A LEARNING THEORETIC PERSPECTIVE ON LOCAL EXPLAINABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we explore connections between interpretable machine learning and learning theory through the lens of *local approximation* explanations. First, we tackle the traditional problem of *performance generalization* and bound the test-time accuracy of a model using a notion of how locally explainable it is. Second, we explore the novel problem of *explanation generalization* which is an important concern for a growing class of *finite sample-based* local approximation explanations. Finally, we validate our theoretical results empirically and show that they reflect what can be seen in practice.

1 INTRODUCTION

There has been a growing interest in interpretable machine learning, which seeks to help people understand their models. While interpretable machine learning encompasses a wide range of problems, it is a fairly uncontroversial hypothesis that there exists a trade-off between a *model's complexity* and *general notions of interpretability*. This hypothesis suggests a seemingly natural connection to the field of learning theory, which has thoroughly explored relationships between a *function class's complexity* and *generalization*. However, formal connections between interpretability and learning theory remain relatively unstudied.

Though there are several notions of conveying interpretability, one common and flexible approach is to use local approximations. Formally, *local approximation explanations* (which we will refer to as "local explanations") provide insight into a model's behavior as follows: for any black-box model $f \in \mathcal{F}$ and input x, the explanation system produces a simple function, $g_x(x') \in \mathcal{G}_{\text{local}}$, which approximates f in a specified neighborhood, $x' \sim N_x$. Crucially, the freedom to specify both $\mathcal{G}_{\text{local}}$ and N_x grants local explanations great versatility. In this paper, we provide two connections between learning theory and how well f can be approximated locally (i.e. the *fidelity* of local explanations).

Our first result studies the standard problem of *performance generalization* by relating test-time performance to a notion of local interpretability. As it turns out, our focus on local explanations leads us to unique tools and insights from a learning theory point of view. Our second result identifies and addresses an unstudied – yet important – question regarding *explanation generalization*. This question pertains to a growing class of explanation systems, such as MAPLE (Plumb et al., 2018) and RL-LIM (Yoon et al., 2019), which we call *finite sample-based* local explanations[1]. These methods learn their local approximations using a common finite sample drawn from D (in contrast to local approximation methods such as LIME (Ribeiro et al., 2016)) and, as a result, run the risk of overfitting to this finite sample. In light of this, we answer the following question: for these explanation-learning systems, how well do they generalize to data not seen during training?

We address these questions with two bounds, which we outline now. Regarding performance generalization, we derive our first main result, Theorem 1, which bounds the expected test mean squared error (MSE) of any f in terms of its MSE over the m samples in the training set, $S = \{(x_i, y_i)\}_{i=1}^m$:

$$\underbrace{\mathbb{E}_{(x,y)\sim D}[(f(x)-y)^2]}_{\text{Test MSE}} \leq \tilde{\mathcal{O}}\Big(\underbrace{\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2}_{\text{Train MSE}} + \underbrace{\mathbb{E}_{\substack{x\sim D,\\ x'\sim N_x}}\left[(g_{x'}(x)-f(x))^2\right]}_{\text{Interpretability Term (MNF)}} + \underbrace{\rho_S \hat{\mathcal{R}}_S(\mathcal{G}_{\text{local}})}_{\text{Complexity Term}} \Big)$$

[1]This terminology is not to be confused with example-based explanations where the explanation itself is in the form of data instances rather than a function.

Regarding explanation generalization for finite sample-based explanation-learning systems, we apply a similar proof technique to obtain Theorem 2, which bounds the quality of the system's explanations on unseen data in terms of their quality on the data on which the system was trained:

$$\underbrace{\mathbb{E}_{\substack{x\sim D,\\ x'\sim N_x}}\left[(g_{x'}(x)-f(x))^2\right]}_{\text{Test MNF}} \leq \underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{x'\sim N_{x_i}}\left[(f(x_i)-g_{x'}(x_i))^2\right]}_{\text{Train MNF}} + \underbrace{\tilde{\mathcal{O}}\left(\rho_S\hat{\mathcal{R}}_S(\mathcal{G}_{\text{local}})\right)}_{\text{Complexity Term}}$$

Before summarizing our contributions, we discuss the key new terms and their relationship.

- *Interpretability terms:* The terms involving MNF correspond to Mirrored Neighborhood Fidelity, a metric we use to measure local explanation quality. As we discuss in Section 3, this is a reasonable modification of the commonly used Neighborhood Fidelity (NF) metric (Ribeiro et al., 2016; Plumb et al., 2018). Intuitively, we generally expect MNF to be larger when the neighborhood sizes are larger since the $g_{x'}$ are required to extrapolate farther.
- *Complexity term:* This term measures the complexity of the local explanation system g in terms of (a) the complexity of the local explanation class $\mathcal{G}_{\text{local}}$ and (b) ρ_S, a quantity that we define and refer to as *neighborhood disjointedness factor*. As we discuss in Section 4, ρ_S is a value in $[1, \sqrt{m}]$ (where $m = |S|$) that is proportional to the level of disjointedness of the neighborhoods for points in the sample S. Intuitively, we expect ρ_S to be larger when the neighborhoods sizes are smaller since smaller neighborhoods will overlap less.

