An Analysis of LIME for Text Data

Dina Mardaoui, Damien Garreau

An Analysis of LIME for Text Data

Dina Mardaoui[1]
dina.mardaoui@etu.univ-cotedazur.fr

Damien Garreau[2]
damien.garreau@unice.fr

[1]Polytech Nice
[2]Université Côte d'Azur, Inria, CNRS, LJAD, France

Abstract

Text data are increasingly handled in an automated fashion by machine learning algorithms. But the models handling these data are not always well-understood due to their complexity and are more and more often referred to as "black-boxes." Interpretability methods aim to explain how these models operate. Among them, LIME has become one of the most popular in recent years. However, it comes without theoretical guarantees: even for simple models, we are not sure that LIME behaves accurately. In this paper, we provide a first theoretical analysis of LIME for text data. As a consequence of our theoretical findings, we show that LIME indeed provides meaningful explanations for simple models, namely decision trees and linear models.

1 Introduction

Natural language processing has progressed at an accelerated pace in the last decade. This time period saw the second coming of artificial neural networks, embodied by the apparition of recurrent neural networks (RNNs) and more particularly long short-term memory networks (LSTMs). These new architectures, in conjunction with large, publicly available datasets and efficient optimization techniques, have allowed computers to compete with and sometime even beat humans on specific tasks.

More recently, the paradigm has shifted from recurrent neural networks to *transformers networks* (Vaswani et al., 2017). Instead of training models specifically for a task, large *language models* are trained on supersized datasets. For instance, `Webtext2` contains the text data associated to 45 millions links (Radford et al., 2019). The growth in complexity of these models seems to know no limit, especially with regards

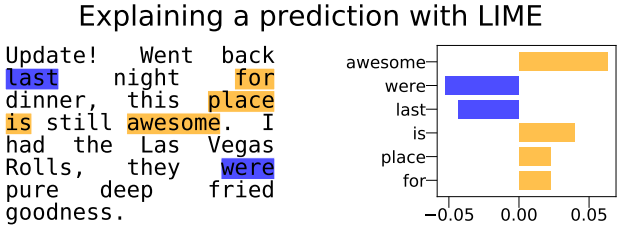

Figure 1: Explaining the prediction of a random forest classifier on a Yelp review. *Left panel:* the document to explain. The words deemed important for the prediction are highlighted, in orange (positive influence) and blue (negative influence). *Right panel:* values of the largest 6 interpretable coefficients, ranked by absolute value.

to their number of parameters. For instance, BERT (Devlin et al., 2018) has roughly 340 millions of parameters, a meager number compared to more recent models such as GTP-2 (Radford et al., 2019, 1.5 billions) and GPT-3 (Brown et al., 2020, 175 billions).

Faced with such giants, it is becoming more and more challenging to understand how particular predictions are made. Yet, *interpretability* of these algorithms is an urgent need. This is especially true in some applications such as healthcare, where natural language processing is used for instance to obtain summaries of patients records (Spyns, 1996). In such cases, we do not want to deploy in the wild an algorithm making near perfect predictions on the test set but for the wrong reasons: the consequences could be tragic.

In this context, a flourishing literature proposing *interpretability methods* emerged. We refer to the survey papers of Guidotti et al. (2018) and Adadi and Berrada (2018) for an overview, and to Danilevsky et al. (2020) for a focus on natural language processing. With the notable exception of SHAP (Lundberg and Lee, 2017), these methods do not come with any guarantees. Namely, given a simple model already interpretable to some extent, we cannot be sure that these methods provide meaningful explanations. For

instance, explaining a model that is based on the presence of a given word should return an explanation that gives high weight to this word. Without such guarantees, using these methods on the tremendously more complex models aforementioned seems like a risky bet.

In this paper, we focus on one of the most popular interpretability method: *Local Interpretable Model-agnostic Explanations* Ribeiro et al. (2016, LIME), and more precisely its implementation for text data. LIME's process to explain the prediction of a model f for an example ξ can be summarized as follows:

(i). from a corpus of documents $\mathcal{C}$, create a TF-IDF transformer ϕ embedding documents into $\mathbb{R}^D$;

(ii). create n perturbed documents $x_1, \dots, x_n$ by deleting words at random in ξ;

(iii). for each new example, get the prediction of the model $y_i := f(\phi(x_i))$;

(iv). train a (weighted) linear surrogate model with inputs the absence / presence of words and responses the y_is.

The user is then given the coefficients of the surrogate model (or rather a subset of the coefficients, corresponding to the largest ones) as depicted in Figure 1. We call these coefficients the *interpretable coefficients*.

The model-agnostic approach of LIME has contributed greatly to its popularity: one does not need to know the precise architecture of f in order to get explanations, it is sufficient to be able to query f a large number of times. The explanations provided by the user are also very intuitive, making it easy to check that a model is behaving in the appropriate way (or not!) on a particular example.

Contributions. In this paper, we present the first theoretical analysis of LIME for text data. In detail,

- we show that, when the number of perturbed samples is large, **the interpretable coefficients concentrate with high probability around a fixed vector** β that depends only on the model, the example to explain, and hyperparameters of the method;
- we provide an **explicit expression of** β, from which we gain interesting insights on LIME. In particular, **the explanations provided are linear in** f;
- for simple decision trees, we go further into the computations. We show that **LIME provably provides meaningful explanations**, giving large coefficients to words that are pivotal for the prediction;
- for linear models, we come to the same conclusion by showing that the interpretable coefficient associate to a given word is approximately equal to **the product of the coefficient in the linear model and the TF-IDF transform of the word** in the example.

We want to emphasize that all our results apply to the default implementation of LIME for text data[1] (as of October 12, 2020), with the only caveat that we do not consider any feature selection procedure in our analysis. All our theoretical claims are supported by numerical experiments, the code thereof can be found at `https://github.com/dmardaoui/lime_text_theory`.

Related work. The closest related work to the present paper is Garreau and von Luxburg (2020a), in which the authors provided a theoretical analysis of a variant of LIME in the case of tabular data (that is, unstructured data belonging to $\mathbb{R}^N$) when f is linear. This line of work was later extended by the same authors (Garreau and von Luxburg, 2020b), this time in a setting very close to the default implementation and for other classes of models (in particular partition-based classifiers such as CART trees and kernel regressors built on the Gaussian kernel). While uncovering a number of good properties of LIME, these analyses also exposed some weaknesses of LIME, notably cancellation of interpretable features for some choices of hyperparameters.

The present work is quite similar in spirit, however we are concerned with *text data*. The LIME algorithm operates quite differently in this case. In particular, the input data goes first through a TF-IDF transform (a non-linear transformation) and there is no discretization step since interpretable features are readily available (the words of the document). Therefore both the analysis and our conclusions are quite different, as it will become clear in the rest of the paper.

2 LIME for text data

In this section, we lay out the general operation of LIME for text data and introduce our notation in the process. From now on, we consider a model f and look at its prediction for a fixed example ξ belonging to a corpus $\mathcal{C}$ of size N, which is built on a dictionary $\mathcal{D}$ of size D. We let $\|\cdot\|$ denote the Euclidean norm, and S^{D-1} the unit sphere of $\mathbb{R}^D$.

Before getting started, let us note that LIME is usually used in the *classification* setting: f takes values in $\{0, 1\}$ (say), and $f(\phi(\xi))$ represents the class at-

[1] `https://github.com/marcotcr/lime`

tributed to ξ by f. However, behind the scenes, LIME requires f to be a real-valued function. In the case of classification, this function is the probability of belonging to a certain class according to the model. In other words, the *regression* version of LIME is used, and this is the setting that we consider in this paper. We now detail each step of the algorithm.

2.1 TF-IDF transform

LIME works with a vector representation of the documents. The TF-IDF transform (Luhn, 1957; Jones, 1972) is a popular way to obtain such a representation. The idea underlying the TF-IDF is quite simple: to any document, associate a vector of size D. If we set $w_1, \dots, w_D$ to be our dictionary, the jth component of this vector represents the importance of word w_j. It is given by the product of two terms: the term frequency (TF, how frequent the word is in the document), and the inverse term frequency (IDF, how rare the word is in our corpus). Intuitively, the TF-IDF of a document has a high value for a given word if this word is frequent in the document and, at the same time, not so frequent in the corpus. In this way, common words such as "the" do not receive high weight.

Formally, let us fix $\delta \in \mathcal{C}$. For each word $w_j \in \mathcal{D}$, we set m_j the number of times w_j appears in δ. We also set $v_j := \log \frac{N+1}{N_j+1} + 1$, where N_j is the number of documents in $\mathcal{C}$ containing w_j. When presented with $\mathcal{C}$, we can pre-compute all the v_js and at run time we only need to count the number of occurrences of w_j in δ. We can now define the normalized TF-IDF:

Definition 1 (Normalized TF-IDF). We define the *normalized TF-IDF* of δ as the vector $\phi(\delta) \in \mathbb{R}^D$ defined coordinate-wise by

$$\forall 1 \leq j \leq D, \quad \phi(\delta)_j := \frac{m_j v_j}{\sqrt{\sum_{j=1}^D m_j^2 v_j^2}}. \qquad (1)$$

Note that there are many different ways to define the TF and IDF terms, as well as normalization choices. We restrict ourselves to the version used in the default implementation of LIME, with the understanding that different implementation choices would not change drastically our analysis. For instance, normalizing by the ℓ_1 norm instead of the ℓ_2 norm would lead to slightly different computations in Proposition 4.

Finally, note that this transformation step does not take place for tabular data, since the data already belong to $\mathbb{R}^D$ in this case.

2.2 Sampling

Let us now fix a given document ξ and describe the sampling procedure of LIME. Essentially, the idea is to sample new documents similar to ξ in order to see how f varies in a neighborhood of ξ.

