2025 IEEE 64th Conference on Decision and Control (CDC)
December 10-12, 2025. Rio de Janeiro, Brazil

AutoLIME and PWA-LIME: towards robust explanations of deep dynamical models

Federico Porcari[1], Valentina Breschi[2], Simone Formentin[1]

2025 IEEE 64th Conference on Decision and Control (CDC) | 979-8-3315-2627-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/CDC57313.2025.11311973

***Abstract*— The increasing complexity of machine learning models highlights the need for interpretability, especially in critical domains requiring trust and transparency. Local Interpretable Model-agnostic Explanations (LIME) is a popular eXplainable AI (XAI) method that provides localized, instance-specific explanations using an interpretable surrogate model. However, its effectiveness is limited by the lack of systematic guidelines for tuning its hyperparameters. This paper addresses this limitation by proposing *Automatic LIME (AutoLIME)*, a bi-level optimization framework to tune LIME's kernel width. Additionally, we introduce *PieceWise Affine LIME (PWA-LIME)*, a clustering-based extension of LIME for multi-instance explanations, particularly useful for interpreting black-box models of dynamical systems. Preliminary numerical results validate the potential of these methods in explaining opaque dynamical models.**

***Index Terms*— XAI, LIME, PWA systems**

I. Introduction

Machine learning has advanced significantly in the past decade, with complex models like deep neural networks achieving state-of-the-art performance across various domains. However, their complexity raises concerns about interpretability and trust. To address these issues, eXplainable AI (XAI) methods have gained prominence in machine learning for explaining specific predictions [1], and they have started to catch the eye of the control community [2]. Nevertheless, these tools are not yet widely used when the interest shifts from describing static phenomena to dynamic systems, with only a few examples where tools from XAI are borrowed to enhance the explainability of networks of dynamical systems [3] and data-driven controllers [4].

Among available XAI approaches, Local Interpretable Model-agnostic Explanations (LIME) [5] generates synthetic data around an input instance, weights them using a proximity measure, and fits an interpretable surrogate model to approximate the black-box model's local behavior. Despite its utility, LIME faces limitations, particularly in selecting the neighborhood for explanations. Indeed, broad neighborhoods result in unspecific explanations, while narrow ones lead to instability due to insufficient data points, as discussed in [6]. Balancing locality and stability requires careful tuning of LIME's hyperparameters. Efforts to enhance LIME's stability and neighborhood selection include hierarchical clustering and k-nearest neighbors [7], as well as autoencoder-based sampling [8]. However, these methods still rely on approach-specific hyperparameters, whose selection remains manual and critical to achieving meaningful explanations. Another limitation of LIME is its focus on single-instance explanations. Extending LIME to provide multi-instance explanations would enhance the understanding of how models approximate complex relationships, particularly when looking at models of dynamical systems, where each instance can represent an operating condition of the actual system.

In this paper, we address these challenges by proposing:

1) *Automatic LIME (AutoLIME)*, a *bi-level optimization framework* for systematically tuning LIME's kernel width, leveraging ideas from information theory and kernel-weighted regression to balance between explanation fidelity and robustness;
2) *PieceWise Affine LIME (PWA-LIME)*, an extension of LIME to *multi-instance explanations* via a clustering-based approach.

The paper is organized as follows. Section II provides an overview of LIME and its limitations, with an emphasis on challenges related to neighborhood selection. Section III introduces the AutoLIME bi-level framework for tuning the kernel width. In Section IV, we present PWA-LIME, our multi-instance extension to LIME. Section V validates the proposed methods on a numerical case study on the dynamical system considered in [9]. The paper is ended by some concluding remarks.

Notation: $\mathbb{N}_0$, $\mathbb{R}$, $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ denote the set of natural numbers (including zero), real numbers, real (column) vectors with n columns and real matrices of dimensions $n \times m$, respectively. For a given positive definite matrix $A \in \mathbb{R}^{n\times n}$, $A^{-\frac{1}{2}}$ denotes the inverse of its unique positive definite square root. We denote with $\mathrm{diag}(a_1, a_2, \dots, a_n)$ an $n \times n$ diagonal matrix with entries $a_1, \dots, a_n$. For any positive semidefinite matrix $Q \in \mathbb{R}^{n\times n}$, $\|v\|_Q = \sqrt{v^\top Q v}$ denotes the weighted norm of v by Q. Given a set $\mathcal{A} \subseteq \mathbb{R}^n$, $\mathcal{A}^c$ indicates its complement and $\mathbb{I}_{\mathcal{A}}$ denotes the indicator function associated with $\mathcal{A}$. Given $v \in \mathbb{R}^n$, we denote its j-th component as $[v]_j$, with $j = 1, \dots, n$ and, with a slight abuse of notation, we denote with $H_\delta(v)$ the sum of Huber losses [10] of each component of v, i.e.,

This work is partially supported by the FAIR project (NextGenerationEU, PNRR-PE-AI, M4C2, Investment 1.3), the 4DDS project (Italian Ministry of Enterprises and Made in Italy, grant F/310097/01-04/X56), and the PRIN PNRR project P2022NB77E (NextGenerationEU, CUP: D53D23016100001). It is also partly supported by the ENFIELD project (Horizon Europe, grant 101120657).