Notably, both our bounds capture the following key trade-off: as neighborhood widths increase, MNF increases but ρ_S decreases. As such, our bounds are non-trivial only if N_x can be chosen such that MNF remains small but ρ_S grows slower than $\tilde{\mathcal{O}}(\sqrt{m})$ (since $\hat{\mathcal{R}}_S(\mathcal{G}_{\text{local}})$ decays as $\tilde{\mathcal{O}}(1/\sqrt{m})$).

We summarize our main contributions as follows:

(1) We make a novel connection between performance generalization and local explainability, arriving at Theorem 1. Given the relationship between MNF and ρ_S, this bound roughly captures that an easier-to-interpret f enjoys better generalization guarantees, a potentially valuable result when reasoning about $\mathcal{F}$ is difficult (e.g. for neural networks). Further, our proof technique may be of independent theoretical interest as it provides a new way to bound the Rademacher complexity of a *randomized* function (see Section 4).

(2) We motivate and explore an important generalization question about expected explanation quality. Specifically, we arrive at Theorem 2, a bound for test MNF in terms of training MNF. This bound suggests that practitioners can better guarantee good local explanation quality (measured by MNF) using methods which encourage the neighborhood widths to be wider (see Section 5).

(3) We verify empirically on UCI Regression datasets that our results non-trivially reflect the two types of generalization in practice. First, we demonstrate that ρ can indeed exhibit slower than $\tilde{\mathcal{O}}(\sqrt{m})$ growth without significantly increasing the MNF terms. Also, for Theorem 2, we show that the generalization gap indeed improves with larger neighborhoods (see Section 6).

(4) To aid in our theoretical results, we propose MNF as a novel measure of local explainability. This metric naturally complements NF and offers a unique advantage over NF when evaluating local explanations on "realistic" on-distribution data (see Section 3).

2 Related Work

Interpretability meets learning theory: Semenova et al. (2019) study the performance generalization of models learned from complex classes when they can be globally well-approximated by simpler (e.g. interpretable) classes. In such cases, their theory argues that if the complex class has many models that perform about as optimally on training data, generalization from the complex class can be expressed in terms of the simpler class's complexity. In our corresponding results, we similarly aim to avoid involving the larger class's complexity. However, we directly study generalization via a function's local explainability, rather than instantiate abstract "complex" and "simple" classes for global approximations. The two are fundamentally different technical problems; standard learning theory results cannot be directly applied as they are for single-function global approximations.

Statistical localized regression: (Fan, 1993; Fan & Gijbels) are canonical results which bound the squared error of a nonparametric function defined using locally fit models. These local models are both simple (e.g. linear) and similarly trained by weighting real examples with a kernel (i.e. neighborhood). However, in these works, each local model is only used to make a prediction at its source point and the theory requires shrinking the kernel width towards 0 as the sample size grows. We instead fit local models *as explanations* for a trained model (i.e. which is considered the "true regression function") and more importantly, care about the performance of each local model over whole (non-zero) neighborhoods. Unlike localized regression, this allows us to use uniform convergence to bound test error with empirical and generalization terms. While the previous results do not have empirical terms, the learning rates are exponential in the number of samples.

Learning Theory: One line of related work also studies how to explain generalization of overparameterized classes. As standard uniform convergence on these classes often leads to vacuous bounds, a general approach that has followed from (Nagarajan & Kolter, 2019; Zhang et al.; Neyshabur et al., 2014) has been to study implications of different biases placed on the learned models. We study what would happen if an overparameterized model had an unexplored type of bias, one that is inspired by local explainability. Additionally, our work's technical approach also parallels another line of existing results which likewise try to apply uniform convergence on a separate surrogate class. This includes PAC-Bayesian bounds, a large family of techniques that come from looking at a stochastic version of in parameter space (McAllester, 1998; 2003; Langford & Caruana, 2002; Langford & Shawe-Taylor, 2003). In a different vein, some results in deep learning look at compressed/sparsified/explicitly regularized surrogates of neural networks (Arora et al., 2018; Dziugaite & Roy, 2017). In our case, the surrogate class is a collection of local explanations.

3 Mirrored Neighborhood Fidelity

In order to connect local explanations to generalization, recall that we study a measure of local interpretability which we call "mirrored neighborhood fidelity" (MNF). As we explain below, this quantity comes from a slight modification to an existing measure of interpretability, namely, that of neighborhood fidelity (NF).