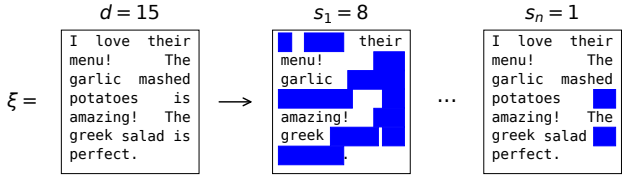

Figure 2: The sampling scheme of LIME for text data. To the left, the document to explain ξ, which contains $d = 15$ distinct words. The new samples $x_1, \dots, x_n$ are obtained by removing s_i random words from ξ (in blue). In the nth sample, one word is removed, yielding two deletions in the original document.

More precisely, let us denote by d the number of distinct words in ξ and set $D_\ell := \{w_1, \dots, w_d\}$ the *local dictionary*. For each new sample, LIME first draws uniformly at random in $\{1, \dots, d\}$ a number s_i of words to remove from ξ. Subsequently, a subset $S_i \subseteq \{1, \dots, d\}$ of size s_i is drawn uniformly at random: all the words with indices contained in S_i are *removed* from ξ. Note that the multiplicity of removals is independent from s_i: if the word "good" appears 10 times in ξ and its index belongs to S, then all the instances of "good" are removed from ξ (see Figure 2). This process is repeated n times, yielding n new samples $x_1, \dots, x_n$. With these new documents come n new binary vectors $z_1, \dots, z_n \in \{0,1\}^d$, marking the absence or presence of a word in x_i. Namely, $z_{i,j} = 1$ if w_j belongs to x_i and 0 otherwise. We call the z_is the *interpretable features*. Note that we will write $\mathbf{1} := (1, \dots, 1)^\top$ for the binary feature associated to ξ: all the words are present.

Already we see a difficulty appearing in our analysis: when removing words from ξ at random, $\phi(\xi)$ is modified in a non-trivial manner. In particular, the denominator of Eq. (1) can change drastically if many words are removed.

In the case of tabular data, the interpretable features are obtained in a completely different fashion, by discretizing the dataset.

2.3 Weights

Let us start by defining the *cosinus distance*:

Definition 2 (Cosinus distance). For any $u, v \in \mathbb{R}^D$, we define

$$d_{\cos}(u, v) := 1 - \frac{u \cdot v}{\|u\| \cdot \|v\|}. \qquad (2)$$

Intuitively, the cosinus distance between u and v is small if the *angle* between u and v is small. Each new

sample x_i receives a positive weight π_i, defined by

$$\pi_i := \exp\left(\frac{-d_{\cos}(\mathbf{1}, z_i)^2}{2\nu^2}\right), \tag{3}$$

where ν is a positive *bandwidth parameter*. The intuition behind these weights is that x_i can be far away from ξ if many words are removed (in the most extreme case, $s = d$, all the words from ξ are removed). In that case, z_i has mostly 0 components, and is far away from $\mathbf{1}$.

Note that the cosinus distance in Eq. (3) is actually multiplied by 100 in the current implementation of LIME. Thus there is the following correspondence between our notation and the code convention: $\nu_{\text{LIME}} = 100\nu$. For instance, the default choice of bandwidth, $\nu_{\text{LIME}} = 25$, corresponds to $\nu = 0.25$.

We now make the following important remark: **the weights only depends on the number of deletions.** Indeed, conditionally to S_i having exactly s elements, we have $z_i \cdot \mathbf{1} = d - s$ and $\|z_i\| = \sqrt{d-s}$. Since $\|\mathbf{1}\| = \sqrt{d}$, using Eq. (3), we deduce that $\pi_i = \psi(s/d)$, where we defined the mapping

$$\begin{aligned} \psi\colon [0,1] &\longrightarrow \mathbb{R} \\ t &\longmapsto \exp\left(\frac{-(1-\sqrt{1-t})^2}{2\nu^2}\right). \end{aligned} \tag{4}$$

We can see in Figure 3 how the weights are given to observations: when s is small, then $\psi(s/d) \approx 1$ and when $s \approx d$, $\psi(s/d)$ which is a small quantity depending on ν. Note that the complicated dependency of the weights in s brings additional difficulty in our analysis, and that we will sometimes restrict ourselves to the large bandwidth regime (that is, $\nu \to +\infty$). In that case, $\pi_i \approx 1$ for any $1 \leq i \leq n$.

Euclidean distance between the interpretable features is used instead of the cosine distance in the tabular data version of the algorithm.

2.4 Surrogate model

The next step is to train a surrogate model on the interpretable features $z_1, \ldots, z_n$, trying to approximate the responses $y_i := f(\phi(x_i))$. In the default implementation of LIME, this model is linear and is obtained by weighted ridge regression (Hoerl and Kennard, 1970). Formally, LIME outputs

$$\hat{\beta}_n^\lambda \in \underset{\beta \in \mathbb{R}^{d+1}}{\arg\min} \left\{ \sum_{i=1}^n \pi_i (y_i - \beta^\top z_i)^2 + \lambda \|\beta\|^2 \right\}, \tag{5}$$

where $\lambda > 0$ is a regularization parameter. We call the components of $\hat{\beta}_n^\lambda$ the *interpretable coefficients*, the 0th coordinate in our notation is by convention the intercept. Note that some feature selection mechanism is often used in practice, limiting the number of interpretable features in output from LIME. We do not consider such mechanism in our analysis.

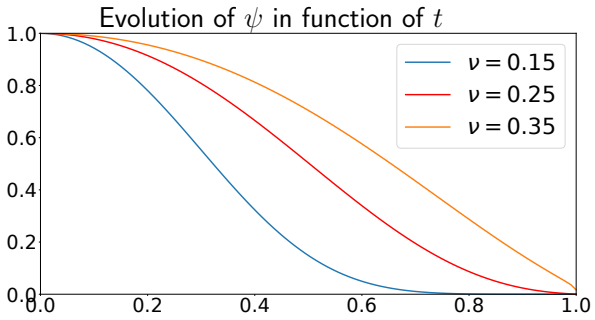

Figure 3: The mapping ψ as a function of t for different bandwidth parameters ($\nu = 0.25$ is default). LIME gives more weights to documents with few deletions ($s/d \approx 0$ means that $\psi(s/d) \approx 1$ regardless of the bandwidth).

We now make a fundamental observation. In its default implementation, LIME uses the default setting of `sklearn` for the regularization parameter, that is, $\lambda = 1$. Hence the first term in Eq. (5) is roughly of order n and the second term of order d. Since we experiment in the large n regime ($n = 5000$ is default) and with documents that have a few dozen distinct words, $n \gg d$. To put it plainly, we can consider that $\lambda = 0$ in our analysis and still recover meaningful results. We will denote by $\hat{\beta}_n$ the solution of Eq. (5) with $\lambda = 0$, that is, ordinary least-squares.

We conclude this presentation of LIME by noting that the main free parameter of the method is the bandwidth ν. As far as we know, there is no principled way of choosing ν. The default choice, $\nu = 0.25$, does not seem satisfactory in many respects. In particular, other choices of bandwidth can lead to different values for interpretable coefficients. In the most extreme cases, they can even change sign, see Figure 4. This phenomenon was also noted for tabular data in Garreau and von Luxburg (2020b).

3 Main result

Without further ado, let us present our main result. For clarity's sake, we split it in two parts: Section 3.1 contains the concentration of $\hat{\beta}_n$ around β^f whereas Section 3.2 presents the exact expression of β^f.

3.1 Concentration of $\hat{\beta}_n$

When the number of new samples n is large, we expect LIME to stabilize and the explanations not to vary too much. The next result supports this intuition.

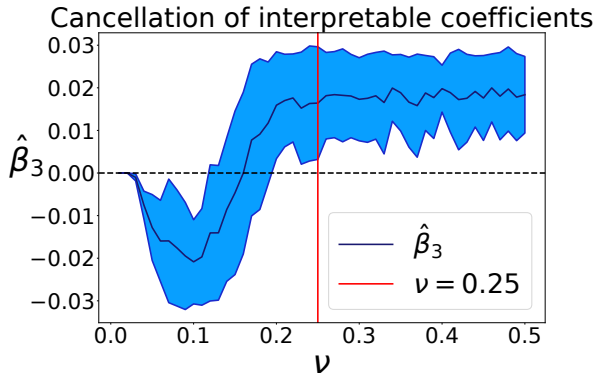

Figure 4: In this experiment, we plot the interpretable coefficient associated to the word "came" as a function of the bandwidth parameter. The red vertical line marks the default bandwidth choice ($\nu = 25$). We can see that LIME gives a negative influence for $\nu \approx 0.1$ and a positive one for $\nu > 0.2$.

Theorem 1 (Concentration of $\hat{\beta}_n$). *Suppose that the model f is bounded by a positive constant M on S^{D-1}. Recall that we let d denote the number of distinct words of ξ, the example to explain. Let $0 < \epsilon < M$ and $\eta \in (0,1)$. Then, there exist a vector $\beta^f \in \mathbb{R}^d$ such that, for every*

$$n \gtrsim \max\left\{M^2 d^9 \mathrm{e}^{\frac{10}{\nu^2}}, M d^5 \mathrm{e}^{\frac{5}{\nu^2}}\right\} \frac{\log \frac{8d}{\eta}}{\epsilon^2},$$

we have $\mathbb{P}\left(\|\hat{\beta}_n - \beta^f\| \geq \epsilon\right) \leq \eta$.

We refer to the supplementary material for a complete statement (we omitted numerical constants here for clarity) and a detailed proof. In essence, Theorem 1 tells us that we can focus on β^f in order to understand how LIME operates, provided that n is large enough. The main limitation of Theorem 1 is the dependency of n in d and ν. The control that we achieve on $\|\hat{\beta}_n - \beta\|$ becomes quite poor for large d or small ν: we would then need n to be unreasonably large in order to witness concentration.