[1]Federico Porcari and Simone Formentin are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, via G. Ponzio 34/5, 20133 Milano, Italy. Email: {federico.porcari, simone.formentin}@polimi.it

[2]Valentina Breschi is with the Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. Email: v.breschi@tue.nl

$$H_\delta(v) = \sum_{j=1}^{n} H_\delta([v]_j),$$
$$H_\delta([v]_j) = \begin{cases} \frac{1}{2}[v]_j^2, & |[v]_j| \le \delta, \\ \delta\left(|[v]_j| - \frac{1}{2}\delta\right), & |[v]_j| > \delta, \end{cases} \quad (1)$$

where $\delta > 0$ is a design parameter.

II. An overview of LIME and its limitations

Let $z \in \mathbb{R}^n$ be a feature in input to a nonlinear, *unknown* and yet accessible function $f^o : \mathbb{R}^n \to \mathbb{R}^m$, where accessibility entails that we can observe the output $y = f^o(z)$.

Consider the general supervised learning problem of approximating such an unknown function with a finite set of feature-output measurements $\{z_i^d, y_i^d\}_{i=1}^{N^d}$. The solution to this problem consists of selecting a (suitable) model class $\mathcal{F} = \{f : \mathbb{R}^n \to \mathbb{R}^m\}$ and learning a specific instance $\hat{y} = f(z)$ of it from the available data such that

$$\hat{y}_i = f(z_i^d) \approx y_i^d, \quad \forall i = 1, \dots, N^d, \quad (2)$$

i.e., the output y_i^d is accurately predicted for all samples. Complex models, such as neural networks, are knowingly capable of achieving high accuracy in approximating a wide range of functions. However, attaining accuracy often entails increasing the complexity of the model class. In turn, this makes it harder to explain the model's behavior, with the role of each parameter becoming almost impossible to discern.

To allow for the interpretability of complex functions, LIME [5] generates a post-hoc explanation of a single instance of the feature z fed to the model $f(\cdot)$. Hence, LIME focuses on providing *local interpretability* rather than describing the global behavior through the following steps.

1) An N-long synthetic dataset $\mathcal{D}_N$ centered around z is generated, which contains both perturbed features z_i, $i = 1, \dots, N$, and the corresponding black-box predictions $\hat{y}_i = f(z_i)$, namely $\mathcal{D}_N = \{z_i, \hat{y}_i\}_{i=1}^N$.
2) To enforce locality around the specific instance z, each pair $(z_i, \hat{y}_i) \in \mathcal{D}_N$ is weighted based on its relevance in describing the instance z. This relevance is quantified by a proximity measure $\pi_z : \mathbb{R}^n \to \mathbb{R}$, which is typically implemented as a kernel function.
3) By using these artificial data, the function $f(\cdot)$ is explained locally by constructing a simple surrogate model $g : \mathbb{R}^n \to \mathbb{R}^m$, drawn from an inherently interpretable model class $\mathcal{G}$. This surrogate is learned seeking for the "best" fit of the synthetic dataset according to a loss $\mathcal{L} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ that measures the accuracy of $g(\cdot)$ in explaining $f(\cdot)$ in the neighborhood defined by $\pi_z(\cdot)$. This ultimately translates into the solution of the optimization problem

$$\min_{g \in \mathcal{G}} \quad \sum_{i=1}^{N} \mathcal{L}(f(z_i), g(z_i), \pi_z(z_i)) + \Omega(g), \quad (3)$$

where $\Omega(g)$ is a measure of the complexity of g, which is used to favor simpler explanations, e.g., sparser models.

This procedure implies the construction of an interpretable surrogate model $g(\cdot)$ in the proximity of z, which requires the definition of a neighborhood of such a feature instance. The latter, in turn, depends on (i) how the synthetic dataset $\mathcal{D}_N$ has been constructed, and (ii) the selected proximity measure $\pi_z(\cdot)$. However, constructing $\mathcal{D}_N$ and selecting $\pi_z(\cdot)$ is delicate and far from trivial. Indeed, if the defined neighborhood is too broad, LIME produces non-specific explanations. On the other hand, if it is too narrow, LIME focuses only on a few data points, making the explanations unstable, where stability is defined as follows.

