A Novel Active Learning Framework for Parametric Design Dataset Generation

Advaith Narayanan
Leigh High School

IMECE2024-144594
Poster #144594

Topic 15-01-01: Undergrad R&D Expo Poster

Project Objectives and Goals

- To create an **adaptable Active Learning framework/ algorithm** for **optimal design dataset generation (ALFD** algorithm**: A**ctive **L**earning **f**or **D**esign**)**.
- Considerations include:
  - Multiple Performance Objectives
  - Redundant Performance Objective Detection
  - Infeasible Performance Value Detection
  - Compatibility with both Categorical and Continuous Parameters
  - Simulation/Query Failures
- To evaluate the ALFD against uniform sampling using **constrained design regression benchmarks** [1]

Background

- Commonly used for design datasets, [pseudo]random design parameter value generation ensures design space coverage but potentially **overrepresent undesired performance values, increase fail rates, and decrease surrogate regressor accuracies [2,3]**.
- **Active Learning** (AL) selectively queries informative points for labeling, improving data efficiency.

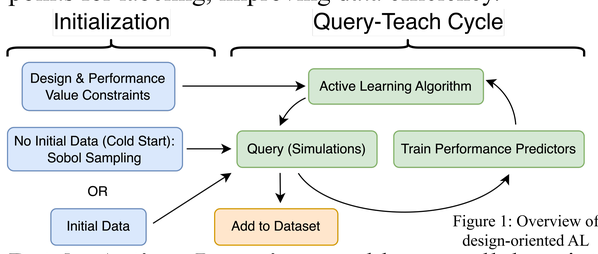


Figure 1: Overview of design-oriented AL

- **Batch Active Learning** enables parallel point labeling/simulations and less frequent model reconstruction.

	Uniform Sampling	Sequential Active Learning	Batch Active Learning
Accuracy	Low	High	**Medium**
Query Speed	High	Low	**Medium**
Cost (Query) Efficiency	Low	High	**Medium**