We notice that Theorem 1 is very similar in its form to Theorem 1 in Garreau and von Luxburg (2020b) except that (i) the dimension is replaced by the number of distinct words in the document to explain, and (ii) there is no discretization parameter in our case. The differences with the analysis in the tabular data framework will be more visible in the next section.

3.2 Expression of β^f

Our next result shows that we can derive an explicit expression for β^f. Before stating our result, we need to introduce more notation. From now on, we set x a random variable such that $x_1, \ldots, x_n$ are i.i.d. copies of x. Similarly, π corresponds to the draw of the π_is and z to that of the z_is.

Definition 3 (α coefficients). Define $\alpha_0 := \mathbb{E}[\pi]$ and, for any $1 \leq p \leq d$,

$$\alpha_p := \mathbb{E}[\pi \cdot z_1 \cdots z_p]\,. \tag{6}$$

Intuitively, when ν is large, α_p corresponds to the probability that p distinct words are present in x. The sampling process of LIME is such that α_p does not depend on the exact set of indices considered. In fact, α_p only depends on d and ν. We show in the supplementary material that it is possible to compute the α coefficients in closed-form as a function of d and ν:

Proposition 1 (Computation of the α coefficients). *Let $0 \leq p \leq d$. For any $d \geq 1$ and $\nu > 0$, it holds that*

$$\alpha_p = \frac{1}{d}\sum_{s=1}^{d}\prod_{k=0}^{p-1}\frac{d-s-k}{d-k}\psi\left(\frac{s}{d}\right).$$

From these coefficients, we form the normalization constant

$$c_d := (d-1)\alpha_0\alpha_2 - d\alpha_1^2 + \alpha_0\alpha_1\,. \tag{7}$$

We will also need the following.

Definition 4 (σ coefficients). For any $d \geq 1$ and $\nu > 0$, define

$$\begin{cases} \sigma_1 &:= -\alpha_1\,, \\ \sigma_2 &:= \frac{(d-2)\alpha_0\alpha_2 - (d-1)\alpha_1^2 + \alpha_0\alpha_1}{\alpha_1 - \alpha_2}\,, \\ \sigma_3 &:= \frac{\alpha_1^2 - \alpha_0\alpha_2}{\alpha_1 - \alpha_2}\,. \end{cases} \tag{8}$$

With these notation in hand, we have:

Proposition 2 (Expression of β^f). *Under the assumptions of Theorem 1, we have $c_d > 0$ and, for any $1 \leq j \leq d$,*

$$\beta_j^f = c_d^{-1}\Big\{\sigma_1\mathbb{E}[\pi f(\phi(x))] + \sigma_2\mathbb{E}[\pi z_j f(\phi(x))] + \sigma_3\sum_{\substack{k=1\\k\neq j}}^{d}\mathbb{E}[\pi z_k f(\phi(x))]\Big\}. \tag{9}$$

We also have an expression for the intercept which can be found in the supplementary material, as well as the proof of Proposition 2. At first glance, Eq. (9) is quite similar to Eq. (6) in Garreau and von Luxburg (2020b), which gives the expression of β_j^f in the tabular data case. The main difference is the TF-IDF transform in the expectation, personified by ϕ, and the additional terms (there is no σ_3 factor in the tabular data

case). In addition, the expression of the σ coefficients is much more complicated than in the tabular data case. We now present some immediate consequences of Proposition 2.

Linearity of explanations. Perhaps the most striking feature of Eq. (9) is that it is **linear in** f. More precisely, the mapping $f \mapsto \beta^f$ is linear in f: for any given two functions f and g, we have

$$\beta^{f+g} = \beta^f + \beta^g \,.$$

Therefore, because of Theorem 1, the explanations $\hat{\beta}_n$ obtained for a finite sample of new examples are also approximately linear in the model to explain. We illustrate this phenomenon in Figure 5. This is remarkable: many models used in machine learning can be written as a linear combination of smaller models (*e.g.*, generalized linear models, kernel regressors, decision trees and random forests). In order to understand the explanations provided by these complicated models, one can try and understand the explanations for the elementary elements of the models first.

Large bandwidth. It can be difficult to get a good sense of the values taken by the σ coefficients, and therefore of β. Let us see how Proposition 2 simplifies in the large bandwidth regime and what insights we can gain. We denote by β_∞ the limit of β when $\nu \to +\infty$. When $\nu \to +\infty$, we prove in the supplementary material that, for any $1 \leq j \leq d$, up to $\mathcal{O}(1/d)$ terms and a numerical constant, the j-th coordinate of β_∞ is then approximately equal to

$$\left(\beta^f_\infty\right)_j \approx \mathbb{E}\left[f(\phi(x))|w_j \in x\right] - \frac{1}{d}\sum_{k\neq j}\mathbb{E}\left[f(\phi(x))|w_k \in x\right].$$

Intuitively, the interpretable coefficient associated to the word w_j is high if **the expected value of the model when word w_j is present is significantly higher than the typical expected value when other words are present**. We think that this is reasonable: if the model predicts much higher values when w_j belongs to the example, it surely means that w_j being present is important for the prediction.

3.3 Sketch of the proof

We conclude this section with a brief sketch of the proof of Theorem 1, the full proof can be found in the supplementary material.

Since we set $\lambda = 0$ in Eq. (5), $\hat{\beta}_n$ is the solution of a weighted least-squares problem. Denote by $W \in \mathbb{R}^{n\times n}$ the diagonal matrix such that $W_{i,i} = \pi_i$, and set $Z \in \{0,1\}^{n\times(d+1)}$ the matrix such that its ith line is $(1, z_i^\top)$. Then the solution of Eq. (5) is given by

$$\hat{\beta}_n = \left(Z^\top W Z\right)^{-1} Z^\top W y \,,$$

where we defined $y \in \mathbb{R}^n$ such that $y_i = f(\phi(x_i))$ for all $1 \leq i \leq n$. Let us set $\hat{\Sigma}_n := \frac{1}{n} Z^\top W Z$ and $\hat{\Gamma}^f_n := \frac{1}{n} Z^\top W y$. By the law of large numbers, we know that both $\hat{\Sigma}_n$ and $\hat{\Gamma}^f_n$ converge in probability towards their population counterparts $\Sigma := \mathbb{E}[\hat{\Sigma}_n]$ and $\Gamma^f := \mathbb{E}[\hat{\Gamma}_n]$. Therefore, provided that Σ is invertible, $\hat{\beta}_n$ is close to $\beta^f := \Sigma^{-1}\Gamma^f$ with high probability.

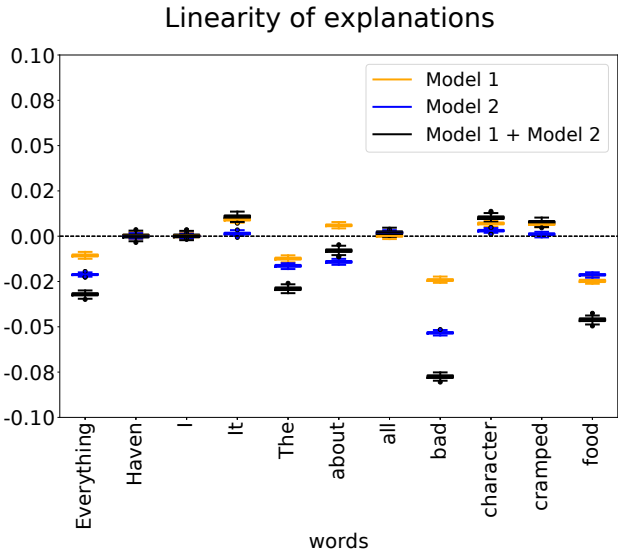

Figure 5: The explanations given by LIME for the sum of two models (here two random forests regressors) are the sum of the explanations for each model, up to noise coming from the sampling procedure.

As we have seen in Section 2, the main differences with respect to the tabular data implementation are (i) the interpretable features, and (ii) the TF-IDF transform. The first point lead to a completely different Σ than the one obtained in Garreau and von Luxburg (2020b). In particular, it has no zero coefficients, leading to more complicated expression for β^f and additional challenges when controlling $\left\|\Sigma^{-1}\right\|_{\text{op}}$. The second point is quite challenging since, as noted in Section 2.1, **the TF-IDF transform of a document changes radically when deleting words at random in the document.** This is the main reason why we have to resort to approximations when dealing with linear models.

4 Expression of β^f for simple models

In this section, we see how to specialize Proposition 2 to simple models f. Recall that our main goal in doing so is to investigate whether it makes sense or not to use LIME in these cases. We will focus on two classes of

models: decision trees (Section 4.1) and linear models (Section 4.2).

4.1 Decision trees

In this section we focus on simple decision trees built on the presence or absence of given words. For instance, let us look at the model returning 1 if the word "food" is present, or if "about" and "everything" are present in the document. Ideally, LIME would give high positive weights to "food," "about," and "everything," if they are present in the document to explain, and small weight to all other words.

We first notice that such simple decision trees can be written as sums of products of the binary features. Indeed, recall that we defined $z_j = \mathbf{1}_{w_j \in x}$. For instance, suppose that the first three words of our dictionary are "food," "about," and "everything." Then the model from the previous paragraph can be written

$$g(x) = z_1 + (1 - z_1) \cdot z_2 \cdot z_3 \,. \tag{10}$$

Now it is clear that the z_js can be written as function of the TF-IDF transform of a word, since $w_j \in x$ if, and only if, $\phi(x)_j > 0$. Therefore this class of models falls into our framework and we can use Theorem 1 and Proposition 2 in order to gain insight on the explanations provided by LIME. For instance, Eq. (10) can be written as $f(\phi(x))$ with, for any $\zeta \in \mathbb{R}^D$,

$$f(\zeta) := \mathbf{1}_{\zeta_1 > 0} + (1 - \mathbf{1}_{\zeta_1 > 0}) \cdot \mathbf{1}_{\zeta_2 > 0} \cdot \mathbf{1}_{\zeta_3 > 0} \,.$$

By linearity, it is sufficient to know how to compute β^f when f is a product of indicator functions.

