# Green LIME: Improving AI Explainability through Design of Experiments

Alexandra Stadler[1], Werner G. Müller[1], Radoslav Harman[2]

[1]Institute of Applied Statistics, Johannes Kepler University Linz, Austria

[2]Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava, Slovakia

JOHANNES KEPLER UNIVERSITY LINZ

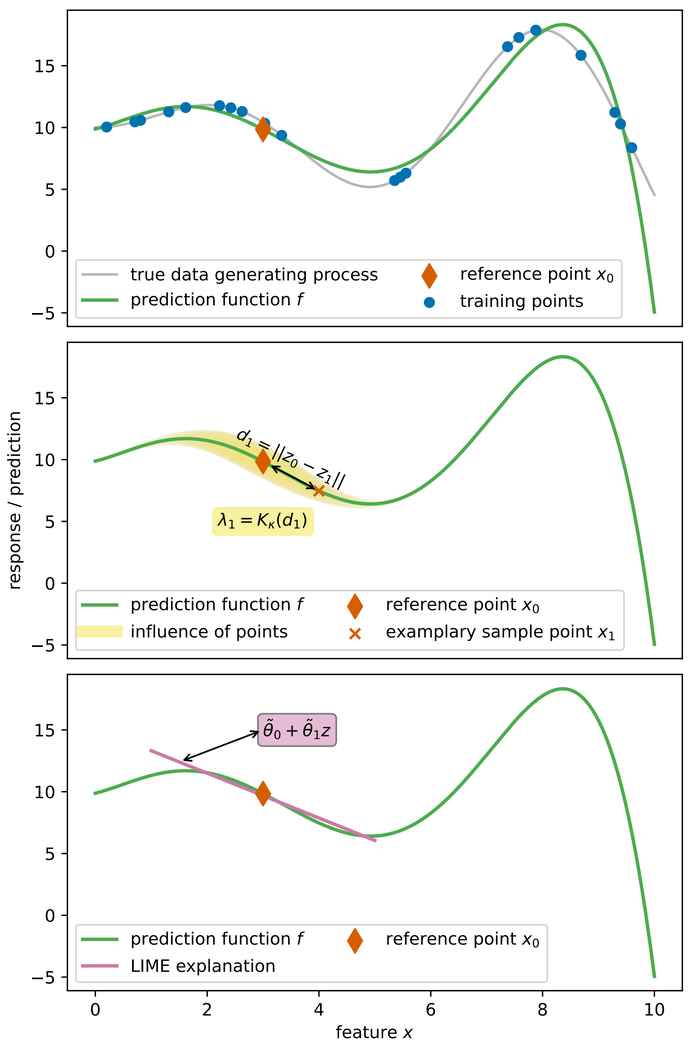
## Local Interpretable Model-agnostic Explanations (LIME)

- Many AI models are too complex for human interpretability, making predictions difficult to understand.
- This is crucial in critical fields like healthcare, where trust in AI predictions is essential.
- Current methods to improve model interpretability are sometimes resource-intensive.
- Optimal Experimental Design aims to maximize information from limited observations.
- LIME generates explanations by creating new data points near an instance and passing them through the model.
- LIME can be computationally expensive due to the need for multiple samples.
- Initial focus: Application of LIME to tabular data, regression tasks, and local linear approximations.

**By utilizing optimal design of experiments' techniques, we reduce the computational effort of LIME by a significant amount, thus providing a version that is much more energy-efficient or "green".**

## Local Linear Models as Explanations

The model's prediction function $f$ shall be approximated by a linear model in the vicinity of the reference point. The prediction at a point $\mathbf{x}$ is denoted by $y = f(\mathbf{x})$. An explanation is computed on the standardized instance $\mathbf{z}_0$, which is centered and scaled by the training data's empirical mean and standard deviation. We showcase the process by the following toy example from Visani et al. (2020).




LIME SIMPLE MODEL FIT

A model has been fit to some training data set and the model's prediction of a reference point $\mathbf{x}_0$ shall be explained.

A random sample is rescaled by the training data's mean and standard deviation, resulting in samples $\mathbf{z}_i$. The distance $d_i$ to the rescaled reference point $\mathbf{z}_0$ is computed via some distance metric, e.g. Euclidean distance. Then, locality weights $\lambda_i$ are assigned by a kernel function with parameter $\kappa$, i.e. $\lambda_i = K_\kappa(d_i)$.

The local linear model's coefficients are estimated by minimizing the sum of squared deviations from the true values, weighted by $\lambda_i$.

The explanation provides contributions to the prediction w.r.t. the input variables, i.e. $\widetilde{\theta}_j z_{0j}$. This reflects the behavior of the model in a small neighborhood around the instance, but not over the whole model.

