Surrogate-Based Optimization Using Machine Learning

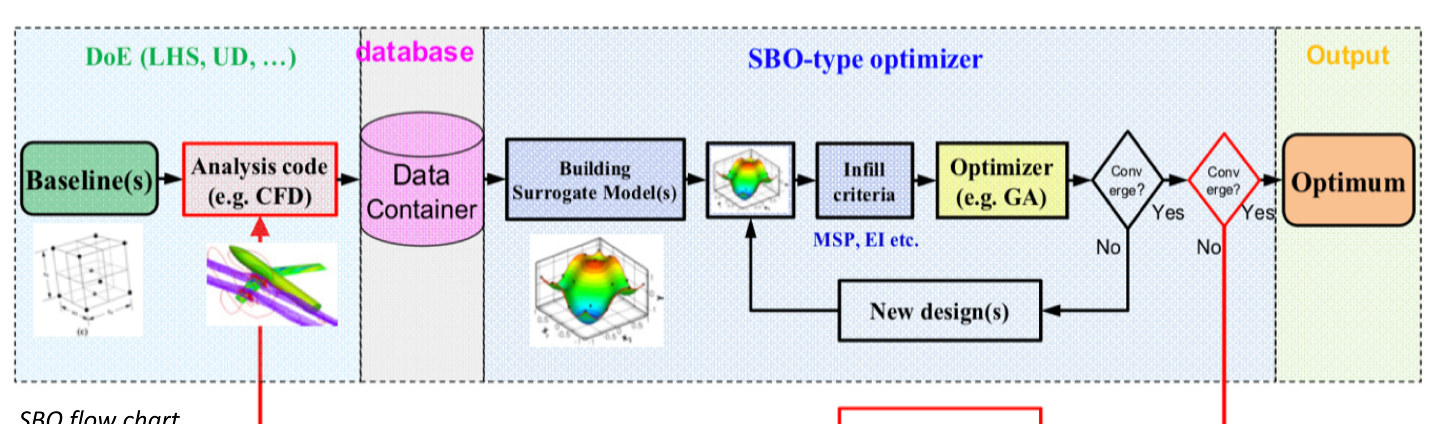
Mohammad (Jabs) Aljubran, Christian Tae

CS229A– Applied Machine Learning (Final Project)

Introduction

Using a solver, e.g. genetic algorithm (GA), to optimize over a computational fluid dynamics (CFD) numerical simulator is computationally expensive. Hence, surrogate-based optimization (SBO), seen in figure below, is a common approach to efficiently find the optimal solution. Machine learning (ML) techniques hold enormous potential in building models that provide accurate, fast, and computationally inexpensive proxies that feed the SBO loop and map the objective function with minimal sample points, allowing the optimization solver to locate the global optimum. This project will explore the viability of using ML techniques to build robust SBO proxies to optimize CFD reservoir well design. An SBO algorithm loop involves the following steps:

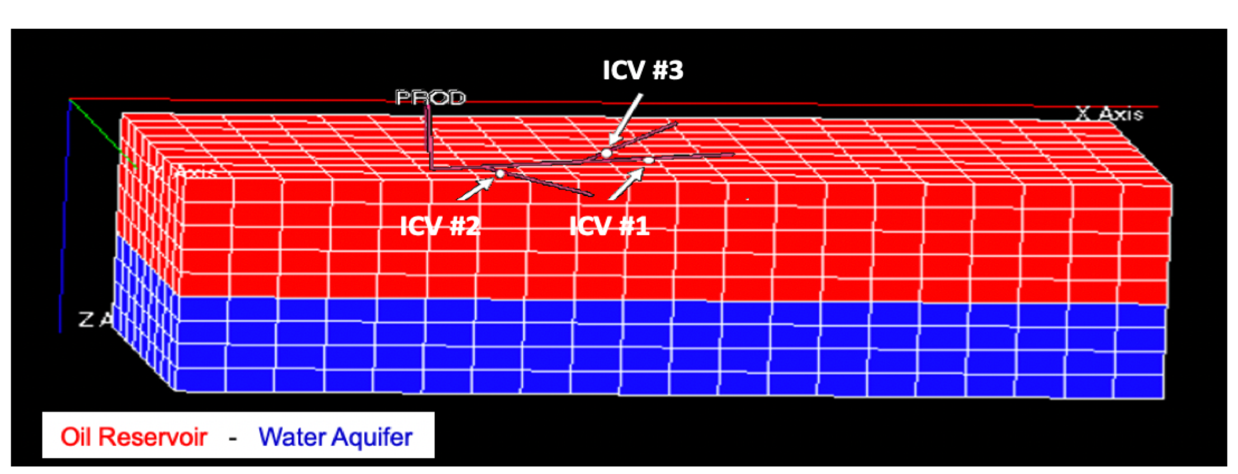
- Latin Hypercube Sampling (LHS) is conducted to create an initial sample
- Numerical simulator evaluates the function at these sample points
- ML approach is used to fit surrogate model (surface response)
- GA is run on the ANN proxy to locate the global optimum