We now make an important remark: since the new example $x_1, \dots, x_n$ are created by deleting words at random from the text ξ, x **only contains words that are already present in** ξ. Therefore, without loss of generality, we can restrict ourselves to the local dictionary (the distinct words of ξ). Indeed, for any word w not already in ξ, $\mathbf{1}_{w \in x} = 0$ almost surely. As before, we denote by D_ℓ the local dictionary associated to ξ, and we denote its elements by $w_1, \dots, w_d$. We can compute in closed-form the interpretable coefficients for a product of indicator functions:

Proposition 3 (Computation of β^f, product of indicator functions). *Let $J \subseteq \{1, \dots, d\}$ be a set of p distinct indices and set $f(x) = \prod_{j \in J} \mathbf{1}_{x_j > 0}$. Then, for any $j \in J$,*

$$\beta_j^f = c_d^{-1}\big[\sigma_1\alpha_p + \sigma_2\alpha_p + (d-p)\sigma_3\alpha_{p+1} + (p-1)\sigma_3\alpha_p\big]$$

and, for any $j \in \{1, \dots, d\} \setminus J$,

$$\beta_j^f = c_d^{-1}\big[\sigma_1\alpha_p + \sigma_2\alpha_{p+1} + (d-p-1)\sigma_3\alpha_{p+1} + p\sigma_3\alpha_p\big].$$

In particular, when $p = 0$, Proposition 3 simplifies greatly and we find that $1 \leq k \leq d$, $\beta_k^f = \mathbf{1}_{k=j}$. It is already a reassuring result: when the model is just indicating if a given word is present, **the explanation given by LIME is one for this word and zero for all the other words**.

It is slightly more complicated to see what happens when $p \geq 1$. To this extent, let us set $j \in J$ and $k \notin J$. Then it follows readily from Proposition 14 that

$$\beta_j^f - \beta_k^f = c_d^{-1}(\sigma_2 + \sigma_3)(\alpha_p - \alpha_{p+1}) \,.$$

Since $\alpha_p \approx 1/(p+1)$ and $\sigma_2 + \sigma_3 \approx 6$, we deduce that $\beta_j^f \gg \beta_k^f$. Moreover, from Definition 3 and 4 one can show that $\beta_k^f = \mathcal{O}(1/d)$ when ν is large. Thus Proposition 14 tells us that **LIME gives large positive coefficients to words that are in the support of f and small coefficients to all the other words**. This is a satisfying property.

Together with the linearity property, Proposition 14 allows us to compute β^f for any decision tree that can be written as in Eq. (10). We give an example of our theoretical predictions in Figure 6. As predicted, **the words that are pivotal in the prediction have high interpretable coefficients, whereas the other words receive near-zero coefficients**. It is interesting to notice that words that are near the root of the tree receive a greater weight. We present additional experiments in the supplementary material.

4.2 Linear models

We now focus on linear models, that is, for any document x,

$$f(\phi(x)) := \sum_{j=1}^{d} \lambda_j \phi(x)_j \,, \tag{11}$$

where $\lambda_1, \dots, \lambda_d$ are arbitrary fixed coefficients. We have to resort to approximate computations in this case: from now on, we assume that $\nu = +\infty$. We start with the simplest linear function: all coefficients are zero except one, that is, $\lambda_k = 1$ if $k = j$ and 0 otherwise in Eq. (11), for a fixed index j. We need to introduce additional notation before stating or result. For any $1 \leq j \leq d$, define

$$\omega_k := \frac{m_j^2 v_j^2}{\sum_{k=1}^d m_k^2 v_k^2} \,,$$

where the m_ks and v_ks were defined in Section 2.1. For any J that is a strict subset of $\{1, \dots, d\}$, define $H_S := \sum_{j \in J} \omega_j$. Recall that S denotes the random subset of indices chosen by LIME in the sampling step (see Section 2.2). Define $E_j = \mathbb{E}\left[(1 - H_S)^{-1/2} \middle| S \not\supset j\right]$

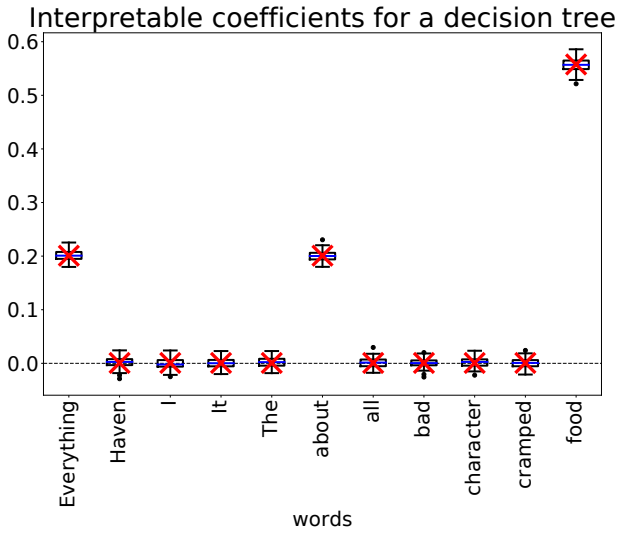

Figure 6: Theory *vs* practice for the tree defined by Eq. (10). The black whisker boxes correspond to 100 runs of LIME with default settings ($n = 5000$ new examples and $\nu = 0.25$) whereas the red crosses correspond to the theoretical predictions given by our analysis. The example to explain is a Yelp review with $d = 35$ distinct words.

and for any $k \neq j$, $E_{j,k} = \mathbb{E}\left[(1 - H_S)^{-1/2} \middle| S \not\ni j, k\right]$. Then we have the following:

Proposition 4 (Computation of β^f, linear case). *Let $1 \leq j \leq d$ and assume that $f(\phi(x)) = \phi(x)_j$. Then, for any $1 \leq k \leq d$ such that $k \neq j$,*

$$\left(\beta^f_\infty\right)_k = \left[2E_{j,1} - \frac{2}{d}\sum_{\ell \neq k,j} E_{j,\ell}\right]\phi(\xi)_j + \mathcal{O}\left(\frac{1}{d}\right),$$

and

$$\left(\beta^f_\infty\right)_j = \left[3E_j - \frac{2}{d}\sum_{k \neq j} E_{j,k}\right]\phi(\xi)_j + \mathcal{O}\left(\frac{1}{d}\right).$$

Proposition 4 is proved in the supplementary material. The main difficulty is to compute the expected value of $\phi(x)_j$: this is the reason for the E_j terms, for which we find an approximate expression as a function of the ω_ks. Assuming that the ω_k are small, we can further this approximation and show that $E_j \approx 1.22$ and $E_{j,k} \approx 1.15$. In particular, **these expressions do not depend on j and k**. Thus we can drastically simplify the statement of Proposition 4: for any $k \neq j$, $\left(\beta^f_\infty\right)_k \approx 0$ and $\left(\beta^f_\infty\right)_j \approx 1.36\phi(\xi)_j$. We can now go back to our original goal, Eq. (11). By linearity, we deduce that

$$\forall 1 \leq j \leq d, \quad \left(\beta^f_\infty\right)_j \approx 1.36 \cdot \lambda_j \cdot \phi(\xi)_j\,. \tag{12}$$

In other words, up to a numerical constant and small error terms depending on d, **the explanation for a linear f is the TF-IDF value of the word multiplied by the coefficient of the linear model.** We believe that this behavior is desirable for an interpretability method: large coefficients in the linear model should intuitively be associated to large interpretable coefficients. But at the same time the TF-IDF of the term is taken into account.

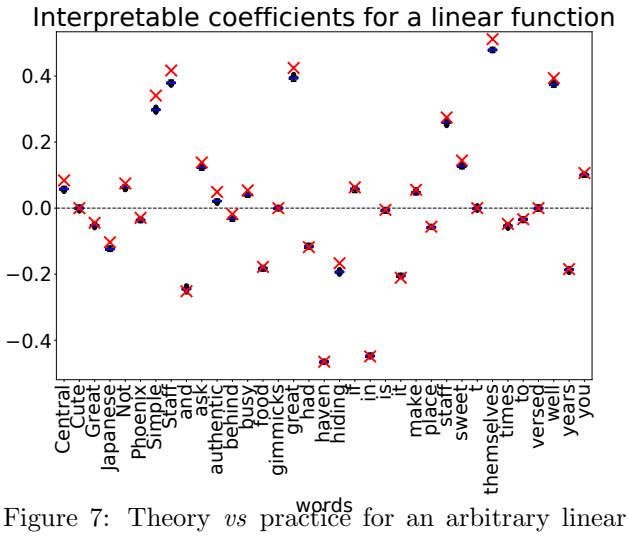

Figure 7: Theory *vs* practice for an arbitrary linear model. The black whisker boxes correspond to 100 runs of LIME with default settings ($n = 5000$ and $\nu = 0.25$). The red crosses correspond to our theoretical predictions: $\beta_j \approx 1.36\lambda_j\phi(\xi)_j$.

We observe a very good match between theory and practice (see Figure 7). Surprisingly, this is the case even though we assume that ν is large in our derivations, whereas ν is chosen by default in all our experiments. We present experiments with other bandwidths in the supplementary.

5 Conclusion

In this work we proposed the first theoretical analysis of LIME for text data. In particular, we provided a closed-form expression for the interpretable coefficients when the number of perturbed samples is large. Leveraging this expression, we exhibited some desirable behavior of LIME such as the linearity with respect to the model. In specific cases (simple decision trees and linear models), we derived more precise expression, showing that LIME outputs meaningful explanations in these cases.