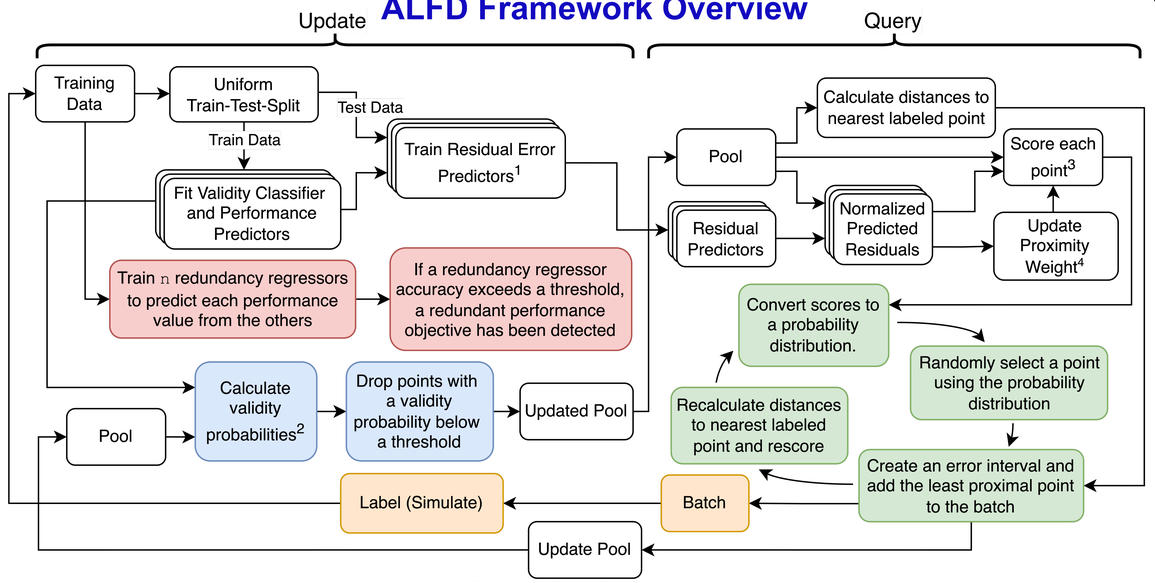
Table I: Comparison of query methods

- Some have introduced ranked-batch and proxy-model-based AL methods [4,5], those studies do not account for multi-objective problems.
- Others have created AL algorithms to achieve a pareto-optimal set of designs [6]. Our work, however, aims to create a model with a low harmonic mean of the MAPEs over the entire design space.

Terminology

- **MAPE:** A metric for measuring a model's prediction accuracy.
- **Acquisition function:** A method of scoring points, usually a function of uncertainty and proximity to the nearest labeled point.

ALFD Framework Overview



[1] We use a K-Nearest-Neighbors (KNN) model as the Residual Error Predictor to preserve residual normalization.

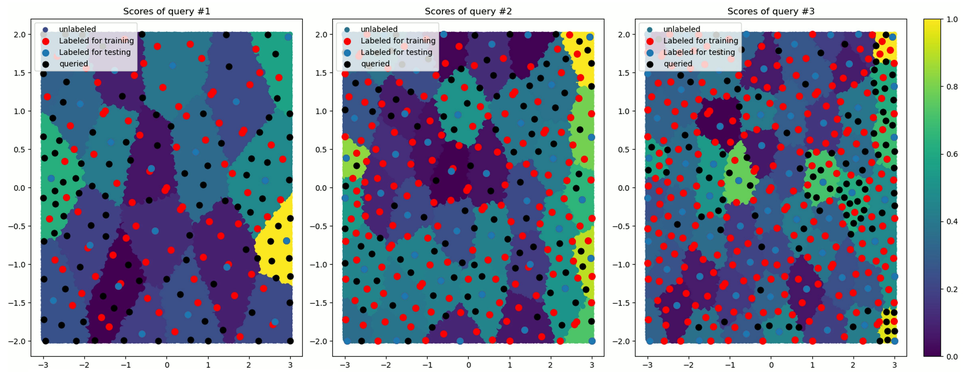
[2] We use the following conservatively high estimate of the validity probabilities: $\prod_{i=1}^{n} P_i^{C(P_i,r)}$ where P_i is the validity probability of the ith performance value, r is the distance to the nearest labeled point, and $C(P_i, r)$ is a classifier confidence function, such as $1 - H(P_i)$.

[3] We score each point with the following experimentally determined acquisition function:

$\text{scores} = (\text{proximity\_weight} + (1 - \text{proximity\_weight}) \cdot \text{error})^{\frac{1}{\text{proximity\_weight}}}$ where error is the mean normalized predicted residuals.

[4] We dynamically update the proximity weight based on the harmonic mean of the MAPEs for all performance predictors.

Data and Results



ALFD densely queries places of higher predicted error, but still maximizes distance from labeled points.

Figure 2: 3 Consecutive Queries. Lighter colors indicate higher predicted error. Plots are flattened visualizations of a higher dimensional space.

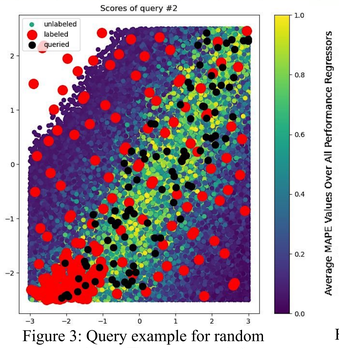


Figure 3: Query example for random regression problem. The empty top left corner demonstrates ALFD's **design-dropping strategy** with **minimal invalid queries**.