To define our terms, we use the following notations. Let $\mathcal{X}$ be an input space and let D be a distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} \subseteq \mathbb{R}$. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathcal{Y}$. For our theoretical results, we specifically assume that $\mathcal{Y}$ is bounded as $\mathcal{Y} = [-B, B]$ for some $B > 0$ (though this does not matter until following sections). In order to provide local explanations, we need to fix a local neighborhood around each $x \in \mathcal{X}$. To this end, for any x, let N_x correspond to some distribution denoting a local neighborhood at x e.g., typically this is chosen to be a distribution centered at x. For any distribution N, we use $p_N(x)$ to denote its density at x. Now, let $\mathcal{G}$ be a class of explainers $g : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ such that for each $x \in \mathcal{X}$, the local explanation $g(x, \cdot) : \mathcal{X} \to \mathcal{Y}$ belongs to a class of (simple) functions (e.g. linear), $\mathcal{G}_{\text{local}}$. In short, we denote $g(x, \cdot)$ as $g_x(\cdot)$ and we'll use $g(x, \cdot)$ to locally approximate f in the neighborhood defined by N_x.

The accuracy of this local approximation is usually quantified by a term called "neighbhorhood fidelity" which is defined as follows (Ribeiro et al., 2016; 2018; Plumb et al., 2018; 2020)

$$\mathsf{NF}(f, g) := \mathbb{E}_{x \sim D} \left[\mathbb{E}_{x' \sim N_x} \left[(f(x') - g_x(x'))^2 \right] \right].$$

To verbally interpet this, let us call x as the "source" point which gives rise to a local explanation $g_x(\cdot)$ and x' the "target" point that we try to fit using g. To compute $\mathsf{NF}(f, g)$, we need to do the following: for each source point x, we first compute the average error in the fit of $g_x(\cdot)$ over target points x' in the local neighborhood of the source point x (i.e., N_x); then, we globally average this error across draws of the source point $x \sim D$.

Now, to define MNF, we take the same expression as NF but swap x and x' within the innermost expectation (without modifying the expectations). In other words, we now sample *a target point* x from D, and sample *a source point* x' from a distribution over points near x. Since this distribution is over source points rather than target points, just for the sake of distinguishing, we'll call this a *mirrored* neighborhood distribution and denote it as N_x^{mir}. Formally we define this measure of local interpretability below, following which we explain how to understand it:

Definition 3.1. ***(Mirrored Neighborhood Fidelity)*** *We define* $\mathsf{MNF} : \mathcal{F} \times \mathcal{G} \to \mathbb{R}$ *as*

$$\mathsf{MNF}(f, g) := \mathbb{E}_{x \sim D} \left[\mathbb{E}_{x' \sim N_x^{\mathsf{mir}}} \left[(f(x) - g_{x'}(x))^2 \right] \right].$$

and with an abuse of notation, we let $\mathsf{MNF}(f, g, x) := \mathbb{E}_{x' \sim N_x^{\mathsf{mir}}} \left[(f(x) - g_{x'}(x))^2 \right]$.

Understanding MNF. It is helpful to parse the expression for MNF in two different ways. First, we can think of it as measuring the error in approximating every target point $x \in \mathcal{X}$ through a *randomized* locally-approximating function $g_{x'}(\cdot)$ where x' is randomly drawn from the local neighborhood N_x^{mir}. A second way to parse this is in a manner similar to how we parsed NF. To do this, first we note that the expectations in MNF can be swapped around and rewritten equivalently as follows:

$$\mathsf{MNF}(f, g) = \mathbb{E}_{x' \sim D^\dagger} \left[\mathbb{E}_{x \sim N_{x'}^\dagger} \left[(f(x) - g_{x'}(x))^2 \right] \right],$$

where $D^\dagger$ and $N_{x'}^\dagger$ are suitably defined distributions (defined in Appendix A) that can be thought of as modified counterparts of D and $N_{x'}^{\mathsf{mir}}$ respectively. With this rewritten expression, one can read MNF like NF: for each source point (here that is x'), we compute the average error in the fit of the corresponding local function ($g_{x'}(\cdot)$) over target points (x) in the local neighborhood of the source point ($N_{x'}^\dagger$); this error is then globally averaged over different values of the source point ($x' \sim D^\dagger$).

While both NF and MNF are closely related measures of local interpretability for f, studying MNF allows us to make connections between local interpretability and different notions of generalization (Sections 4 and 5). Furthermore, MNF may also be of interest to the interpretability community, as it offers a unique advantage over NF when the intended usage of local explanations is centered around understanding how the model works on the specific learning task it was trained on.