Definition 1 (Algorithmic stability): An XAI method is stable if repeated applications of the algorithm under the same conditions, e.g., similar dataset distribution, result in consistent explanations across each experiment. ■

Stability is a critical issue for LIME, as the random sampling used to define the neighborhood of an instance z may generate out-of-distribution points or ill-defined regions when the proximity measure $\pi_z(\cdot)$ is too restrictive. To address this, several works propose improved sampling strategies [7], [11], but these focus on dataset generation and give little guidance on choosing $\pi_z(\cdot)$.

Most proximity definitions rely either on trial-and-error tuning [6] or on fixed distance-based weighting, which does not ensure that all samples are informative for the surrogate model. A few studies treat the problem more systematically. For example, [12] jointly optimizes sample generation and kernel hyperparameters to align distributions, but without constraints on neighborhood size, the resulting explanations risk being overly broad. Conversely, [13] adjusts kernel size to maximize stability while enforcing a prescribed adherence to the black-box function $f(\cdot)$. However, fixing this adherence level enforces a rigid trade-off between local fidelity and stability, preventing more flexible compromises.

III. Automatic LIME (AutoLIME): bi-level optimization for proximity measure calibration

By focusing on kernels as proximity measures, we take a different perspective on their calibration with respect to the one in the literature. Indeed, while still aiming to attain a trade-off between the accuracy of local explanations and algorithmic stability as in [13], we propose to calibrate the proximity measures to maximize the probability of the surrogate model generating the true output subject to the available data. This choice does not require fixing a priori an adherence level, allowing the explainability algorithm to directly select the "most suitable" trade-off value.

Let $\kappa \ge 0$ define the width of the kernel π_z and let the class of surrogate models $\mathcal{G}$ be parameterized by some parameters $\xi \in \Xi$. Then, the cost function in (3) can be recast explicitly showing its dependence on κ as

$$J(\kappa, \xi, \mathcal{D}_N) = \sum_{i=1}^{N} \mathcal{L}(f(z_i), g(z_i; \xi), \pi_z(z_i; \kappa)) + \Omega(\xi), \quad (4)$$

where ξ is the variable optimized by LIME and $\Omega : \Xi \to \mathbb{R}$ is, e.g., a nuclear-norm regularization enforcing sparsity

in the surrogate model [14]. To select the most suitable kernel width, we draw inspiration from the literature on kernel-weighted regression (see, e.g., [15], [16]) and cast the selection of κ into the bi-level optimization problem

$$\max_{\kappa} \quad M(\kappa, \xi^*(\kappa), \mathcal{D}_N), \tag{5a}$$

$$\text{s.t.} \quad \xi^*(\kappa) \in \underset{\xi(\kappa)}{\operatorname{argmin}} \; J(\kappa, \xi(\kappa), \mathcal{D}_N), \tag{5b}$$

where the inner problem coincides with LIME and seeks to find the optimal ξ for a given κ, whereas the outer cost $M(\kappa, \xi^*(\kappa), \mathcal{D}_N)$ allows for the selection of the "best" kernel width κ^* given the optimal model parameters ξ^*. Nonetheless, differently from the choices usually made in kernel-weighted regression for the outer loss[1] $M(\cdot)$, we select it to achieve a trade-off between local accuracy of explanation, which is maximized by narrow kernels, and algorithmic stability, which is maximized by large kernels, as discussed in [13]. We propose to achieve this trade-off by solving a maximum likelihood estimation problem, i.e., we select $M(\cdot)$ as

$$M(\kappa, \xi^*, \mathcal{D}_N) = \mathbb{P}\{g(Z, \xi^*) = f(Z) \,|\, Z, \xi^*\}, \tag{6}$$

with $Z = \{z_i : z_i \in \mathcal{D}_N\}$, thus searching for the kernel width maximizing the probability that the surrogate model $g(z;\xi)$ generates the true prediction $f(z)$ given the dataset $\mathcal{D}_N$ and the optimal parameter ξ^*. By defining the surrogate model's residuals $\varepsilon_i(\kappa) = f(z_i) - g(z_i; \xi^*(\kappa))$, $i = 1, \ldots, N$, and denoting as $\phi(\varepsilon_i(\kappa))$ the associated probability density functions, i.e., $\phi(\varepsilon_i(\kappa)) = \mathbb{P}\{g(z_i, \xi^*) = f(z_i) \,|\, z_i, \xi^*\}$, an explicit expression for the outer cost in (6) can be found by making an assumption about the structure of $\phi(\varepsilon_i(\kappa))$, as well as one on the surrogate model residuals.

Assumption 1 (i.i.d. residuals): The surrogate model residuals ε_i are independent and identically distributed.