SBO flow chart

Data Collection

A numerical CFD simulator is used to construct a reservoir model that simulates a homogenous oil reservoir with two-phase flow. A segmented, trilateral producer well is located at the grid center and completed across the reservoir top layer with three ICV devices (each holds 11 discrete settings from 0 to 10 indicating constriction area). The initial state of a cartesian reservoir model, shown in the figure below, is 100% oil saturation on top (red) and 100% water saturation on bottom (blue). A branched producer is located at the center with three ICV devices.



Numerical Reservoir CFD Model

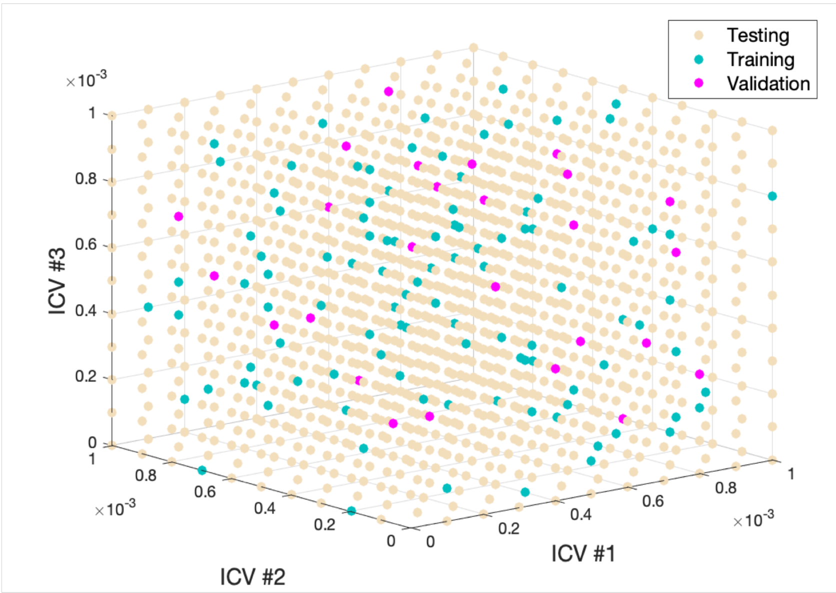
A total of 1,330 simulation runs are conducted, generating 667,660 temporal sample points. Below is a simplified tabulated example where the oil and water production of cases 1,2 and 3 are generated using the numerical simulator which then are used to train an ML model. The latter acts as a surrogate to predict the cumulative production at the cases 5,6, and 7 which are unseen by the ML model.

Variables Over Time (x_1)	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
ICV #1 (x_2)	10	8	7	0	3	2
ICV #2 (x_3)	10	10	5	6	4	0
ICV #3 (x_4)	10	0	7	2	5	1
Oil (y_1) and Water (y_2) Production	Given Data			Predicted Output		

Data Partitioning

LHS is commonly used in Monte Carlo simulation sampling to reduce the variance given a fixed number of sample points from a multidimensional domain. Compared to random sampling, LHS provides a more representative space sample which is critical to fit a representative SBO surface response.