Specifically, we argue that selecting the target point distribution to be D rather than D perturbed by N_x (as for NF) better emphasizes the ability for explanations to accurately convey how well g will predict at *realistic* points. This is relevant for ML (and deep learning particularly) because (a) high-dimensional datasets often exhibit significant feature dependencies and adherence to lower dimensional manifolds; (b) f can often be highly unpredictable and unstable when extrapolating beyond the training data. As such, when one measures NF with standard neighborhood choices that ignore feature dependencies (i.e. most commonly $N_x = \mathcal{N}(x, \sigma I)$), the resulting target distribution may concentrate significantly on regions that are non-relevant to the actual task at hand. As can be shown, this can lead to overemphasis on fitting noisy off-manifold behavior, deteriorating the fit of explanations relative to task-relevant input regions (we defer a more detailed presentation of this point, as well as other trade-offs between NF and MNF to Appendix A).

4 Generalization of Model Performance via MNF

The generalization error of the function f is typically bounded by some notion of the representational capacity/complexity of f. While standard results bound complexity in terms of parameter counts, there is theoretical value in deriving bounds involving other novel terms. By doing so, we can understand how regularizing for those terms can affect the representation capacity, and in turn, the generalization error of f. Especially when f's complexity may be intractable to bound on its own, introducing these terms provides a potentially useful new way to understand f's generalization.

Here specifically, we are interested in establishing a general connection between the representation complexity and the local intrepretability of *any* f. This naturally requires coming up with a notion that appropriately quantifies the complexity of $\mathcal{G}$, which we discuss in the first part of this section. In the second part, we then relate this quantity to the generalization of f to derive our first main result.

Key technical challenge: bounding the complexity of $\mathcal{G}$. The overall idea behind how one could tie the notions of generalization and local interpretability is fairly intuitive. For example, consider a simplified setting where we approximate f by dividing $\mathcal{X}$ into K disjoint pieces/neighborhoods, and then approximating each neighborhood via a simple (say, linear) model. Then, one could bound the generalization error of f as the sum of two quantities: first, the error in approximating f via the piecewise linear model, and second, a term involving the complexity of the piecewise linear model. It is straightforward to show that this complexity grows polynomially with the piece-count, K, and also the complexity of the simple local approximator (see Appendix C.0.1). Similarly, one could hope to bound the generalization error of f in terms of $\mathsf{MNF}(f, g)$ and the complexity of $\mathcal{G}$. However, the key challenge here is that the class $\mathcal{G}$ is a much more complex class than the above class of piecewise linear models. For example, a straightforward piece-count-based complexity bound would be infinitely large since there are effectively infinitely many unique pieces in g.

Our core technical contribution here is to bound the Rademacher complexity of $\mathcal{G}$ in this more complex local-interpretability setting. At a high level, the resulting bound (which will be stated shortly) grows with "the level of overlap" between the neighborhoods $\{N_x^{\mathsf{mir}} | x \in \mathcal{X}\}$, quantified as:

Definition 4.1. *Given a dataset $S \in (\mathcal{X} \times \mathcal{Y})^m$, we define the **disjointedness factor** ρ_S as*

$$\rho_S := \int_{x' \in \mathcal{X}} \sqrt{\frac{1}{m} \sum_{i=1}^{m} (p_{N_{x_i}^{\mathsf{mir}}}(x'))^2} dx'$$

Understanding the disjointedness factor. ρ_S can be interpreted as bounding the "effective number" of pieces induced by the set of neighborhood distributions $\{N_x^{\mathsf{mir}} | x \in \mathcal{X}\}$. This turns out to be a quantity that lies in $[1, \sqrt{m}]$ (shown formally in Appendix Fact B.1). To intuit about this quantity, it is helpful to consider its behavior in extreme scenarios. First, consider the case where N_x^{mir} is the same distribution (say N) regardless of x i.e., neighborhoods are completely overlapping. Then, $\rho_S = \int_{x' \in \mathcal{X}} (p_N(x')) dx' = 1$. In the other extreme, consider if neighborhoods centered on the training data are all disjoint with supports $\mathcal{X}_1, \ldots, \mathcal{X}_{|S|}$. Here the integral splits into m summands as: $\rho_S = \sum_{i=1}^{m} \int_{x' \in \mathcal{X}_i} \frac{1}{\sqrt{m}} p_{N_{x_i}^{\mathsf{mir}}}(x') dx' = \sqrt{m}$. Thus, intuitively ρ_S grows from 1 to $\sqrt{m}$ as the level of overlap between the neighborhoods $N_{x_1}^{\mathsf{mir}}, \ldots, N_{x_{|S|}}^{\mathsf{mir}}$ reduces. For intuition at non-extreme values, we show in Appendix B.2 that in a simple setting, $\rho = \sqrt{m^{1-k}}$ (where $0 \leq k \leq 1$) if every neighborhood is just large enough to encompass a $1/m^{1-k}$ fraction of mass of the distribution D.