Indeed, under Assumption 1, maximizing the outer loss $M(\kappa, \xi^*(\kappa), \mathcal{D}_N)$ corresponds to the maximization of

$$\log \mathbb{P}\left\{g(Z, \xi^*(\kappa)) = f(Z) | Z, \xi^*(\kappa)\right\} = \sum_{i=1}^{N} \log \phi(\varepsilon_i(\kappa)).$$

While several possible choices can be made on the structure of posterior $\phi(\cdot)$ as well as on the surrogate class $\mathcal{G}$ and of the proximity measure π_z, we now focus our discussion on choices for $\mathcal{G}$, π_z and $\phi(\varepsilon_i)$ to specifically provide explanations for dynamical systems.

A. Bi-level optimization for dynamical systems

When selecting the surrogate model class $\mathcal{G}$, one must consider that its role is to provide an interpretable, yet accurate, description of the black-box model $f(z)$. Since in this work we address the explanation of dynamical systems, it is reasonable to select $\mathcal{G}$ following classical procedures in control theory, where complex models are often locally approximated with their Taylor expansions around an operating condition [18]. Following this rationale and defining the instance $z \in \mathbb{R}^n$ as a vector stacking the state $x \in \mathbb{R}^{n_x}$ and the input $u \in \mathbb{R}^{n_u}$ of a dynamical system, we select the class of surrogate models as the set of fully observable affine time-invariant systems in state-space form

$$\mathcal{G} = \left\{ g : g(z) = \gamma + Ax + Bu = \xi \begin{bmatrix} 1 \\ z \end{bmatrix} \right\}, \; z = \begin{bmatrix} x \\ u \end{bmatrix}, \tag{7}$$

where the LIME optimization parameter $\xi = \begin{bmatrix} \gamma & A & B \end{bmatrix}$ comprises all state-space parameters. This specific choice of surrogate model is consistent with classical control approaches that simplify nonlinear systems with affine maps [19]. Moreover, the assumption on full observability implies that we have prior knowledge of the complexity of the surrogate model, allowing us not to impose any regularization on its structure, i.e., $\Omega(\xi) = 0$ in (4).

To weigh the surrogate prediction errors in the cost function (4), we take as loss $\mathcal{L}(\cdot)$ the squared L_2-norm of the residuals weighted by the proximity measure π_z, namely

$$\mathcal{L}(f(z_i), g(z_i;\xi), \pi_z(z_i)) = \pi_z(z_i) \| f(z_i) - g(z_i;\xi) \|^2, \tag{8}$$

thus searching for the surrogate that best fits the available outputs in a mean square sense. Meanwhile, we focus on Gaussian kernels [20] as proximity measure, namely

$$\pi_z(z_i) = \exp\left\{ -\kappa \left\| z_i - z \right\|^2_{\sigma^{-1}} \right\}, \tag{9}$$

where z_i is a sample of the synthetic dataset $\mathcal{D}_N$ and $\sigma > 0$ is the sample covariance of $\mathcal{D}_N$. Apart from the fact that Gaussian proximity measures are often used both in LIME [6], [11] and in kernel-weighted regression, our choice is motivated by their simplicity and supported by the statement in [16] indicating that the shape of the chosen kernel function has a limited impact on accuracy.

Lastly, we define the outer cost function $M(\kappa, \xi, \mathcal{D}_N)$ as

$$M(\kappa, \xi, \mathcal{D}_N) = \sum_{i=1}^{N} H_\delta\left(\tilde{\varepsilon}_i\right), \tag{10}$$

with $H_\delta(\cdot)$ defined as in (1) and $\tilde{\varepsilon}_i$ is the i-th normalized surrogate model's residual, i.e., $\tilde{\varepsilon}_i = \Lambda^{-\frac{1}{2}} \varepsilon_i$, with Λ being the sample residual error covariance defined as $\Lambda = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \varepsilon_i^\top$.

Using the Huber loss as the outer cost enhances the robustness of the optimization. For small normalized residuals it behaves like the L_2-norm, ensuring smooth convergence, while for large residuals it transitions to the L_1-norm, reducing sensitivity to outliers. Normalizing residuals further shifts the penalty from error magnitude to the frequency of non-negligible relative errors, aligning with the objective of the outer layer in the bi-level problem (5). This prevents LIME from selecting excessively small kernels, which overfit the synthetic dataset, or overly large ones, which yield generic explanations with frequent small errors.

The choice of the Huber loss also complies with (6). Indeed, by selecting (10) we make the implicit assumption that the residuals' distribution $\phi(\varepsilon_i)$ follow a zero-mean

[1] $M(\cdot)$ is usually chosen as an information complexity criterion, e.g., Akaike information criterion, and then optimized through grid search [17].

