# A Learning Theoretic Perspective on Local Explainability

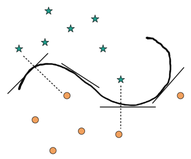
Jeffrey Li*[1] Vaishnavh Nagarajan*[2] Gregory Plumb[2] Ameet Talwalkar[23]

[1]University of Washington [2]Carnegie Mellon University [3]Determined AI

## Our Contributions

Interpretability has been a largely empirical field.

We establish one of the first connections between interpretability and learning theory:



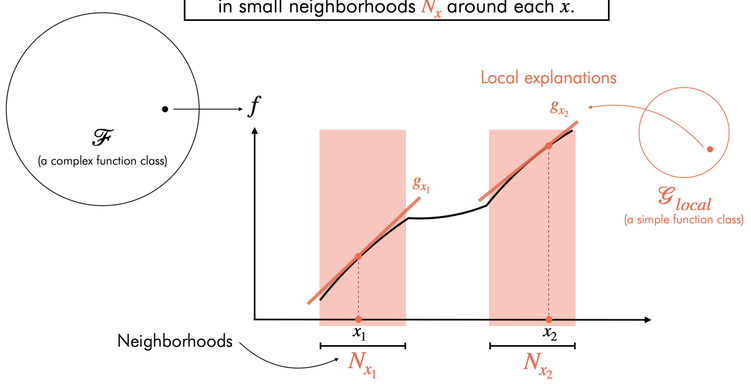
1. We derive a generalization bound on test performance of a model in terms of its local explainability.

2. We address a new question: how well does the "quality" of local explanations generalize?

**KEY IDEA** Interpretability of a model ⟷ Complexity of the model ⟷ Generalization properties of the model

## Background: Local Explanations

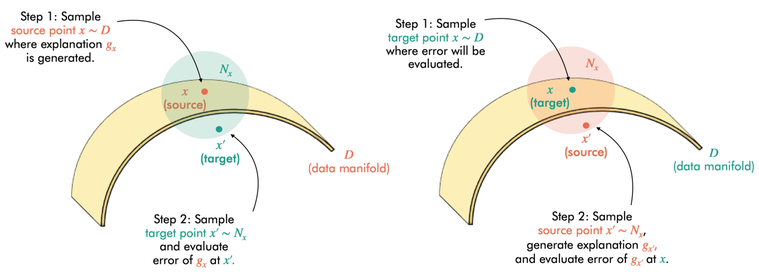
Approximate a function $f$ via simple functions $g_x$ in small neighborhoods $N_x$ around each $x$.




## Local Explanation Quality

Standard measure: Neighborhood Fidelity (NF)
$\mathbb{E}_{x\sim D}\mathbb{E}_{x'\sim N_x}[(g_x(x') - f(x'))^2]$

We propose: Mirrored Neighborhood Fidelity (MNF)
$\mathbb{E}_{x\sim D}\mathbb{E}_{x'\sim N_x}[(g_x(x) - f(x))^2]$




Unlike NF, MNF does not evaluate fit of explanations on off-manifold data!

$\Longrightarrow$ more amenable to theoretical analysis
$\Longrightarrow$ more robust to irregular off-manifold behavior of $f$.

## Performance Generalization

Motivation: Bounds based on complexity($\mathcal{F}$) may not capture how "simple" $f$ is!

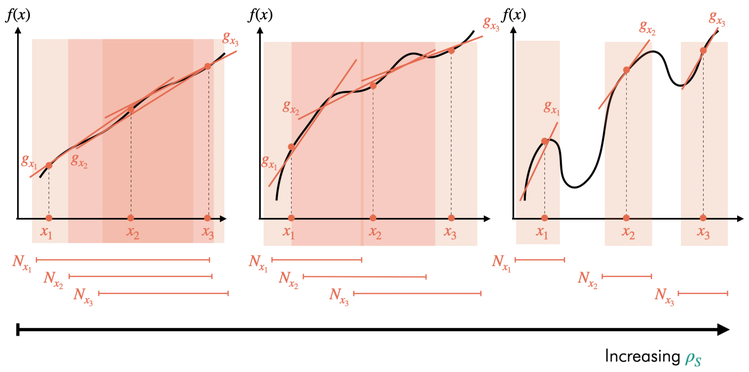
**Theorem 1:**

$$\mathbb{E}_D[(f(x)-y)^2] \le \hat{\mathbb{E}}_S[(f(x)-y)^2] + \mathbb{E}_{x\sim D}\mathbb{E}_{x'\sim N_x}[(g_x(x)-f(x))^2] + \rho_S \cdot \mathcal{R}(\mathcal{G}_{local})$$