As future work, we want to tackle more complex models. More precisely, we think that it is possible to obtained approximate statements in the spirit of Eq. (12) for models that are not linear. In the long run, we also want to analyze LIME for images, which is a much more challenging task.

References

A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. A survey of the state of Explainable AI for Natural Language Processing. *arXiv preprint arXiv:2010.00711*, 2020.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

S. Filipovski. Improved Cauchy-Schwarz inequality and its applications. *Turkish journal of inequalities*, 3(2):8–12, 2019.

D. Garreau and U. von Luxburg. Explaining the explainer: A first theoretical analysis of LIME. In *Proceedings of the 33rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1287–1296, 2020a.

D. Garreau and U. von Luxburg. Looking Deeper into Tabular LIME. *arXiv preprint arXiv:2008.11092*, 2020b.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

D. M. W. Powers. Applications and explanations of Zipf's law. In *New methods in language processing and computational natural language learning*, 1998.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

M. T. 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*, pages 1135–1144, 2016.

P. Ross. Generalized hockey stick identities and n-dimensional blockwalking. *The College Mathematics Journal*, 28(4):325, 1997.

P. Spyns. Natural language processing in medicine: an overview. *Methods of information in medicine*, 35(4-5):285–301, 1996.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Supplementary material for the paper: "An Analysis of LIME for Text Data"

Organization of the supplementary material

In this supplementary material, we collect the proofs of all our theoretical results and additional experiments. We study the covariance matrix in Section 1 and the responses in Section 2. The proof of our main results can be found in Section 3. Combinatorial results needed for the approximation formulas obtained in the linear case are collected in Section 4, while other technical results can be found in Section 5. Finally, we present some additional experiments in Section 6.

Notation. First, let us quickly recall our notation. We consider x, z, π the generic random variables associated to the sampling of new examples by LIME. To put it plainly, the new examples $x_1, \dots, x_n$ are i.i.d. samples from the random variable x. Also remember that we denote by $S \subseteq \{1, \dots, d\}$ the random subset of indices removed by LIME when creating new samples for a text with d distinct words. For any finite set R, we write $\#R$ the cardinality of R. Recall that we denote by S the random set of indices deleted in the sampling. We write $\mathbb{E}_s$ the expectation conditionally to $\#S = s$. Since we consider vectors belonging to $\mathbb{R}^{d+1}$ with the zero-th coordinate corresponding to an intercept, we will often start the numbering at 0 instead of 1. For any matrix M, we set $\|M\|_{\mathrm{F}}$ the Frobenius norm of M and $\|M\|_{\mathrm{op}}$ the operator norm of M.

1 The study of Σ

We begin by the study of the covariance matrix. We show in Section 1.1 how to compute Σ. We will see how the α coefficients defined in the main paper appear. In Section 1.2, we show that it is possible to invert Σ in closed-form: it can be written in function of c_d and the σ coefficients. We show how $\hat{\Sigma}_n$ concentrates around Σ in Section 1.3. Finally, Section 1.4 is dedicated to the control of $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$.

1.1 Computation of Σ

In this section, we derived a closed-form expression for $\Sigma := \mathbb{E}[\hat{\Sigma}_n]$ as a function of d and ν. Recall that we defined $\hat{\Sigma} = \frac{1}{n} Z^\top W Z$. By definition of Z and W, we have

$$\hat{\Sigma} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \pi_i & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} & \cdots & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} & \cdots & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} z_{i,d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} z_{i,d} & \cdots & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Taking the expectation in the last display with respect to the sampling of new examples yields

$$\Sigma = \begin{pmatrix} \mathbb{E}[\pi] & \mathbb{E}[\pi z_1] & \cdots & \mathbb{E}[\pi z_d] \\ \mathbb{E}[\pi z_1] & \mathbb{E}[\pi z_1] & \cdots & \mathbb{E}[\pi z_1 z_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\pi z_d] & \mathbb{E}[\pi z_1 z_d] & \cdots & \mathbb{E}[\pi z_d] \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{13}$$

An important remark is that $\mathbb{E}[\pi z_j]$ does not depend on j. Indeed, there is no privileged index in the sampling of S (the subset of removed indices). Thus we only have to look into $\mathbb{E}[\pi z_1]$ (say). For the same reason, $\mathbb{E}[\pi z_j z_k]$ does not depend on the 2-uple (j, k), and we can limit our investigations to $\mathbb{E}[\pi z_1 z_2]$. This is the reason why we

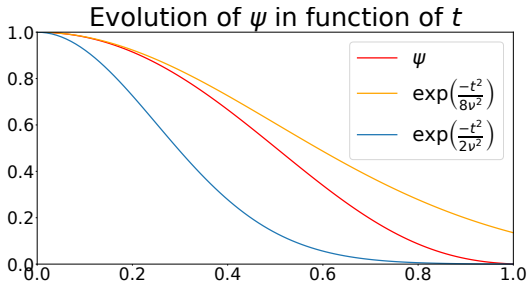

Figure 8: The function ψ defined by Eq. (15) with bandwidth parameter $\nu = 0.25$. In orange (resp. blue), one can see the upper (resp. lower) bound given by Eq. (16).

defined $\alpha_0 = \mathbb{E}[\pi]$ and, for any $1 \leq p \leq d$,

$$\alpha_p = \mathbb{E}[\pi \cdot z_1 \cdots z_p] \tag{14}$$

in the main paper. We recognize the definition of the α_ps in Eq. (13) and we write

$$\Sigma_{j,k} = \begin{cases} \alpha_0 & \text{if } j = k = 0, \\ \alpha_1 & \text{if } j = 0 \text{ and } k > 0 \text{ or } j > 0 \text{ and } k = 0 \text{ or } j = k > 0, \\ \alpha_2 & \text{otherwise.} \end{cases}$$

As promised, we can be more explicit regarding the α coefficients. Recall that we defined the mapping

$$\begin{aligned} \psi \colon [0,1] &\longrightarrow \mathbb{R} \\ t &\longmapsto \exp\left(-(1-\sqrt{1-t})^2/(2\nu^2)\right) . \end{aligned} \tag{15}$$

It is a decreasing mapping (see Figure 8). With this notation in hand, we have the following expression for the α coefficients (this is Proposition 1 in the paper):

Proposition 5 (Computation of the α coefficients). *For any $d \geq 1$, $\nu > 0$, and $p \geq 0$, it holds that*

$$\alpha_p = \frac{1}{d} \sum_{s=1}^{d} \prod_{k=0}^{p-1} \frac{d-s-k}{d-k} \psi\left(\frac{s}{d}\right) .$$

In particular, the first three α coefficients can be written

$$\alpha_0 = \frac{1}{d} \sum_{s=1}^{d} \psi\left(\frac{s}{d}\right) , \quad \alpha_1 = \frac{1}{d} \sum_{s=1}^{d} \left(1 - \frac{s}{d}\right) \psi\left(\frac{s}{d}\right) , \quad \text{and} \quad \alpha_2 = \frac{1}{d} \sum_{s=1}^{d} \left(1 - \frac{s}{d}\right) \left(1 - \frac{s}{d-1}\right) \psi\left(\frac{s}{d}\right) .$$

Proof. The idea of the proof is to use the law of total expectation with respect to the collection of events $\{\#S = s\}$ for $s \in \{1, \ldots, d\}$. Since $\mathbb{P}(\#S = s) = \frac{1}{d}$ for any $1 \leq s \leq d$, all that is left to compute is the expectation of $\pi z_1 \cdots z_p$ conditionally to $\#S = s$. According to the remark in Section 2.3 of the main paper, $\pi = \psi(s/d)$ conditionally to $\{\#S = s\}$. We can conclude since, according to Lemma 4,

$$\mathbb{P}_s(w_1 \in x, \ldots, w_p \in x) = \frac{(d-s)(d-s-1)\cdots(d-s-p+1)}{d(d-1)\cdots(d-p+1)} .$$

□

It is important to notice that, when $\nu \to +\infty$, $\psi(t) \to 0$ for any $t \in (0,1]$. As a consequence, in the large bandwidth regime, the $\psi(s/d)$ weights are arbitrarily close to one. We demonstrate this effect in Figure 9. In this situation, the α coefficients take a simpler form.

Corollary 1 (Large bandwidth approximation of α coefficients). *For any $0 \leq p \leq d$, it holds that*

$$\lim_{\nu \to +\infty} \alpha_p = \frac{d-p}{(p+1)d} .$$

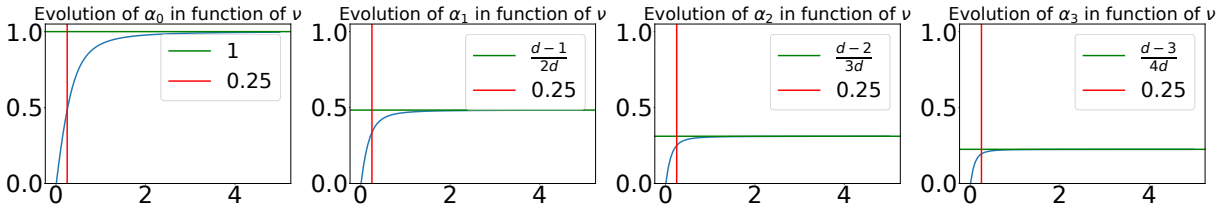

Figure 9: Behavior of the first α coefficients with respect to the bandwidth parameter ν. The red vertical lines mark the default bandwidth choice ($\nu = 0.25$). The green horizontal line denotes the limits for large d given by Corollary 1.

We report these approximate values in Figure 9. In particular, when both ν and d are large, we can see that $\alpha_p \approx 1/(p+1)$. Thus $\alpha_0 \approx 1$, $\alpha_1 \approx \frac{1}{2}$, and $\alpha_2 \approx \frac{1}{3}$.