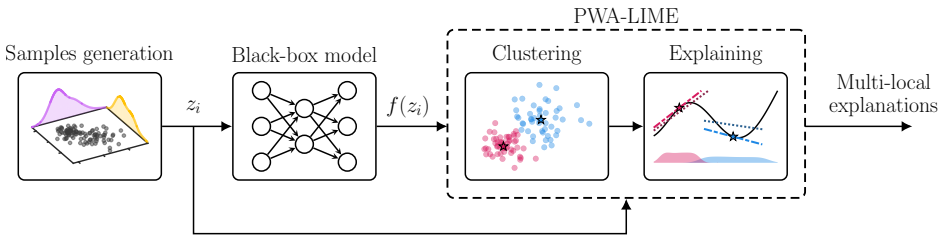


Fig. 1: Block scheme of the explanation process of PWA-LIME.

Huber density function[2]

$$h_\delta(\tilde{\varepsilon}_i) \propto \exp\left\{-\frac{1}{2}\tilde{\varepsilon}_i^2\right\}\mathbb{I}_{[-\delta,\delta]} + \exp\left\{-\delta|\tilde{\varepsilon}_i| + \frac{1}{2}\delta^2\right\}\mathbb{I}_{[-\delta,\delta]^c} = \exp\left\{-H_\delta(\tilde{\varepsilon}_i)\right\}, \quad i = 1,\dots,N, \tag{11}$$

where $H_\delta(\cdot)$ is defined as in (1). Accordingly, it holds that

$$\mathbb{P}\{g(Z,\xi) = f(Z)|Z,\xi\} \propto \sum_{i=1}^{N} H_\delta\left(\tilde{\varepsilon}_i\right), \tag{12}$$

whose right-handside can be taken as outer cost function $M(\kappa,\xi,\mathcal{D}_N)$. Based on these choices, the AutoLIME bi-level optimization problem thus becomes

$$\begin{aligned} \min_{\kappa} \quad & \sum_{i=1}^{N} H_\delta\left(\Lambda^{-\frac{1}{2}}\left(f(z_i) - \xi\begin{bmatrix}1\\ z_i\end{bmatrix}\right)\right), \\ \text{s.t.} \quad & \xi^*(\kappa) \in \underset{\xi(\kappa)}{\operatorname{argmin}}\ J(\kappa,\xi,\mathcal{D}_N), \\ & (7),\ (8),\ (9),\ \Omega(\xi) = 0. \end{aligned} \tag{13}$$

Remark 1 (On the robustness of the Huber loss): As the Huber density in (11) follows a normal distribution inside $[-\delta,\delta]$, δ can be selected to tune the robustness of the algorithm. For example, if $\delta = 1$, roughly 68% of data (one standard deviation) is weighted by the L_2-norm. The remaining 32% is robustified against outliers by the L_1-norm.

IV. PWA-LIME: A CLUSTERING-BASED APPROACH FOR MULTI-INSTANCE EXPLANATIONS

From a control-theoretic perspective, LIME has remarkable similarities with system identification and model order reduction, as both strive to find the best fit for a complex system within a model class of reduced complexity [21]. However, unlike these approaches, LIME looks at explaining the behavior of a black-box model only in the proximity of the instance z rather than providing global explanations. This feature ultimately limits the applicability of LIME when considering models for dynamical systems.

To address this limitation, we introduce PieceWise Affine LIME (PWA-LIME), which gives insights into multiple instances of the black-box model at once. The key idea of PWA-LIME is to divide the explanation process into two steps, as schematized in Figure 1:

[2]The zero-mean Huber density function is defined as the mixture of a normal distribution $\mathcal{N}(0,I)$ truncated to the interval $[-\delta,\delta]$ and a Laplace distribution with scale δ truncated to $[-\delta,\delta]^c$.

Parameter	T	m	g	l	J	τ	K_m
Value	0.005	0.07	9.81	0.042	$2.2{\cdot}10^{-4}$	0.6	15.3
Unit	[s]	[kg]	[m/s^2]	[m]	[kg m^2]	[-]	[-]

TABLE I: Parameters of the unbalanced disk.

1) divide the (large) synthetic dataset $\mathcal{D}_N$ into η clusters, hence automatically defining the instance of interest as the centroid of the cluster;
2) explain each centroid by solving (5), thus providing multiple explanations at different locations.

Note that, since the number of clusters η remains a user-chosen hyperparameter, PWA-LIME produces *multi-local explanations* rather than global ones. At the same time, it allows for a further level of interpretability of the black-box opaque model with respect to conventional LIME.