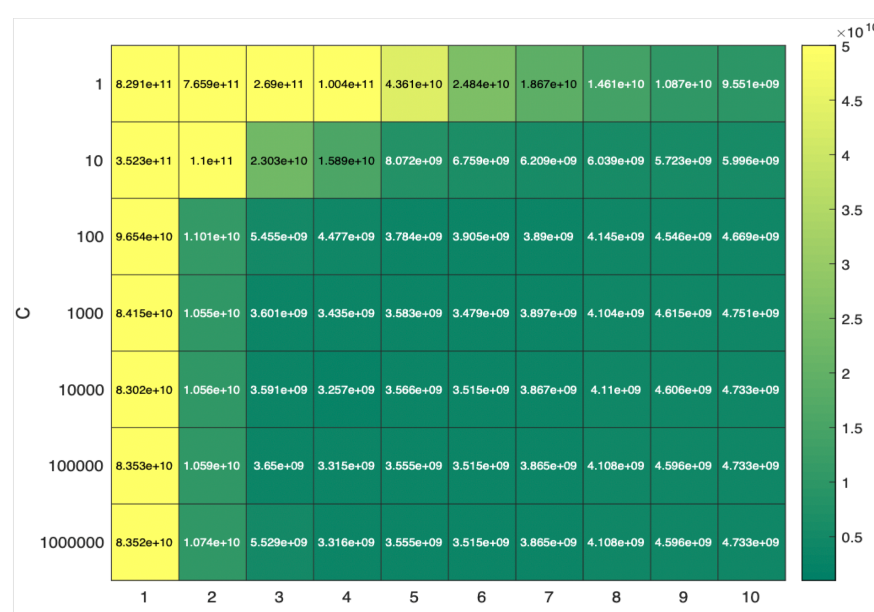
For SBO purposes, the training/validation sample must be as small as possible. Hence, only 10% of the simulation runs are used for training and validation. Of that, 80% and 20% are used for training and validation, respectively. Meanwhile, the remaining are unseen testing samples.



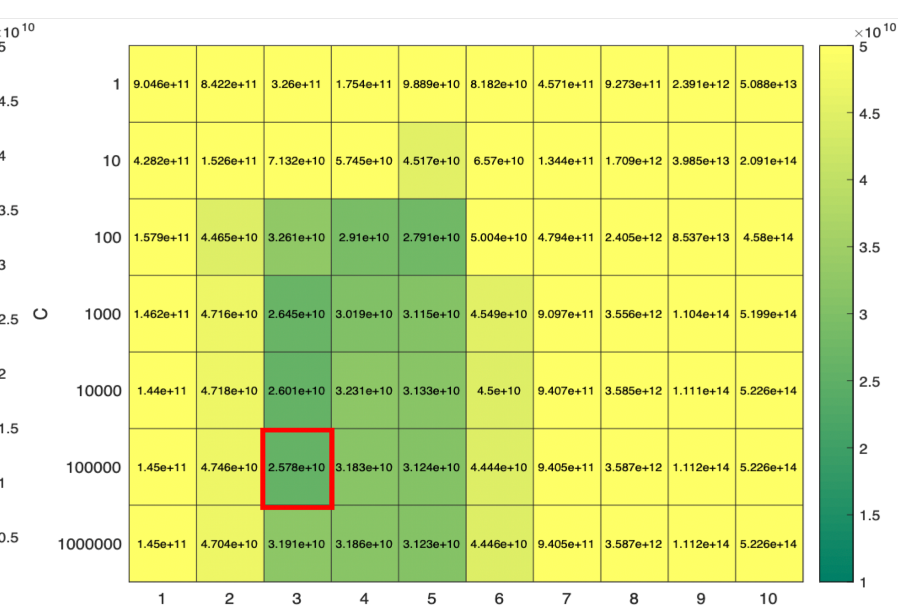
Space-Filling LHS Partition

Results

Support Vector Machine Regression with Linear and Polynomial Kernels:

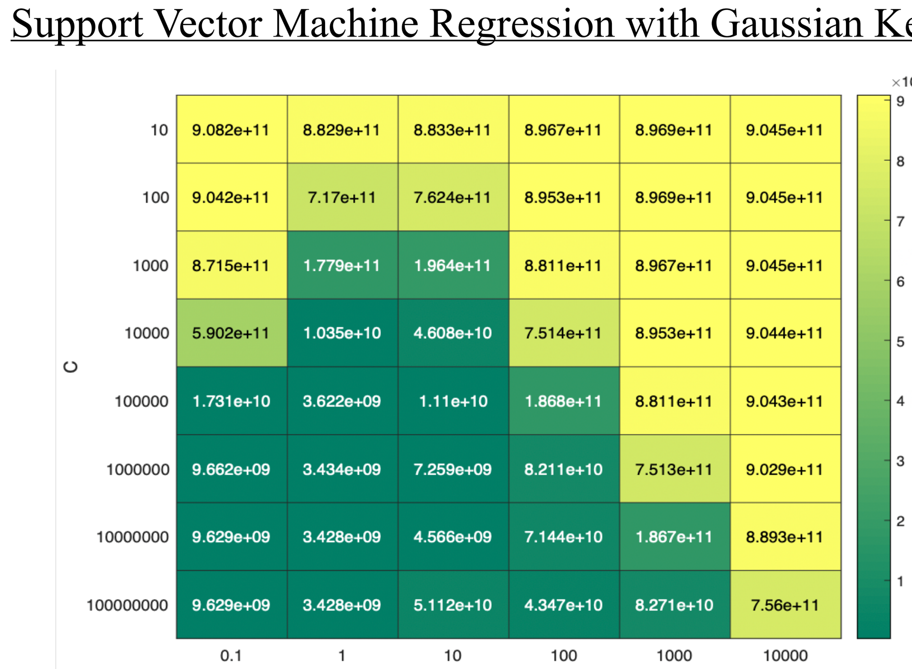


Polynomial SVM Training

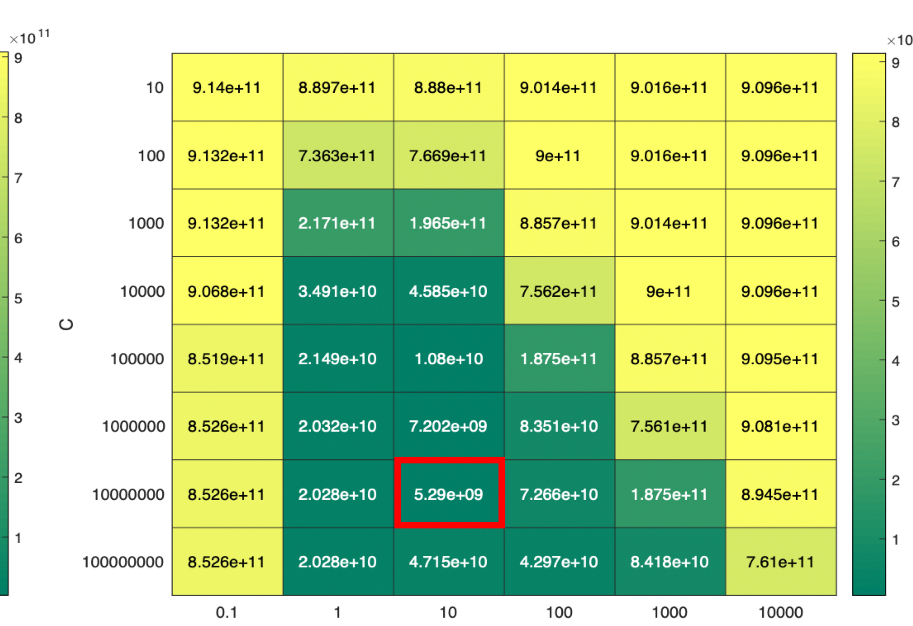


Polynomial SVM Validation

Support Vector Machine Regression with Gaussian Kernel:

