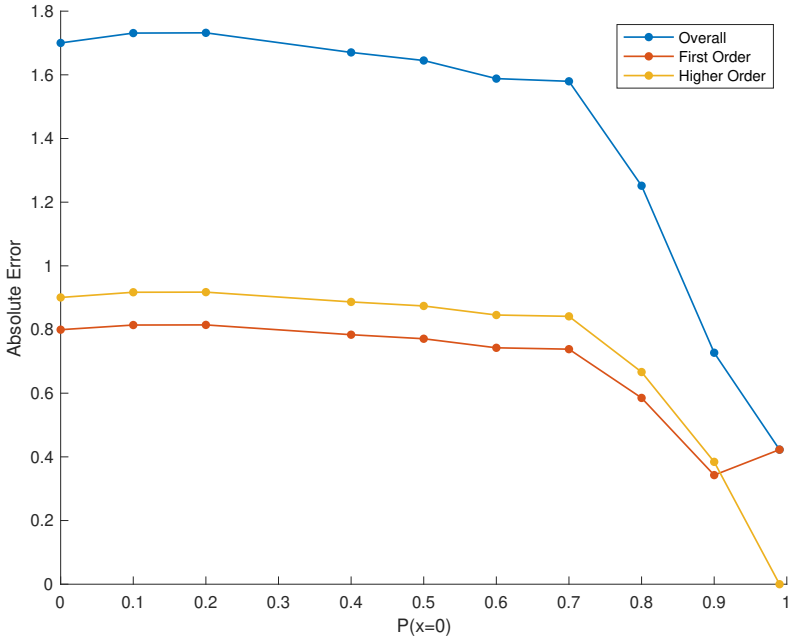

Figure 15: Surrogates for six dimensional polynomial with no higher order interactions using different probabilities, trained with 1000 training points and 300 neurons.

7.3 Only Higher Order Interactions

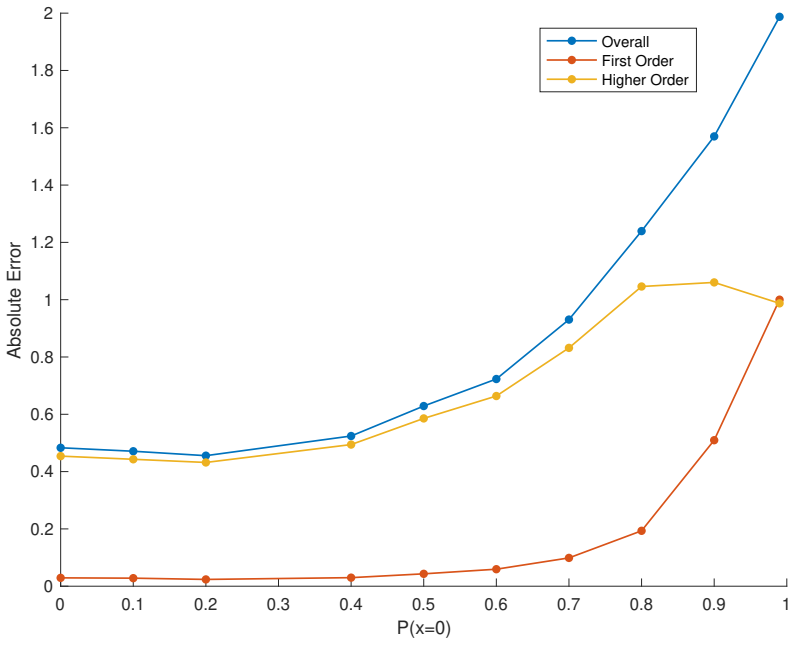

Figure 16: Surrogates for six dimensional polynomial with only higher order interactions using different probabilities, trained with 1000 training points and 300 neurons.