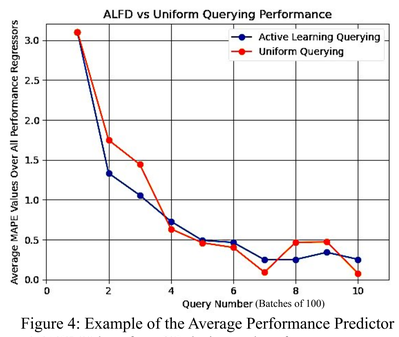


Figure 4: Example of the Average Performance Predictor MAPE Values for a 2D design and performance space (JLH1 and JLH2) [1]. ALFD **initially outperforms uniform sampling and later matches it**.

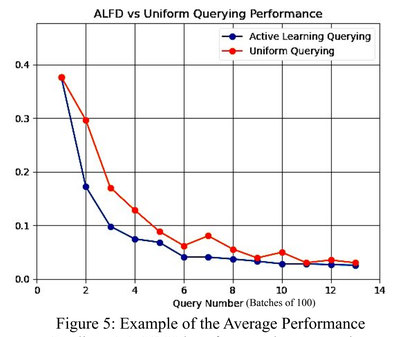


Figure 5: Example of the Average Performance Predictor MAPE Values for a random regression problem set. In this test regression problem, ALFD **consistently outperforms uniform sampling.**

Conclusion

- ALFD outperforms uniform sampling for some problems and matches its performance for others.
- ALFD successfully reduced the number of invalid queries and detected redundant performance objectives.
- Fluctuations in accuracy, as shown in Figure 5, indicate potential overfitting, which could be addressed with a higher quality regressor.

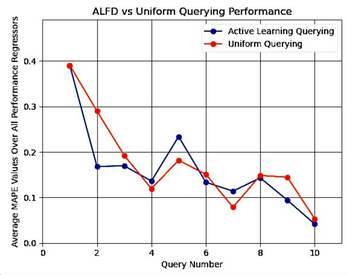


Figure 5: Performance of ALFD vs Uniform Sampling for a 2D design space and 3D performance space (JLH1, JLH2, Ackley 2D).

Future Work

- Investigate non-stochastic alternatives to the query strategy to improve reproducibility and reliability
- Simplify the query strategy to eliminate hyperparameters such as acquisition function exponents, error interval widths, etc to improve reliability and interpretability
- Support for categorical performance values
- Further experimentation with uncertainty calculations and query strategies, such as query-by-committee, to improve convergence rates of ALFD
- Extensive testing with complex design benchmark problems, such as GKXWC1/2, Three Truss, Reinforced Concrete Beam, etc.

GitHub: **PyPI Package:**

References

[1] R. Yu, C. Picard, and F. Ahmed, "Fast and accurate bayesian optimization with pre-trained transformers for constrained engineering problems," 2024.

[2] A. Narayanan, "A data-driven recommendation framework for optimal walker designs," *ArXiv*, 2023. [Online]. Available: https://arxiv.org/pdf/2310.18772, https://idetc.secure-platform.com/a/solicitations/228/sessiongallery/17287/application/145053

[3] N. J. Bagazinski and F. Ahmed, "Ship-d: Ship hull dataset for design optimization using machine learning," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 87301. American Society of Mechanical Engineers, 2023, p. V03AT03A028.

[4] T. N. Cardoso, R. M. Silva, S. Canuto, M. M. Moro, and M. A. Gonçalves, "Ranked batch-mode active learning," *Information Sciences*, vol. 379, pp. 313–337, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025516313949

[5] T. Evans, S. Pathak, H. Merzic, J. Schwarz, R. Tanno, and O. J. H´enaff, "Bad students make great teachers: Active learning accelerates large-scale visual understanding," *arXiv*, vol. abs/2312.05328, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266162309

[6] M. Zuluaga, G. Sergent, A. Krause, and M. P¨uschel, "Active learning for multi-objective optimization," in *Proceedings of the 30th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research*, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 1. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 462–470. [Online]. Available: https://proceedings.mlr.press/v28/zuluaga13.html