Proof. When $\nu \to +\infty$, we have $\psi(s/d) \to 1$ and we can conclude directly by using Lemma 5. □

Notice that we can be slightly more precise than Corollary 1. Indeed, ψ is decreasing on $[0,1]$, thus for any $t \in [0,1]$, $\exp\left(-1/(2\nu^2)\right) \leq \psi(t) \leq 1$. Therefore we can present some efficient bounds for the α coefficients when ν is large.

Corollary 2 (Bounds on the α coefficients). *For any $0 \leq p \leq d$, it holds that*

$$\frac{d-p}{(p+1)d} \mathrm{e}^{\frac{-1}{2\nu^2}} \leq \alpha_p \leq \frac{d-p}{(p+1)d} .$$

One can further show that, for any $0 \leq t \leq 1$,

$$\exp\left(\frac{-t^2}{2\nu^2}\right) \leq \psi(t) \leq \exp\left(\frac{-t^2}{8\nu^2}\right) . \tag{16}$$

Using Eq. (16) together with the series-integral comparison theorem would yield very accurate bounds for the α coefficients and related quantities, but we will not follow that road.

1.2 Computation of Σ^{-1}

In this section, we present a closed-form formula for the matrix inverse of Σ as a function of d and ν.

Proposition 6 (Computation of Σ^{-1}). *For any $d \geq 1$ and $\nu > 0$, recall that we defined*

$$c_d = (d-1)\alpha_0\alpha_2 - d\alpha_1^2 + \alpha_0\alpha_1 .$$

Assume that $c_d \neq 0$ and $\alpha_1 \neq \alpha_2$. Define $\sigma_0 := (d-1)\alpha_2 + \alpha_1$ and recall that we set

$$\begin{cases} \sigma_1 &= -\alpha_1 , \\ \sigma_2 &= \frac{(d-2)\alpha_0\alpha_2 - (d-1)\alpha_1^2 + \alpha_0\alpha_1}{\alpha_1 - \alpha_2} , \\ \sigma_3 &= \frac{\alpha_1^2 - \alpha_0\alpha_2}{\alpha_1 - \alpha_2} . \end{cases}$$

Then it holds that

$$\Sigma^{-1} = \frac{1}{c_d} \begin{pmatrix} \sigma_0 & \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_3 \\ \sigma_1 & \sigma_3 & \sigma_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \sigma_3 \\ \sigma_1 & \sigma_3 & \cdots & \sigma_3 & \sigma_2 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)} . \tag{17}$$

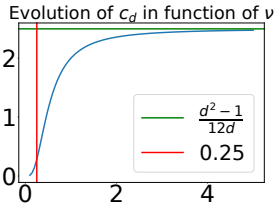

Figure 10: Evolution of the normalization constant c_d as a function of the bandwidth for $d = 30$. In red, the default bandwidth $\nu = 0.25$, in green the limit for large bandwidth given by Corollary 3.

We display the evolution of the σ_i/c_d coefficients with respect to ν in Figure 11.

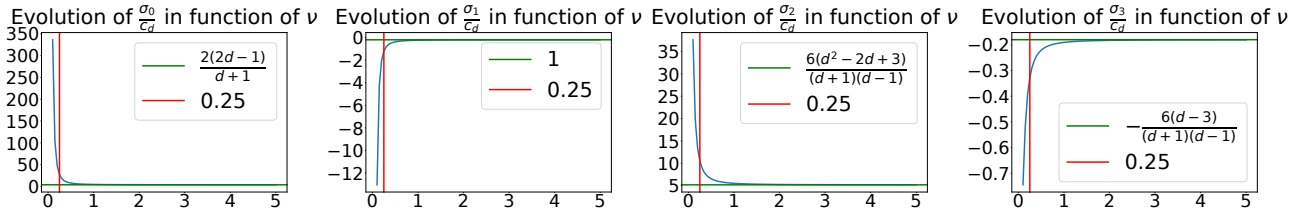

Figure 11: Evolution of σ_i/c_d as a function of ν for $1 \leq i \leq 4$ for $d = 30$. In red the default value of the bandwidth. In green the limits given by Corollary 3. We can see that the σ coefficients are close to these limit values for the default bandwidth.

Proof. From Eq. (13), we can see that Σ is a block matrix. The result follows from the block matrix inversion formula and one can check directly that $\Sigma \cdot \Sigma^{-1} = I_{d+1}$. □

Our next result shows that the assumptions of Proposition 6 are satisfied: $\alpha_1 - \alpha_2$ and c_d are positive quantities. In fact, we prove a slightly stronger statement which will be necessary to control the operator norm of Σ^{-1}.

Proposition 7 (Σ is invertible). *For any $d \geq 2$,*

$$\alpha_1 - \alpha_2 \geq \frac{\mathrm{e}^{\frac{-1}{2\nu^2}}}{6} > 0\,, \qquad \textit{and} \qquad c_d \geq \frac{\mathrm{e}^{\frac{-2}{\nu^2}}}{40} > 0\,.$$

Proof. By definition of the α coefficients (Eq. (14)), we have

$$\alpha_1 - \alpha_2 = \frac{1}{d}\sum_{s=1}^{d}\left(1 - \frac{s}{d}\right)\frac{s}{d-1}\psi\left(\frac{s}{d}\right)\,.$$

Since $\mathrm{e}^{\frac{-1}{2\nu^2}} \leq \psi(t) \leq 1$ for any $t \in [0,1]$, we have

$$\mathrm{e}^{\frac{-1}{2\nu^2}} \cdot \frac{1}{d}\sum_{s=1}^{d}\left(1 - \frac{s}{d}\right)\frac{s}{d-1} = \frac{d+1}{6d} \cdot \mathrm{e}^{\frac{-1}{2\nu^2}} \leq \alpha_1 - \alpha_2 \leq \frac{d+1}{6d}\,. \tag{18}$$

The right-hand side of Eq. (18) yields the promised bound. Note that the same reasoning gives

$$\frac{d+1}{2d} \cdot \mathrm{e}^{\frac{-1}{2\nu^2}} \leq \alpha_0 - \alpha_1 \leq \frac{d+1}{2d}\,. \tag{19}$$

Let us now find a lower bound for c_d. We first start by noticing that

$$\begin{aligned} c_d &= d\alpha_1(\alpha_0 - \alpha_1) - (d-1)\alpha_0(\alpha_1 - \alpha_2) \qquad (20)\\ &= \sum_{s=1}^{d}\left(1 - \frac{s}{d}\right)\psi\left(\frac{s}{d}\right) \cdot \frac{1}{d}\sum_{s=1}^{d}\frac{s}{d}\psi\left(\frac{s}{d}\right) - \sum_{s=1}^{d}\psi\left(\frac{s}{d}\right) \cdot \frac{1}{d}\sum_{s=1}^{d}\left(1 - \frac{s}{d}\right)\psi\left(\frac{s}{d}\right)\\ c_d &= \frac{1}{d}\left[\sum_{s=1}^{d}\psi\left(\frac{s}{d}\right) \cdot \sum_{s=1}^{d}\frac{s^2}{d^2}\psi\left(\frac{s}{d}\right) - \left(\sum_{s=1}^{d}\frac{s}{d}\psi\left(\frac{s}{d}\right)\right)^2\right]\,. \end{aligned}$$

Therefore, by Cauchy-Schwarz inequality, $c_d \geq 0$. In fact, $c_d > 0$ since the equality case in Cauchy-Schwarz is attained for proportional summands, which is not the case here.

However, we need to improve this result if we want to control $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$ more precisely. To this extent, we use a refinement of Cauchy-Schwarz inequality obtained by Filipovski (2019). Let us set, for any $1 \leq s \leq d$,

$$a_s := \sqrt{\psi\left(\frac{s}{d}\right)}\,, \quad b_s := \frac{s}{d}\sqrt{\psi\left(\frac{s}{d}\right)}\,, \quad A := \sqrt{\sum_{s=1}^{d} a_s^2}\,, \quad \text{and } B := \sqrt{\sum_{s=1}^{d} b_s^2}\,.$$

With these notation,

$$c_d = \frac{1}{d}\left[A^2B^2 - \left(\sum_{s=1}^{d} a_s b_s\right)^2\right],$$

and Cauchy-Schwarz yields $A^2B^2 \geq \left(\sum_{s=1}^{d} a_s b_s\right)^2$. Theorem 2.1 in Filipovski (2019) is a stronger result, namely

$$AB \geq \sum_{s=1}^{d} a_s b_s + \frac{1}{4}\sum_{s=1}^{d} \frac{(a_s^2B^2 - b_s^2A^2)^2}{a_s^4B^4 + b_s^4A^4} a_s b_s \,. \tag{21}$$

Let us focus on this last term. Since all the terms are non-negative, we can lower bound by the term of order d, that is,

$$\frac{1}{4}\sum_{s=1}^{d} \frac{(a_s^2B^2 - b_s^2A^2)^2}{a_s^4B^4 + b_s^4A^4} a_s b_s \geq \frac{1}{4}\frac{(b_d^2A^2 - a_d^2B^2)^2}{b_d^4A^4 + a_d^4B^4} a_d b_d = \frac{1}{4}\frac{(A^2 - B^2)^2}{A^4 + B^4}\psi(1)\,, \tag{22}$$

since $a_d = b_d = \sqrt{\psi(1)}$. On one side, we notice that

$$\begin{aligned}
A^2 - B^2 &= \sum_{s=1}^{d}\left(1 - \frac{s^2}{d^2}\right)\psi\left(\frac{s}{d}\right) \\
&\geq \exp\left(\frac{-1}{2\nu^2}\right)\cdot\sum_{s=1}^{d}\left(1 - \frac{s^2}{d^2}\right) && \text{(for any } t\in[0,1],\ \psi(t) \geq \mathrm{e}^{-1/(2\nu^2)}\text{)} \\
&= \exp\left(\frac{-1}{2\nu^2}\right)\cdot\frac{1}{6}\left(4d - \frac{1}{d} - 3\right) \\
A^2 - B^2 &\geq \frac{3d\cdot\exp\left(\frac{-1}{2\nu^2}\right)}{8}\,,
\end{aligned}$$

where we used $d \geq 2$ in the last display. We deduce that $(A^2 - B^2)^2 \geq 9d^2\mathrm{e}^{\frac{-1}{2\nu^2}}/64$. On the other side, it is clear that $A^2 \leq d$, and

$$B^2 \leq \sum_{s=1}^{d}\frac{s^2}{d^2} = \frac{(d+1)(2d+1)}{6d}\,.$$

For any $d \geq 2$, we have $B^2 \leq 5d/8$, and we deduce that $A^4 + B^4 \leq \frac{89}{64}d^2$. Therefore,

$$\frac{(A^2 - B^2)^2}{A^4 + B^4} \geq \frac{9\mathrm{e}^{\frac{-1}{\nu^2}}}{89}\,.$$