Rademacher complexity of $\mathcal{G}$. We now use ρ_S to bound the Rademacher complexity of $\mathcal{G}$. First, in order to define the complexity of $\mathcal{G}$, it is useful to think of g as *a randomized function*. Specifically, at any target point x, the output of g is a random variable $g_{x'}(x)$ where the randomness comes from $x' \sim N_x^{\mathsf{mir}}$. Then, in Lemma 4.1, we take this randomization into account to define and bound the complexity of $\mathcal{G}$ (which we use prove our main results). To keep our statement general, we consider a generic loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (e.g., the squared error loss is $L(y, y') = (y - y')^2$). Whenever L satisfies a standard Lipschitz assumption, we can bound the complexity of $\mathcal{G}$ composed with the loss function L, in terms of ρ_S, the complexity of $\mathcal{G}_{\text{local}}$ and the Lipschitzness of L:

Lemma 4.1. *(see Appendix Lemma D.1 for full, precise statement) Let $L(\cdot, y')$ be a c-Lipschitz function w.r.t. y' in that for all $y_1, y_2 \in [-B, B]$, $|L(y_1, y') - L(y_2, y')| \leq c|y_1 - y_2|$. Then, the empirical Rademacher complexity of $\mathcal{G}$ under the loss function L is defined and bounded as:*

$$\hat{\mathcal{R}}_S(L \circ \mathcal{G}) := \mathbb{E}_{\vec{\sigma}} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i}^{m} \sigma_i \mathbb{E}_{x' \sim N_{x_i}^{\mathsf{mir}}} [L(g_{x'}(x_i), y_i)] \right] \leq O\left(c \rho_S \hat{\mathcal{R}}_S(\mathcal{G}_{local}) \cdot \ln m \right).$$

We note that the proof technique employed here may be of independent theoretical interest as it provides a novel way to bound the complexity of a randomized function. Although techniques like PAC-Bayes provide ways to do this, they do not apply here since the stochasticity in the function is of a different form.

Main result. With the above key lemma in hand, we are now ready to prove our main result, which bounds the generalization error of f in terms of the complexity of $\mathcal{G}$, thereby establishing a connection between model generalization and local interpretability.

Theorem 1. *(see Appendix Theorem 3 for full, precise statement) With probability over $1 - \delta$ over the draws of $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \sim D^m$, for all $f \in \mathcal{F}$ and for all $g \in \mathcal{G}$, we have (ignoring $\ln 1/\delta$ factors):*

$$\mathbb{E}_{(x,y) \sim D}[(f(x) - y)^2] \leq \frac{4}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + 2 \underbrace{\mathbb{E}_{x \sim D}[\mathbb{E}_{x' \sim N_x^{\mathsf{mir}}} [(f(x) - g_{x'}(x))^2]]}_{\mathit{MNF}(f,g)}$$
$$+ \frac{4}{m} \sum_{i=1}^{m} \underbrace{\mathbb{E}_{x' \sim N_x^{\mathsf{mir}}} [(f(x_i) - g_{x'}(x_i))^2]}_{\mathit{MNF}(f,g,x_i)} + \mathcal{O}(B \rho_S \hat{\mathcal{R}}_S(\mathcal{G}_{local}) \ln m),$$

and $\hat{\mathcal{R}}_S(\mathcal{G}_{local})$ is the empirical Rademacher complexity of $\mathcal{G}_{local}$ defined as $\hat{\mathcal{R}}_S(\mathcal{G}_{local}) := \mathbb{E}_{\vec{\sigma}} \left[\sup_{h \in \mathcal{G}_{local}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right]$ where $\vec{\sigma}$ is uniformly distributed over $\{-1, 1\}^m$.

This result decomposes the test error of f into four quantities. The first quantity corresponds to the training error of f on the training set S. The second and the third correspond to the mirrored neighborhood fidelity of f with respect to g (computed on test and training data respectively). The fourth and final quantity corresponds to a term that bounds the complexity of $\mathcal{G}$ in terms of the "disjointedness factor" and the complexity of the simpler function class $\mathcal{G}_{\text{local}}$.

Takeaway. A key aspect of this bound is the trade-off that it captures with varying neighborhood widths. Consider shrinking the neighborhood widths to smaller and smaller values, in turn creating less and less overlap between the neighborhoods of the training data. Then, on the one hand, we'd observe that the complexity term (the fourth term on the R.H.S) increases. Specifically, since $\hat{\mathcal{R}}_S(\mathcal{G}_{\text{local}})$ typically scales as $O(1/\sqrt{m})$, as we go from the one extreme of full overlap to the other extreme of complete disjointedness, the complexity term would increase from $O(1/\sqrt{m})$ to $O(1)$ (eventually rendering the bound trivial). On the other hand, as the widths decrease, the fidelity terms (the second and the third term) would likely *decrease* – this is because the simple functions in $\mathcal{G}_{\text{local}}$ would find it easier and easier to approximate the shrinking neighborhoods.