## Optimal Design of Experiments for Local Linear Models

Random sampling of new instances is inefficient, especially if the ML model's prediction function is costly, or there are many evaluation requests at once. **The task of ODE is to find the data points $\mathbf{z}_i$ that are most informative.** In this setting, the input is location shifted by the instance of interest, i.e. the design is computed on $\mathbf{u} = \mathbf{z} - \mathbf{z}_0$. We follow theory in Fedorov et al. (1999).

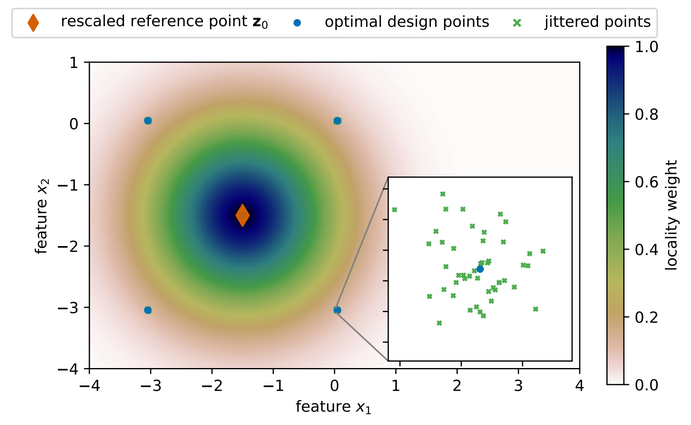
A design consists of a set of candidate design points and their corresponding weights. Let the design weights be defined by $\xi_i = \xi(\mathbf{u}_i)$, with $0 \leq \xi(\mathbf{u}_i) \leq 1$ and $\sum_i \xi(\mathbf{u}_i) = 1$. The weight indicates the amount of experimental units assigned to this point. Suppose the bias of the local approximation to $f$ is negligible, and denote the (approximate) normalized mean squared error matrix for $\widetilde{\boldsymbol{\theta}}$ by $\mathbf{R}(\xi)$.
In order to quantify the "goodness" of a design, we utilize the D-criterion, i.e. $\Psi_D[\mathbf{R}(\xi)] = \log|\mathbf{R}(\xi)|$, which captures the "generalized" MSE of the parameters in the local model.

$\xi^*$ is called the D-optimal (approximate) design if $\xi^* = \operatorname{argmin}_\xi \Psi_D[\mathbf{R}(\xi)]$.

The only information about the contribution of a design point is its distance to the reference point and the other points. We propose a simple design that optimizes the D-criterion (under regularization), with optimized Euclidean distances. The design points are all possible $m$-dimensional vectors of $\begin{pmatrix}\pm u^* & \pm u^* & \dots & \pm u^*\end{pmatrix}$ with

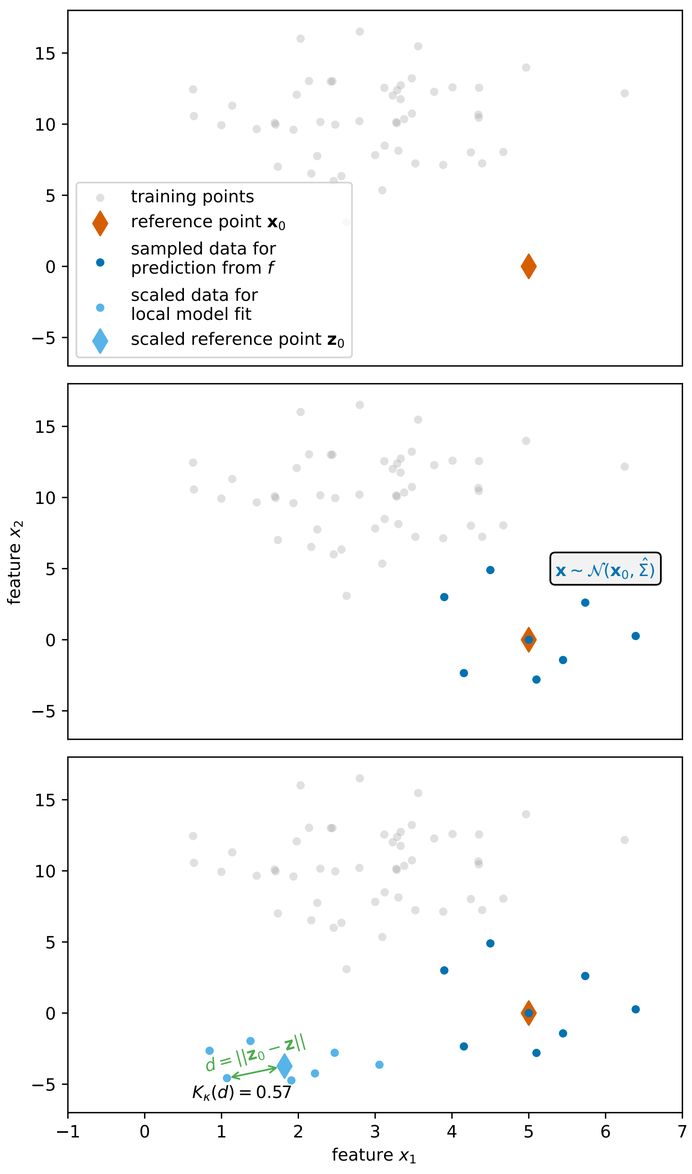
$$u^* = \operatorname*{argmin}_u \log\left(\delta + (1-\delta)K_\kappa(\sqrt{m}u)^2\right) - \log\left(\delta + (1-\delta)K_\kappa(\sqrt{m}u)\right) - 2m\log(u),$$