Coming back to Eq. (22), we proved that

$$\frac{1}{4}\sum_{s=1}^{d} \frac{(a_s^2B^2 - b_s^2A^2)^2}{a_s^4B^4 + b_s^4A^4} a_s b_s \geq \frac{9\mathrm{e}^{\frac{-3}{2\nu^2}}}{356}\,.$$

Plugging into Eq. (21) and taking the square, we deduce that

$$A^2B^2 \geq \left(\sum_{s=1}^{d} a_s b_s\right)^2 + 2\cdot\sum_{s=1}^{d} a_s b_s \cdot \frac{9\mathrm{e}^{\frac{-3}{2\nu^2}}}{356} + \frac{81\mathrm{e}^{\frac{-3}{\nu^2}}}{126736}\,.$$

But $\sum a_s b_s \geq d\mathrm{e}^{\frac{-1}{2\nu^2}}/2$, therefore, ignoring the last term, we have

$$A^2B^2 - \left(\sum_{s=1}^{d} a_s b_s\right)^2 \geq \frac{9d\mathrm{e}^{\frac{-2}{\nu^2}}}{356}\,.$$

We conclude by noticing that $356/9 \leq 40$. □

Remark 1. We suspect that the correct lower bound for c_d is actually of order d, but we did not manage to prove it. Careful inspection of the proof shows that this d factor is lost when considering only the last term of the summation in Eq. (21). It is however challenging to control the remaining terms, since B^2 is roughly half of A^2 and $\frac{s^2}{d^2}B^2 - A^2$ is close to 0 for some values of s.

We conclude this section by giving an approximation of Σ^{-1} for large bandwidth. This approximation will be particularly useful in Section 3.1.

Corollary 3 (Large bandwidth approximation of Σ^{-1}). *For any $d \geq 2$, when $\nu \to +\infty$, we have*

$$c_d \longrightarrow \frac{d^2 - 1}{12d}\,,$$

and, as a consequence,

$$\begin{cases} \frac{\sigma_0}{c_d} & \to \frac{2(2d-1)}{d+1} = 4 - \frac{6}{d} + \mathcal{O}\left(\frac{1}{d^2}\right) \\ \frac{\sigma_1}{c_d} & \to \frac{-6}{d+1} = -\frac{6}{d} + \mathcal{O}\left(\frac{1}{d^2}\right) \\ \frac{\sigma_2}{c_d} & \to \frac{6(d^2-2d+3)}{(d+1)(d-1)} = 6 - \frac{12}{d} + \mathcal{O}\left(\frac{1}{d^2}\right) \\ \frac{\sigma_3}{c_d} & \to \frac{-6(d-3)}{(d+1)(d-1)} = -\frac{6}{d} + \mathcal{O}\left(\frac{1}{d^2}\right)\,. \end{cases} \tag{23}$$

Proof. The proof is straightforward from the definition of c_d and the σ coefficients, and Corollary 1. □

1.3 Concentration of $\hat{\Sigma}_n$

We now turn to the concentration of $\hat{\Sigma}_n$ around Σ. More precisely, we show that $\hat{\Sigma}_n$ is close to Σ in operator norm, with high probability. Since the definition of $\hat{\Sigma}_n$ is identical to the one in the Tabular LIME case, we can use the proof machinery of Garreau and von Luxburg (2020b).

Proposition 8 (Concentration of $\hat{\Sigma}_n$). *For any $t \geq 0$,*

$$\mathbb{P}\left(\left\|\hat{\Sigma}_n - \Sigma\right\|_{\text{op}} \geq t\right) \leq 4d \cdot \exp\left(\frac{-nt^2}{32d^2}\right)\,.$$

Proof. We can write $\hat{\Sigma} = \frac{1}{n}\sum_i \pi_i Z_i Z_i^\top$. The summands are bounded i.i.d. random variables, thus we can apply the matrix version of Hoeffding inequality. More precisely, the entries of $\hat{\Sigma}_n$ belong to $[0,1]$ by construction, and Corollary 2 guarantees that the entries of Σ also belong to $[0,1]$. Therefore, if we set $M_i := \frac{1}{n}\pi_i Z_i Z_i^\top - \Sigma$, then the M_i satisfy the assumptions of Theorem 21 in Garreau and von Luxburg (2020b) and we can conclude since $\frac{1}{n}\sum_i M_i = \hat{\Sigma}_n - \Sigma$. □

1.4 Control of $\left\|\Sigma^{-1}\right\|_{\text{op}}$

We now turn to the control of $\left\|\Sigma^{-1}\right\|_{\text{op}}$. Essentially, our strategy is to bound the entries of Σ^{-1}, and then to derive an upper bound for $\left\|\Sigma^{-1}\right\|_{\text{op}}$ by noticing that $\left\|\Sigma^{-1}\right\|_{\text{op}} \leq \left\|\Sigma^{-1}\right\|_{\text{F}}$. Thus let us start by controlling the σ coefficients in absolute value.

Lemma 1 (Control of the σ coefficients). *Let $d \geq 2$ and $\nu \geq 1.66$. Then it holds that*

$$|\sigma_0| \leq \frac{d}{3}\,, \quad |\sigma_1| \leq 1\,, \quad |\sigma_2| \leq \frac{3d}{2}\mathrm{e}^{\frac{1}{2\nu^2}}\,, \quad \text{and} \quad |\sigma_3| \leq \frac{3}{2}\mathrm{e}^{\frac{1}{2\nu^2}}\,.$$

Proof. By its definition, we know that σ_0 is positive. Moreover, from Corollary 2, we see that

$$\begin{aligned} \sigma_0 &= (d-1)\alpha_2 + \alpha_1 \\ &\leq \frac{(d-1)(d-2)}{3d} + \frac{d-1}{2d} \\ &= \frac{2d^2 - 3d + 3}{6d}\,. \end{aligned}$$

One can check that for any $d \geq 2$, we have $2d^2 - 3d + 3 \leq 2d^2$, which concludes the proof of the first claim.

Since $|\sigma_1| = \alpha_1$, the second claim is straightforward from Corollary 2.

Regarding σ_2, we notice that

$$\sigma_2 = \frac{c_d + \alpha_1^2 - \alpha_0\alpha_2}{\alpha_1 - \alpha_2} \,.$$

Since $\alpha_0 \geq \alpha_1 \geq \alpha_2$, we have

$$-\alpha_1(\alpha_0 - \alpha_1) \leq \alpha_1^2 - \alpha_0\alpha_2 \leq \alpha_0(\alpha_1 - \alpha_2) \,.$$

Using Eqs. (18) and (19) in conjunction with Corollary 2, we find that $\left|\alpha_1^2 - \alpha_0\alpha_2\right| \leq 1/4$. Moreover, from Eq. (20), we see that $c_d \leq d/4$. We deduce that

$$|\sigma_2| \leq \left(\frac{d}{4} + \frac{1}{4}\right) \cdot 6\mathrm{e}^{\frac{1}{2\nu^2}} \,,$$

where we used the first statement of Proposition 7 to lower bound $\alpha_1\alpha_2$. The results follows, since $d \geq 2$.

Finally, we write

$$\begin{aligned} |\sigma_3| &= \frac{\left|\alpha_1^2 - \alpha_0\alpha_2\right|}{\alpha_1 - \alpha_2} \\ &\leq \frac{1/4}{\frac{d+1}{6d} \cdot \mathrm{e}^{\frac{-1}{2\nu^2}}} \end{aligned}$$

according to Proposition 7. □

We now proceed to bound the operator norm of Σ^{-1}.

Proposition 9 (Control of $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$). *For any $d \geq 2$ and any $\nu > 0$, it holds that*

$$\left\|\Sigma^{-1}\right\|_{\mathrm{op}} \leq 70 d^{3/2} \mathrm{e}^{\frac{5}{2\nu^2}} \,.$$

Remark 2. We notice that the control obtained worsens as $d \to +\infty$ and $\nu \to 0$. We conjecture that the dependency in d is not tight. For instance, showing that $c_d = \Omega(d)$ (that is, improving Proposition 7) would yield an upper bound of order d instead of $d^{3/2}$. The discussion after Proposition 7 indicates that such an improvement may be possible. Moreover, we see in experiments that the concentration of $\hat{\beta}_n$ does not degrade that much for large d (see, in particular, Figure 17 in Section 6.2), another sign that Proposition 9 could be improved.