This tradeoff is intuitive. A function f that is hardly amenable to being fit by local explanations would require extremely tiny neighborhoods for $\mathcal{G}_{\text{local}}$ to locally approximate it (i.e. make the MNF terms small). For example, in an extreme case, when the neighborhoods N_x^{mir} are set be point masses at x, it is trivially easy to find $g_x(\cdot) \in \mathcal{G}_{\text{local}}$ with no approximation error. Thus, the complexity term would be too large in this case, implying that a hard-to-interpret f results in bad generalization. On the other hand, when f is easy to interpret, then we'd expect it to be well-approximated by $\mathcal{G}_{\text{local}}$ even with wider neighborhoods. This allows one to afford smaller values for *both* the complexity and MNF terms. In other words, an easy-to-interpret f enjoys better generalization guarantees.

Caveats. Our bound has two limitations worth noting. First, for high-dimensional datasets (like image datasets), practical choices of N_x can lead to almost no overlap between neighborhoods, thus rendering the bound trivial in practice. This potentially poor dimension-dependence is a caveat similarly shared by bounds for non-parametric local regression, whereby increasing d results in an exponential increase in the required sample size (Fan, 1993; Fan & Gijbels). Nevertheless, for low-dimensional datasets, we show in the experiments that for practical choices of the neighborhood distributions, there *is* sufficient neighborhood overlap to achieve values of ρ_S that are $o(\sqrt{m})$.

A second caveat is that the second quantity, $\mathsf{MNF}(f, g)$, requires *unlabeled* test data to be computed, which may be limiting if one is interested in numerically computing this bound in practice. It is however possible to get a bound without this dependence, although only on the test error of g rather than f (see Appendix Theorem 4). Nevertheless, we believe that the above bound has theoretical value in how it establishes a connection between the interpretability of f and its generalization.

5 Generalization of Local Explainability

We now turn our attention to a more subtle kind of generalization that is both unstudied yet important. Typically, the way $g_{x'}$ is learned at any source point x' is by fitting a finite set of points sampled near x', with the hope that this fit generalizes to unseen, neighboring target points. Naturally, we would want to ask: how well do the explanations $g_{x'}$ themselves generalize in this sense?

The subtlety in this question is that it is not always a worthwhile question to ask. In particular, assume that we learn $g_{x'}$ by sampling a set $S_{x'}$ of nearby points from a Gaussian centered at x', and that we care about the fit of $g_{x'}$ generalizing to the same Gaussian. Here, we have access to unlimited amounts of data from the known Gaussian distribution (and free labels using $f(\cdot)$), so we can be sure that with sufficiently large $S_{x'}$, $g_{x'}$ will fit to arbitrarily small error on local neighborhoods. Hence, the above generalization question is neither conceptually nor practically interesting here.

However, consider *finite sample-based* local explanation settings like MAPLE (Plumb et al., 2018) and RL-LIM (Yoon et al., 2019) where the training procedure is vastly different from this: in these procedures, the goal is to learn local explanations $g_{x'}$ in a way that is sensitive to the local structure of the (unknown) underlying data distribution D. So, instead of fitting the $g_{x'}$ to samples drawn from an arbitrarily defined Gaussian distribution, here one first draws a finite sample S from the underlying distribution D (and then labels it using f). Then, across all $x' \in \mathcal{X}$, one reuses a reweighted version of the *same* dataset S (typically, points x in S that are near x' are weighted more) and then learns a $g_{x'}$ that fits this reweighted dataset. Contrast this with the former setting, where for each x', one

has access to a *fresh* dataset (namely, $S_{x'}$) to learn $g_{x'}$. This distinction then makes it interesting to wonder when the reuse of a common dataset S could cause the explanations to generalize poorly.

Motivated by this question, we present Theorem 2. By using Lemma 4.1, we provide a bound on the "test MNF" (which corresponds to the fit of $g_{x'}$ on the unseen data averaged across D) in terms of the "train MNF" (which corresponds to the fit of $g_{x'}$ on S, averaged across x') and the complexity term from Lemma 4.1. We must however caution the reader that this theorem does *not* answer the exact question posed in the above paragraph; it only addresses it indirectly as we discuss shortly.