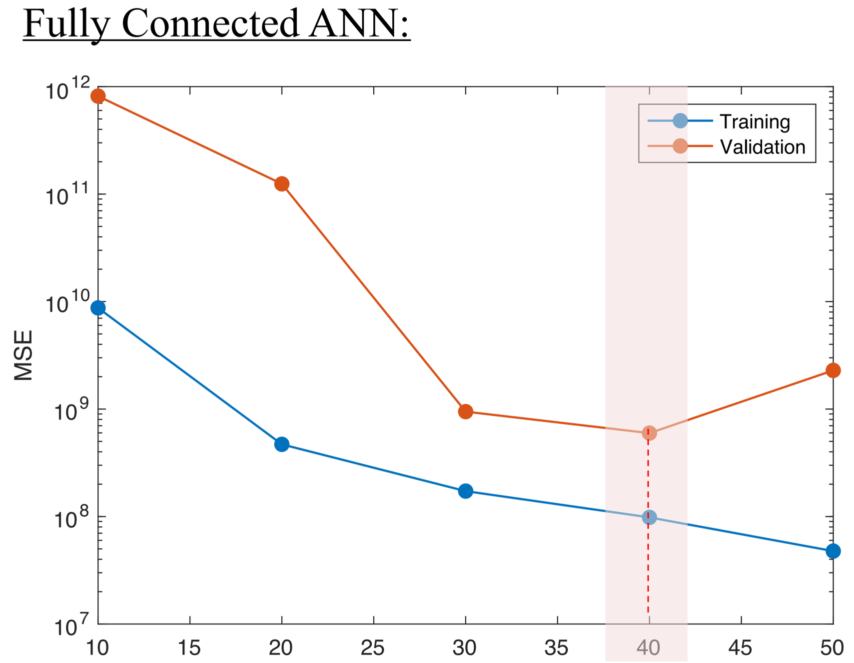


Gaussian SVM Training

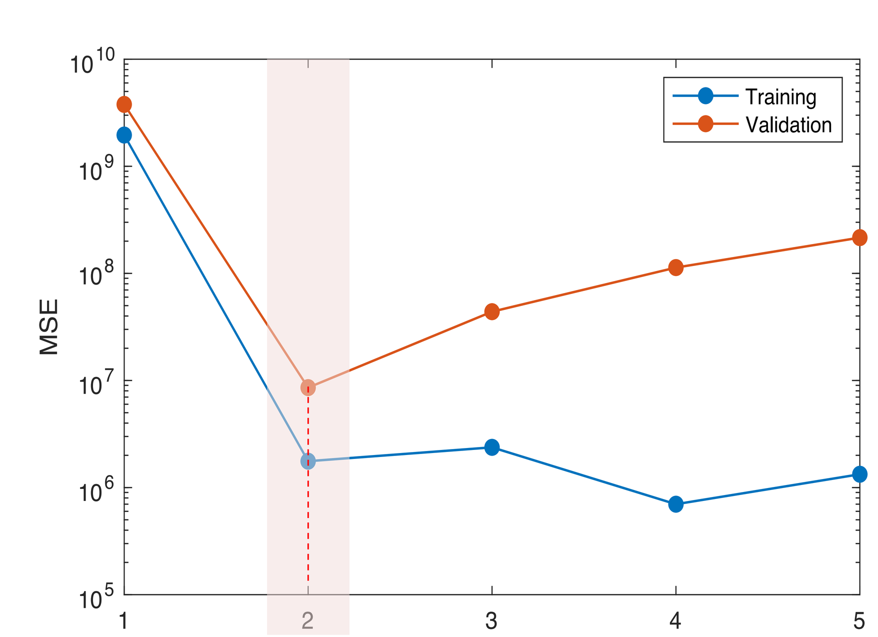


Gaussian SVM Validation

Fully Connected ANN:



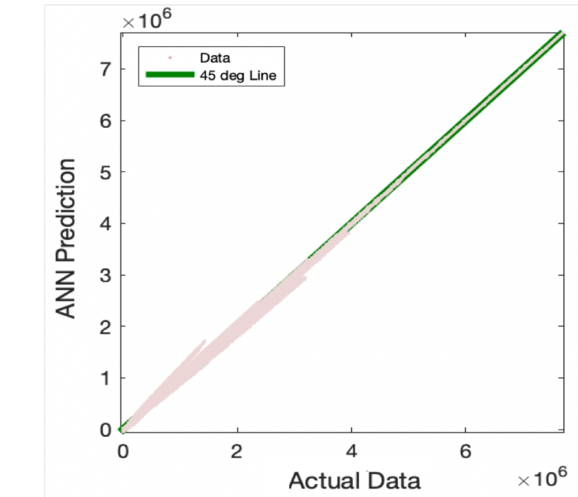
ANN Neuron Diagnostics



ANN Layer Diagnostics

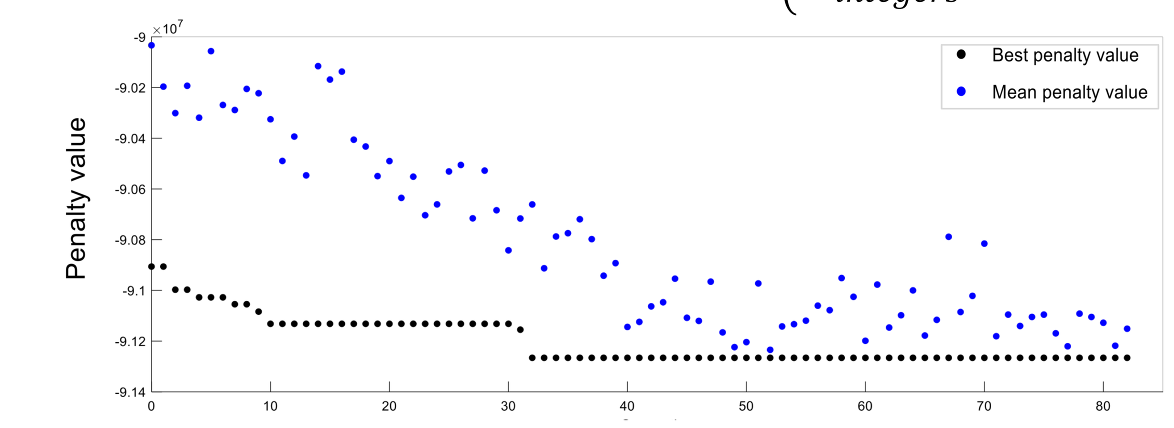
Model Accuracy

Regularized 10-neuron, 3-layer ANN is the best-performing model in predicting two-phase flow. To verify the accuracy of this ANN model in SBO applications, an integer nonlinear programming optimization problem is posed and solved with GA using oil price of $p_o = 50\ \$/barrel$ and water cost of $c_w = 20\ \$/barrel$.



Regression Fit

$$\max_x f(x) = p_o\, g(x) - c_w\, h(x) \quad s.t. \quad \begin{cases} x_2, x_3, x_4 \leq 10 \\ x_2, x_3, x_4 > 0 \\ integers \end{cases}$$



Genetic Algorithm solver

Proxy	ICV – x*	Profit – f(x*)	Time
ANN	[7 2 1]	$ 90,989,317	2.0 hrs
Eclipse	[10 2 1]	$ 90,898,380	16.7 hrs

Conclusion & Future Work

Conclusion:

- ML provides robust means of building a proxy that is fast, accurate, and computationally inexpensive for surrogate-based optimization purposes, saving 88% in optimization time
- 3rd degree polynomial kernel resulted in the best SVM regression fit to the given CFD model
- ANN was found to be superior to kernel-based SVM regression techniques in building a physics-free statistical CFD simulator

Future Work:

- Explore recurrent neural network (RNN) architectures as the simulator to be replicated generates sequential (temporal) data
- Incorporate radial grid, rock heterogeneity, multiphase flows, a large number of ICVs, and multiple wells to examine the ML techniques capability in fitting more complex models

References

1. Abukhamsin, A.Y. (2017). Inflow profiling and production optimization in Smart wells using distributed acoustic and Temperature measurements (Unpublished doctoral dissertation). Stanford University, CA.
2. Ahmadi, M. A., Ebadi, M., Shokrollahi, A., & Majidi, S. M. J. (2013). Evolving artificial neural network and imperialist competitive algorithm for prediction oil flow rate of the reservoir. Applied Soft Computing, 13(2), 1085-1098.
3. Al-Anazi, A., & Gates, I. D. (2010). A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. Engineering Geology, 114(3-4), 267-277.
4. Burden, F., & Winkler, D. (2008). Bayesian regularization of neural networks. In Artificial neural networks (pp. 23-42). Humana Press.
5. El-Sebakhy, E. A. (2009). Forecasting PVT properties of crude oil systems based on support vector machines modeling scheme. Journal of Petroleum Science and Engineering, 64(1-4), 25-34.
6. Foresee, F. D., & Hagan, M. T. (1997, June). Gauss-Newton approximation to Bayesian learning. In Proceedings of International Conference on Neural Networks (ICNN'97) (Vol. 3, pp. 1930-1935). IEEE.