Test loss | Train loss (on dataset $S$) | Explanation quality (MNF) | Complexity of the system of local explanations

$\mathcal{R}(\mathcal{G}_{local})$ = Rademacher complexity of $\mathcal{G}_{local}$ $\approx O\left(\frac{1}{\sqrt{|S|}}\right)$ — How complex each local explanation is

$\rho_S = \int_{x'\in\mathcal{X}} \sqrt{\frac{1}{|S|}\sum_{x\in S}(p_{N_x}(x'))^2 dx'}$ $\in [1, \sqrt{|S|}]$ — How disjoint the neighborhoods of the training points are

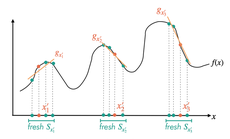



Takeaway: Better generalization when it is easier to locally approximate $f$ on larger neighborhoods.

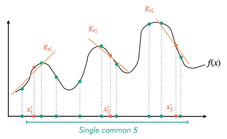
Small neighborhoods ⟷ Key Tradeoff ⟷ Large neighborhoods

MNF is small, $\rho_S$ is large — MNF is large, $\rho_S$ is small

## Explanation Generalization: Motivation




Canonical approaches learn explanations from "infinite" data (e.g., LIME [1])!




Recent approaches (MAPLE [2], RL-LIM [3]) reuse finite dataset (by re-weighting it each time)!

Finite-sample-based approaches could potentially "overfit their explanations"!

What determines the quality of these explanations on unseen data?

## Explanation Generalization: Result

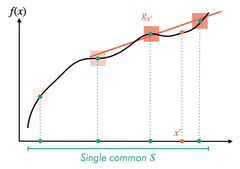
**Theorem 2:**

$$\mathbb{E}_{x\sim D}\mathbb{E}_{x'\sim N_x}[(g_x(x)-f(x))^2] \le \hat{\mathbb{E}}_{x\sim S}\mathbb{E}_{x'\sim N_x}[(g_x(x)-f(x))^2] + \rho_S \cdot \mathcal{R}(\mathcal{G}_{local})$$

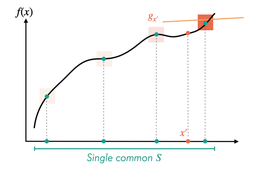
Test explanation quality (Test MNF) | Train explanation quality (Train MNF) | Complexity of the system of explanations

Takeaway: Better generalization when explanations can nicely fit training data that fall in a larger neighborhood.

Larger neighborhood $\Longrightarrow$ explanation is trained on more points $\Longrightarrow$ better generalization

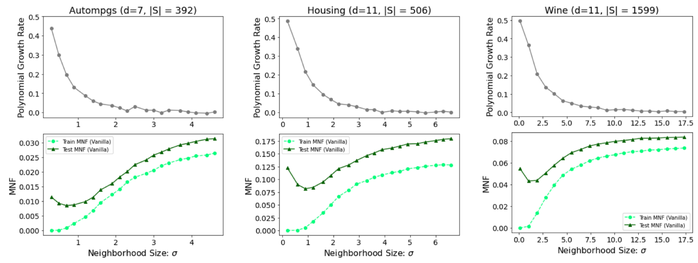



Smaller neighborhood $\Longrightarrow$ explanation is trained on too few points $\Longrightarrow$ poorer generalization




## Experiments

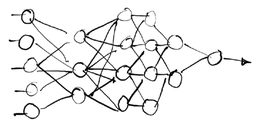
- Are there neighborhood widths s.t. $\rho_S = o(|S|^{0.5})$ while Train MNF is small? Yes!
- Do wider neighborhoods bring the generalization gap down? Yes!




Exponent of $\rho_S$ (top) and test & train MNF (bottom) on UCI datasets

## Future Work

1. Extend bounds to high-dimensional datasets.
2. Explore these bounds for NNs.
3. When is MNF better than NF in practice?



## References

1. *"Why should I trust you?: Explaining the predictions of any classifier."* Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ACM SIGKDD, 2016.
2. "*Model Agnostic Supervised Local Explanations*", Gregory Plumb, Denali Molitor and Ameet S. Talwalkar, NeurIPS 2018
3. *"RL-LIM: Reinforcement learning-based locally interpretable modeling"*, Jinsung Yoon, Sercan O. Arik, and Tomas Pfister, 2019
4. *"Uniform convergence may be unable to explain generalization in deep learning"*. Vaishnavh Nagarajan and J. Zico Kolter, NeurIPS 2019
5. *"Understanding deep learning requires rethinking generalization"*, Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals, ICLR' 17