which is equivalent to optimization of the D-criterion under such a design. The zero vector is included with design weight $0 < \delta \ll 1$ for regularization of the criterion. This leads to $2^m + 1$ support points of the design.




Our sample contains the support of the optimal design. In order to increase the sample up to a specified size $N$, **we add some small Gaussian noise to the support points**. The variance of the random noise is a tuning parameter. Each support point gets an allocation of the remaining sample points proportional to the design weight. We call this approach **"jittering"** (around the support set).

## Neighborhood Sampling in LIME




LIME SAMPLE GENERATION

The prediction of a reference point $\mathbf{x}_0$ shall be explained.

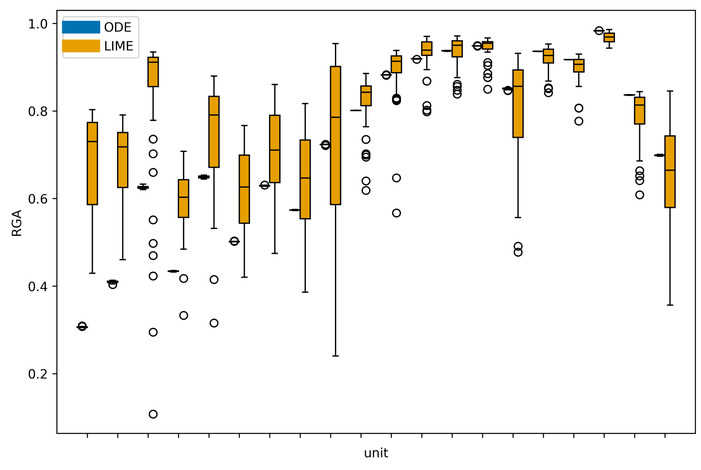
A random sample is drawn from a normal distribution with location $\mathbf{x}_0$ and scale $\widehat{\boldsymbol{\Sigma}}$ corresponding to the (independent) empirical variances of the training data.

The sample is rescaled by the training data's mean and standard deviations. Then the distance $d$ to the rescaled reference point is calculated and put into a kernel function $K_\kappa$ with bandwidth $\kappa$. The result is a "locality" weight.

This new sample is used for the local model fit in the explanation.

## Evaluation of Explanations

Objective evaluation of explanations is not straightforward. The goal is not well-defined and there is no consensus on what constitutes a good explanation. One metric is the rank graduation accuracy (RGA, see Babaei et al., 2025).




The RGA has values between 0 and 1 and is based on the predictions by the ML model as well as the ranks of the simple and the ML model. High values of the RGA indicate a good concordance between the predictions of the two models.
On the left is an example of the RGA values for a complex model with 5 input variables. For each of the 20 randomly drawn reference units, 50 explanations are computed with sample size 50 for the ODE and the original LIME method.

## Conclusion

- Neighborhood sampling is inefficient in LIME.
- The process can be costly if the model's prediction function is costly or if many predictions have to be computed simultaneously.
- Design of experiments can help to limit the number of function evaluations by collecting data more efficiently and thus decreasing energy consumption for a "green" version of LIME.

## References

Babaei, G., Giudici, P. (2025). A statistical package for safe artificial intelligence. DOI: 10.1007/s10260-025-00796-y

Fedorov, V. V., Montepiedra, G., & Nachtsheim, C. J. (1999). Design of experiments for locally weighted regression. DOI: 10.1016/s0378-3758(99)00018-x

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. DOI: 10.1145/2939672.2939778

Stadler, A., Müller, W.G. & Harman, R. (2025). Green LIME: Improving AI Explainability through Design of Experiments. DOI: 10.48550/arXiv.2502.12753

Visani, G., Bagli, E., & Chesani, F. (2020). OptiLIME: Optimized LIME Explanations for diagnostic computer Algorithms. DOI: 10.48550/arxiv.2006.05714