Theorem 2. *(see Appendix Theorem 2-full for full, precise statement) For a fixed function f, with high probability $1-\delta$ over the draws of $S \sim D^m$, for all $g \in \mathcal{G}$, we have (ignoring $\ln 1/\delta$ factors):*

$$\underbrace{\mathbb{E}_{x\sim D}\left[\mathbb{E}_{x'\sim N_x^{mir}}\left[(f(x)-g_{x'}(x))^2\right]\right]}_{\textit{test MNF i.e., MNF}(f,g)} \leq \underbrace{\sum_{i=1}^{m}\mathbb{E}_{x'\sim N_x^{mir}}\left[(f(x_i)-g_{x'}(x_i))^2\right]}_{\textit{train MNF}} + O(\rho_S \mathcal{R}_S(\mathcal{G}_{local})\ln m).$$

Understanding the overall bound. We first elaborate on how this bound provides an (indirect) answer to our question about how well explanations generalize. Consider a procedure like MAPLE that learns g using the finite dataset S. For each $x' \in \mathcal{X}$, we would expect this procedure to have learned a $g_{x'}$ that fits well on at least those target points x in S that are near x'. In doing so, it's reasonable to expect the training procedure to have implicitly controlled the "train MNF" term The reasoning for this is that the train MNF computes the error in the fit of $g_{x'}$ on S for different values of x', and sums these up in a way that errors corresponding to nearby values of (x, x') are weighted more (where the weight is given by $p_{N_x^{\text{mir}}}(x')$). Now, our bound suggests that when this train MNF is minimized, this carries over to test MNF too (provided the complexity term is not large). That is, we can say that the fit of $g_{x'}$ generalizes well to unseen, nearby target points x that lie outside of S.

The indirectness of our result. Existing finite sample-based explainers do not explicitly minimize the train MNF term (e.g., MAPLE minimizes an error based upon NF). However, as argued above, they have implicit control over train MNF. Hence, our bound essentially treats MNF as a surrogate for reasoning about the generalization of the explanations learned by an arbitrary procedure. As such, our bound does *not* comment on how well the exact kind of fidelity metric used during training generalizes to test data. Nevertheless, we hope that this result offers a concrete first step towards quantifying the generalization of explanations. Furthermore, we also note that one could also imagine a novel explanation-learning procedure that does explicitly minimize the train MNF term to learn g; in such a case our bound would provide a direct answer to how well its explanations generalize. Indeed, we derive such a theoretically-principled algorithm in Appendix A.

Takeaway. While the above bound captures a similar trade-off with neighborhood width as the Theorem 1, it is worth pausing to appreciate the distinct manner in which this tradeoff arises here. In particular, when the width is too small, we know that the complexity term approaches $O(\sqrt{m})$ and generalization is poor. Intuitively, this is because in this case, the procedure for learning $g_{x'}$ would have been trained to fit very few datapoints from S that would have fallen in the small neigbhorhood of x'. On the other hand, when the neighborhoods are large, this issue would not persist which is captured by the fact that ρ_S approaches $O(1)$. However, with large neighborhoods, it may also be hard to find functions in $\mathcal{G}_{\text{local}}$ that fit so many points in S. Overall, one practical takeaway from this bound is that it is important to not excessively shrink the neighborhood widths if one wants explanations that generalize well for predicting how f behaves at unseen points (see Section 6).

Caveats. We remark that this particular bound applies only when the dataset S is used to learn only g i.e., f and the neighborhoods must be learned beforehand with separate data. This sort of a framework is typical when deriving theoretical results for models like random forests, where it greatly aids analysis to assume that the trees' splits and their decisions are learned from independent datasets (i.e. two halves of an original dataset) (Arlot & Genuer, 2014). Now, if one is interested in a bound where S is also used to simultaneously learn f, the only change to the bound is an added factor corresponding to the complexity of $\mathcal{F}$. Another caveat is that our bound only tells us how well the explanations $g_{x'}$ generalize *on average* over different values of x'. This does not tell us anything about the quality of the generalization of $g_{x'}$ for an arbitrary value of x'. That being said, just as average accuracy remains a central metric for performance (despite ignoring discrepancies across inputs), average MNF can still be a useful quantity for evaluating an explainer's overall performance.

6 EMPIRICAL RESULTS

We present two sets of empirical results to illustrate the the usefulness of our bounds. First, we demonstrate that ρ_S grows much smaller than $\mathcal{O}(\sqrt{m})$ which, as stated before, establishes that our bounds yield meaningful convergence rates. Second, we show that Theorem 2 accurately reflects the relationship between explanation generalization (Theorem 2) and the width of N_x^{mir} used to both generate and evaluate explanations.

Setup. For both experiments, we use several regression datasets from the UCI collection (Dua & Graff, 2017) and standardize each feature to have mean 0 and variance 1. We train neural networks as our "black-box" models with the same setup as in (Plumb et al., 2020), using both their non-regularized and ExpO training procedures. The latter explicitly regularizes for NF during training, which we find also decreases MNF on all datasets. For generating explanations, we define $\mathcal{G}_{\text{local}}$ to be linear models and optimize each g_x using the empirical MNF minimizer (see Appendix A). Finally, we approximate ρ_S using a provably accurate sampling-based approach (see Appendix E).