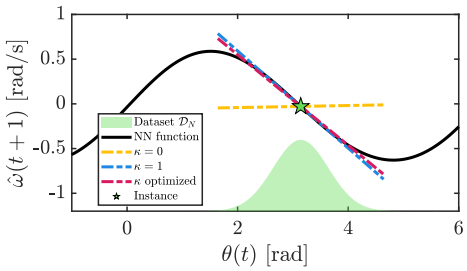
To partition the synthetic dataset, PWA-LIME relies on the piecewise affine (PWA) regression algorithm proposed in [19, Algorithm 1]. Hence, synthetic data are iteratively clustered based on (i) the proximity of z_i, for $i = 1,\dots,N$, to sequentially refined centroids in the input instance space, and (ii) the prediction accuracy of a local, affine model associated to each centroid. During clustering, these affine models are also iteratively refined.

Remark 2 (LIME and multi-local explainations): PWA-LIME is not the first extension of LIME to multiple explanations. For instance, [22] proposed to generate explanations for each synthetic sample and merge similar ones to globally characterize the opaque function. In contrast, PWA-LIME utilizes the surrogate model's structure during clustering, directly yielding meaningful explanations without post-hoc operations.

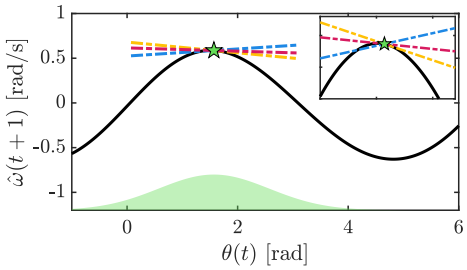
Remark 3 (PWA regression beyond clustering): When considering the surrogate model class $\mathcal{G}$ in (7), the PWA algorithm in [19, Algorithm 1] retrieves a model within $\mathcal{G}$ describing the *average* intra-cluster behavior. Without a proximity measure, all samples within a cluster are weighted equally to build the local surrogate models. This information can be compared with LIME's local explanation for each cluster to gain insights into the quality of the opaque function's explanations, as discussed in Section V-B.

V. A BENCHMARK DYNAMICAL CASE STUDY: THE UNBALANCED DISK

To provide a preliminary validation of our approaches, we take as an illustrative example the unbalanced disk system considered, among others, in [9]. The "true" discretized



(a) First scenario.



(b) Second scenario.

Fig. 2: Explanation of the relationship between $\theta(t)$ and $\hat{\omega}(t+1)$ over two experiments.

dynamics of the system is given by the difference equations:

$$\theta(t+1) = \theta(t) + T\omega(t),$$
$$\omega(t+1) = T\frac{mgl}{J}\sin(\theta(t)) + \left(1 - \frac{T}{\tau}\right)\omega(t) + T\frac{K_m}{\tau}u(t), \tag{14}$$

where $\theta(t)$ [rad] and $\omega(t)$ [rad/s] are the angular position and speed of the disk at time $t \in \mathbb{N}_0$, respectively, $u(t)$ [V] is the input voltage to the disk, while the system's parameters are reported in Table I. Under the assumption that (14) is unknown, we model the unbalanced disk via a feedforward neural network with $\tanh$ activation functions and 3 hidden layers, having 4, 8, and 4 neurons, respectively. This neural network represents the black-box opaque function $\hat{y} = f(z)$, where $\hat{y} = [\hat{\theta}(t+1) \quad \hat{\omega}(t+1)]^\top$ is the one-step-ahead prediction of both the angular position and velocity, and $z = [\theta(t) \quad \omega(t) \quad u(t)]^\top$. To this end, the surrogate model we learn with LIME belongs to the class of fully observable, affine, state-space models.

A. Optimizing the kernel width with AutoLIME

Setting $\delta = 1$ in (13), we first assess the effectiveness of AutoLIME in calibrating the kernel width κ. To do so, we compare the explanations resulting from the surrogate model with κ solving (13) with those obtained with two constant kernel widths: $\kappa = 0$, which equally weights each sample of the synthetic dataset; $\kappa = 1$, leading to a standard, non-optimized Gaussian kernel (9). Such a comparison is performed over two different synthetic dataset distributions $\mathcal{D}_N$ and instances to be explained.

In the first scenario, we explain the neural network at instance $z = [\pi \ 0 \ 0]^\top$ by drawing $N = 1000$ samples from $\mathcal{N}(z, \sigma^2)$ with $\sigma = \text{diag}(0.5, 5, 2)$. As shown in Figure 2a, this choice represents an almost linear region

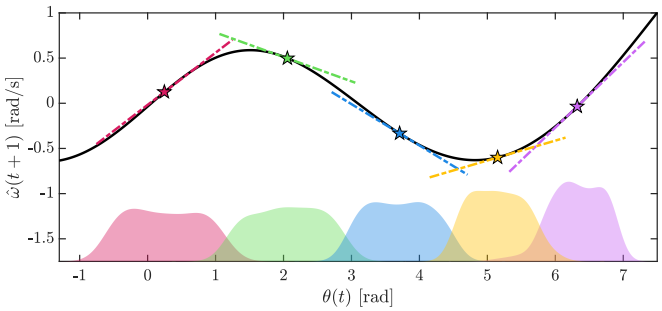


Fig. 3: PWA-LIME explanation of the angular speed dynamics (solid black) with 5 clusters. The shaded areas represent the angular position data distribution within each cluster.

of the neural network function, suggesting that an optimal algorithm should select a small κ^* for a large neighborhood, ensuring both adherence and explanation stability. In the second scenario, we explain $z = [\pi/2 \ 0 \ 0]^\top$ using a wider distribution, with $N = 1000$ samples from $\mathcal{N}(z, \sigma^2)$ and $\sigma = \text{diag}(1, 10, 4)$. Figure 2b shows this instance lies in a highly nonlinear region, requiring a smaller neighborhood and higher κ values for effective explanation.

In the first scenario, the bi-level optimization yields $\kappa^* = 0.256$, which defines a larger neighborhood compared to $\kappa = 1$, i.e., a more stable explanation, without losing local adherence. In the second scenario, AutoLIME aligns again with the expectations for the more complex nonlinear case. In particular, to achieve a higher accuracy compared to $\kappa = 1$, AutoLIME sets the optimal kernel width to $\kappa^* = 3.377$, thus defining a narrower neighborhood. These results show that the proposed optimization effectively tailors the kernel size to the synthetic dataset's explanatory capability.

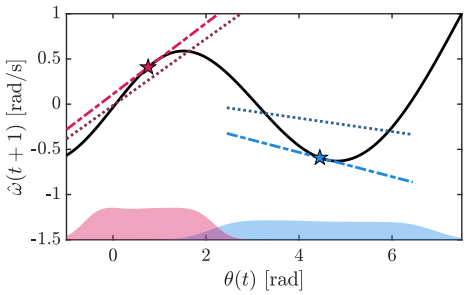
Remark 4: To analyze the effectiveness of XAI methods in explaining time series, [23] introduced the perturbation analysis framework, which relies on the premise that the feature space of each instance z is large and only a small portion of the feature space is relevant for an explanation. However, this assumption does not hold for AutoLIME. Indeed, the instance z stacks the state and input of a dynamical system, and, therefore, already defines a small (and possibly minimal) description of the opaque model.

B. Multiple explanations with PWA-LIME

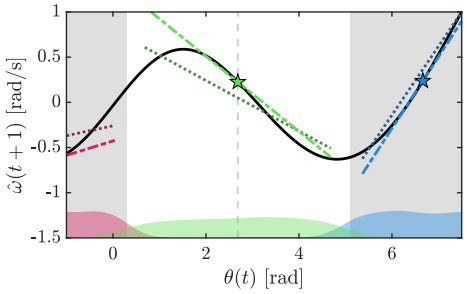
Since the unbalanced disk is a dynamical system, multi-instance explanations of its black-box model are particularly useful for analysis and control. We demonstrate this with PWA-LIME.

A synthetic dataset of $N = 2000$ samples is generated from a uniform distribution over $[-\frac{\pi}{2}, \frac{5\pi}{2}] \times [-0.5, 0.5] \times [-0.5, 0.5]$, focusing on the relation between $\theta(t)$ and $\hat{\omega}(t+1)$. With $\eta = 5$ clusters, PWA regression clustering and the bilevel optimization (5) (with $\delta = 1$) yield the results in Figure 3, where PWA-LIME captures dominant trends and provides accurate local approximations.

Reducing to $\eta = 2$ clusters highlights the limitations of centroid-based surrogates [19]. As shown in Figure 4a,



(a) Offset (blue cluster).



(b) Different slopes (green cluster).

Fig. 4: Examples of cluster explanations obtained by comparing the PWA-LIME surrogate model (dash-dotted lines) with the result of [19, Algorithm 1] (dotted lines).

nonlinearities produce a mismatch between PWA regression (average behavior) and PWA-LIME (local explanations). The discrepancy stems from the shift between the function's center of mass and the cluster centroid, while slopes remain aligned.

Two observations follow. Offsets reveal information on the distribution of the function around the centroid—for instance, in the blue cluster of Figure 4a, the centroid lies at a minimum, shifting the center of mass upward. Moreover, slope similarity indicates local symmetry and limited influence of excluded points (e.g., magenta cluster in Figure 4a, blue cluster in Figure 4b). In contrast, slope differences, as in the green cluster of Figure 4b, signal centroids aligned with inflection points. These mismatches provide structural insights into the black-box model within each cluster.