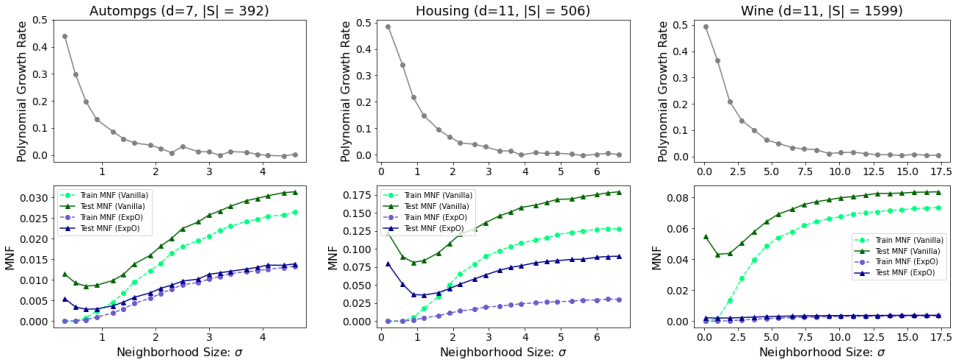

Figure 1: Approximate exponent of ρ_S's polynomial growth rate (top) and train/test MNF (below) plotted for various neighborhood widths across several UCI datasets (see Appendix E for more).

Growth-rate of ρ_S. In Figure 1 (top), we track the sample dependence of ρ_S for various neighborhoods of width σ (setting $N_x^{\text{mir}} = \mathcal{N}(x, \sigma I)$). We specifically approximate the growth rate as polynomial, estimating the exponent by taking the overall slope of a log-log plot of ρ_S over m. To cover a natural range for each dataset, σ is varied to be between the smallest and half the largest inter-example l_2 distances. In these plots, while small σ result in a large exponent for ρ_S and large σ cause g to intuitively saturate towards a global linear model, we observe that there do exist values of σ, where both these terms are in control i.e., we observe that we can achieve a growth rate of approximately $\mathcal{O}(m^{0.2})$ without causing g to saturate and MNF metrics to rise sharply.

Generalization and neighborhood size. As per the setting of Theorem 2, we generate all explanations using data not used to learn the black-box model. Specifically, we split the original test data into two halves, using only the first half for *explanation training* and the second for *explanation testing*. We plot MNF as measured over these two subsets of examples in Figure 1 (bottom). From the results, it is evident that a generalization gap between train and test MNF exists. Further, recall that Theorem 2 predicts that this gap decreases as wider neighborhoods are used, a phenomena reflected in most of these plots. As a result, while training MNF monotonically increases with larger neighborhoods, test MNF always decreases at certain ranges of σ.

7 CONCLUSION AND FUTURE WORK

In this work, we have studied two novel connections between learning theory and local explanations. We believe these results may be of use in guiding the following directions of future work: (1) developing new local explanation algorithms inspired by our theory and the metric of MNF; (2) resolving caveats or otherwise strengthening the theory presented in this paper; and (3) exploring applications of our techniques beyond interpretability, such as the general problem of deep learning generalization or others that require reasoning about the complexity of randomized functions.

REFERENCES

Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias, 2014.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2016*. 2017.

Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21, 03 1993. doi: 10.1214/aos/1176349022.

Jianqing Fan and Irène Gijbels. *Local polynomial modelling and its applications*. Number 66 in Monographs on statistics and applied probability series. Chapman & Hall, London [u.a.]. ISBN 0412983214.

John Langford and Rich Caruana. (Not) bounding the true error. In T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems 14*, pp. 809–816. MIT Press, 2002.

John Langford and John Shawe-Taylor. Pac-bayes & margins. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems 15*, pp. 439–446. MIT Press, 2003.

David McAllester. Simplified pac-bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth (eds.), *Learning Theory and Kernel Machines*, pp. 203–215, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

David A McAllester. Some pac-bayesian theorems. In *11th annual conference on Computational learning theory*, 1998.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems 32*, pp. 11615–11626. Curran Associates, Inc., 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning, 2014. URL https://arxiv.org/abs/1412.6614.

Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2515–2524. Curran Associates, Inc., 2018.

Gregory Plumb, Maruan Al-Shedivat, Angel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. Regularizing black-box models for improved interpretability, 2020. URL https://arxiv.org/abs/1902.06787.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, pages 1135–1144. ACM, 2016*. 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*. 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning, 2019. URL https://arxiv.org/abs/1908.01755.

Jinsung Yoon, Sercan O. Arik, and Tomas Pfister. RL-LIM: Reinforcement learning-based locally interpretable modeling, 2019. URL https://arxiv.org/abs/1909.12367.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.