VI. Conclusions

In this paper, we addressed a key limitation of LIME by formulating a bi-level optimization problem to tune the hyperparameter of its kernel-based proximity measure systematically. Preliminary numerical results demonstrate that AutoLIME seeks a trade-off between accuracy of explanation and algorithmic stability. Additionally, we propose Piece-Wise Affine LIME (PWA-LIME), a novel extension enabling multi-instance explanations. Numerical analysis highlights its ability to provide insights into the local nonlinearity of the interpreted black-box model.

Future work will include the integration of temporal information into PWA-LIME for sequential and time-series data. In this direction, possible extensions include selecting different surrogate classes $\mathcal{G}$, e.g., ARX models.

References

[1] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 3043–3101, 2024.

[2] S. Warnick, D. Materassi, K. Vemuru, F. Vatan, P. Petrillo, A. Kamara, and B. Henz, "Explainable AI: motivations and connections with system identification," *IFAC-PapersOnLine*, vol. 58, no. 15, pp. 502–507, 2024.

[3] D. Biparva and D. Materassi, "Application of eXplainable AI and causal inference methods to estimation algorithms in networks of dynamic systems," in *2023 American Control Conference (ACC)*, 2023, pp. 1889–1894.

[4] G. Riva and S. Formentin, "Towards explainabile data-driven control (XDDC): The property-preserving framework," *IEEE Control Systems Letters*, 2024.

[5] M. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[6] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022.

[7] M. Zafar and N. Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *arXiv preprint arXiv:1906.10263*, 2019.

[8] S. Shankaranarayana and D. Runje, "ALIME: Autoencoder based approach for local interpretability," in *Intelligent Data Engineering and Automated Learning – IDEAL 2019*, 2019, pp. 454–463.

[9] C. Verhoek, H. Abbas, R. Tóth, and S. Haesaert, "Data-driven predictive control for linear parameter-varying systems," *IFAC-PapersOnLine*, vol. 54, no. 8, pp. 101–108, 2021, 4th IFAC Workshop on Linear Parameter Varying Systems LPVS 2021.

[10] P. Huber, "Robust statistics," in *International Encyclopedia of Statistical Science*. Springer, 2011, pp. 1248–1251.

[11] Z. Zhou, G. Hooker, and F. Wang, "S-LIME: Stabilized-lime for model explanation," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2429–2438.

[12] R. Gaudel, L. Galárraga, J. Delaunay, L. Rozé, and V. Bhargava, "s-LIME: Reconciling locality and fidelity in linear explanations," in *Advances in Intelligent Data Analysis XX*, 2022, pp. 102–114.

[13] G. Visani, E. Bagli, and F. Chesani, "OptiLIME: Optimized LIME explanations for diagnostic computer algorithms," *arXiv preprint arXiv:2006.05714*, 2022.

[14] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistical learning with sparsity," *Monographs on statistics and applied probability*, vol. 143, no. 143, p. 8, 2015.

[15] M. Köhler, A. Schindler, and S. Sperlich, "A review and comparison of bandwidth selection methods for kernel regression," *International Statistical Review*, vol. 82, no. 2, pp. 243–274, 2014.

[16] A. Gajewicz-Skretna, S. Kar, M. Piotrowska, and J. Leszczynski, "The kernel-weighted local polynomial regression (KwLPR) approach: an efficient, novel tool for development of QSAR/QSAAR toxicity extrapolation models," *Journal of Cheminformatics*, vol. 13, no. 1, p. 9, 2021.

[17] T. Koç, "Bandwidth selection in geographically weighted regression models via information complexity criteria," *Journal of Mathematics*, vol. 2022, no. 1, p. 1527407, 2022.

[18] H. Khalil, *Nonlinear systems; 3rd ed.* Upper Saddle River, NJ: Prentice-Hall, 2002.

[19] V. Breschi, D. Piga, and A. Bemporad, "Piecewise affine regression via recursive multiple least squares and multicategory discrimination," *Automatica*, vol. 73, pp. 155–162, 2016.

[20] Y. Tong, *The Multivariate Normal Distribution*. Springer New York, NY, 1990.

[21] D. Biparva and D. Materassi, "Interpretation of explainable AI methods as identification of local linearized models," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 2383–2388, 2023, 22nd IFAC World Congress.

[22] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "ILIME: Local and global interpretable model-agnostic explainer of black-box decision," in *Advances in Databases and Information Systems*, 2019, pp. 53–68.

[23] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of XAI methods on time series," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 4197–4201.