Proof. We will use the fact that $\left\|\Sigma^{-1}\right\|_{\mathrm{op}} \leq \left\|\Sigma^{-1}\right\|_{\mathrm{F}}$. We first write

$$\left\|\Sigma^{-1}\right\|_{\mathrm{F}}^2 = \frac{1}{c_d^2}\left(\sigma_0^2 + 2d\sigma_1^2 + d\sigma_2^2 + (d^2 - d)\sigma_3^2\right) \,,$$

by definition of the σ coefficients. On one hand, using Lemma 1, we write

$$\begin{aligned} \sigma_0^2 + 2d\sigma_1^2 + d\sigma_2^2 + (d^2 - d)\sigma_3^2 &\leq \frac{d^2}{9} + 2d + d \cdot (3d/2)^2 \mathrm{e}^{\frac{1}{\nu^2}} + (d^2 - d) \cdot \frac{9}{4}\mathrm{e}^{\frac{1}{\nu^2}} \\ &\leq 3d^3 \mathrm{e}^{\frac{1}{\nu^2}} \,, \end{aligned} \tag{24}$$

where we used $c_d \leq d$ and $d \geq 2$ in the last display. On the other hand, a direct consequence of Proposition 7 is that

$$\frac{1}{c_d^2} \leq 1600 \mathrm{e}^{\frac{4}{\nu^2}} \,. \tag{25}$$

Putting together Eqs. (24) and (25), we obtain the claimed result, since $\sqrt{3 \cdot 1600} \leq 70$. □

2 The study of Γ^f

We now turn to the study of the (weighted) responses. In Section 2.1, we obtain an explicit expression for the average responses. We show how to obtain closed-form expressions in the case of indicator functions in Section 2.2. In the case of a linear model, we have to resort to approximations that are detailed in Section 2.3. Section 2.4 contains the concentration result for $\hat{\Gamma}_n$.

2.1 Computation of Γ^f

We start our study by giving an expression for Γ^f for any f under mild assumptions. Recall that we defined $\hat{\Gamma}_n = \frac{1}{n} Z^\top W y$, where $y \in \mathbb{R}^{d+1}$ is the random vector defined coordinate-wise by $y_i = f(x_i)$. From the definition of $\hat{\Gamma}_n$, it is straightforward that

$$\hat{\Gamma}_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \pi_i f(\phi(x_i)) \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} f(\phi(x_i)) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} f(\phi(x_i)) \end{pmatrix} \in \mathbb{R}^{d+1} .$$

As a consequence, since we defined $\Gamma^f = \mathbb{E}[\hat{\Gamma}_n]$, it holds that

$$\Gamma^f = \begin{pmatrix} \mathbb{E}\left[\pi f(\phi(x))\right] \\ \mathbb{E}\left[\pi z_1 f(\phi(x))\right] \\ \vdots \\ \mathbb{E}\left[\pi z_d f(\phi(x))\right] \end{pmatrix} . \tag{26}$$

Of course, Eq. (26) depends on the model f. These computations can be challenging. Nevertheless, it is possible to obtain exact results in simple situations.

Constant model. As a warm up, let us show how to compute Γ^f when f is constant. Perhaps the simplest model of all: f always returns the same value, whatever the value of $\phi(x)$ may be. By linearity of Γ^f (see Section 3.2 of the main paper), it is sufficient to consider the case $f = 1$. From Eq. (26), we see that

$$\Gamma^f_j = \begin{cases} \mathbb{E}\left[\pi\right] & \text{if } j = 0, \\ \mathbb{E}\left[\pi z_j\right] & \text{otherwise.} \end{cases}$$

We recognize the definitions of the α coefficients, and, more precisely, $\Gamma^f_0 = \alpha_0$ and $\Gamma^f_j = \alpha_1$ if $j \geq 1$.

2.2 Indicator functions

Let us turn to a slightly more complicated class of models: indicator functions, or rather products of indicator functions. As explained in the paper, these functions fall into our framework. We have the following result:

Proposition 10 (Computation of Γ^f, product of indicator functions). *Set $J \subseteq \{1, \dots, d\}$ a set of p distinct indices. Define*

$$f(\phi(x)) := \prod_{j \in J} \mathbf{1}_{\phi(x)_j > 0} .$$

Then it holds that

$$\Gamma^f_\ell = \begin{cases} \alpha_p & \text{if } \ell \in \{0\} \cup J \\ \alpha_{p+1} & \text{otherwise.} \end{cases}$$

Proof. As noticed in the paper, f can be written as a product of z_js. Therefore, we only have to compute

$$\mathbb{E}\left[\pi \prod_{j \in J} z_j\right] \quad \text{and} \quad \mathbb{E}\left[\pi z_k \prod_{j \in J} z_j\right] ,$$

for any $1 \leq k \leq d$. The first term is α_p by definition. For the second term, we notice that if $\ell \in \{0\} \cup J$, then two terms are identical in the product of binary features, and we recognize the definition of α_p. In all other cases, there are no cancellation and we recover the definition of α_{p+1}. □

2.3 Linear model

We now consider a linear model, that is,

$$f(\phi(x)) := \sum_{j=1}^{d} \lambda_j \phi(x)_j \,, \tag{27}$$

where $\lambda_1, \dots, \lambda_d$ are arbitrary fixed coefficients. In order to simplify the computations, we will consider that $\nu \to +\infty$ in this section. In that case, $\pi \xrightarrow{\text{a.s.}} 1$. It is clear that f is bounded on S^{D-1}, thus, by dominated convergence,

$$\Gamma^f \longrightarrow \Gamma_\infty := \begin{pmatrix} \mathbb{E}\left[f(\phi(x)\right] \\ \mathbb{E}\left[z_1 f(\phi(x)\right] \\ \vdots \\ \mathbb{E}\left[z_d f(\phi(x)\right] \end{pmatrix} \in \mathbb{R}^{d+1} \,. \tag{28}$$

By linearity of $f \mapsto \Gamma^f_\infty$, it is sufficient to compute $\mathbb{E}\left[\phi(x)_j\right]$ and $\mathbb{E}\left[z_k \phi(x)_j\right]$ for any $1 \leq j, k \leq d$.

For any $1 \leq j \leq d$, recall that we defined

$$\omega_k = \frac{m_j^2 v_j^2}{\sum_{k=1}^d m_k^2 v_k^2} \,,$$

and $H_S := \sum_{k \in S} \omega_k$, where S is the random subset of indices chosen by LIME. The motivation for the definition of the random variable H_S is the following proposition: it is possible to write the expected TF-IDF as an expression depending on H_S.

Proposition 11 (Expected normalized TF-IDF). *Let w_j be a fixed word of ξ. Then, it holds that*

$$\mathbb{E}\left[\phi(x)_j\right] = \mathbb{E}\left[z_j \phi(x)_j\right] = \frac{d-1}{2d} \cdot \phi(\xi)_j \cdot \mathbb{E}\left[\frac{1}{\sqrt{1-H_S}} \middle| S \not\ni j\right] \,, \tag{29}$$

and, for any $k \neq j$,

$$\mathbb{E}\left[z_k \phi(x)_j\right] = \frac{d-2}{3d} \cdot \phi(\xi)_j \cdot \mathbb{E}\left[\frac{1}{\sqrt{1-H_S}} \middle| S \not\ni j, k\right] \,. \tag{30}$$

Proof. We start by proving Eq (29). Let us split the expectation depending on $w_j \in x$. Since the term frequency is 0 if $w_j \notin x$, we have

$$\mathbb{E}\left[\phi(x)_j\right] = \mathbb{E}\left[\phi(x)_j | w_j \in x\right] \mathbb{P}\left(w_j \in x\right) \,. \tag{31}$$

Lemma 5 gives us the value of $\mathbb{P}\left(w_j \in x\right)$. Let us focus on the TF-IDF term in Eq. (31). By definition, it is the product of the term frequency and the inverse document frequency, normalized. Since the latter does not change when words are removed from ξ, only the norm changes: we have to remove all terms indexed by S. For any $1 \leq j \leq d$, let us set m_j (resp. v_j) the term frequency (resp. the inverse term frequency) of w_j Conditionally to $\{w_j \in x\}$,

$$\phi(x)_j = \frac{m_j v_j}{\sqrt{\sum_{k \notin S} m_k^2 v_k^2}} \,.$$

Let us factor out $\phi(\xi)_j$ in the previous display. By definition of H_S, we have

$$\phi(x)_j = \phi(\xi)_j \cdot \frac{1}{\sqrt{1 - \sum_{k \in S} \frac{m_k^2 v_k^2}{\|\varphi(\xi)\|^2}}} = \phi(\xi)_j \cdot \frac{1}{\sqrt{1-H_S}} \,.$$

Since $\{w_j \in x\}$ is equivalent to $\{j \notin S\}$ by construction, we can conclude. The proof of the second statement is similar; one just has to condition with respect to $\{w_j, w_k \in x\}$ instead, which is equivalent to $\{S \not\ni j, k\}$. □

As a direct consequence of Proposition 11, we can derive $\Gamma^f_\infty = \lim_{\nu \to +\infty} \Gamma^f$ when $f : x \mapsto x_j$. Recall that we set $E_j = \mathbb{E}\left[(1-H_S)^{-1/2} \middle| S \not\ni j\right]$ and $E_{j,k} = \mathbb{E}\left[(1-H_S)^{-1/2} \middle| S \not\ni j, k\right]$. Then

$$\left(\Gamma^f_\infty\right)_k = \begin{cases} \left(\frac{1}{2} - \frac{1}{2d}\right) \cdot E_j \cdot \phi(\xi)_j & \text{if } k = 0 \text{ or } k = j, \\ \left(\frac{1}{3} - \frac{2}{3d}\right) \cdot E_{j,k} \cdot \phi(\xi)_j & \text{otherwise.} \end{cases} \tag{32}$$

In practice, the expectation computations required to evaluate E_j and $E_{j,k}$ are not tractable as soon as d is large. Indeed, in that case, the law of H_S is unknown and approximating the expectation by Monte-Carlo methods requires is hard since one has to sum over all subsets and there are $\mathcal{O}\left(2^d\right)$ subsets S such that $S \subseteq \{1, \dots, d\}$. Therefore we resort to approximate expressions for these expected values computations.

We start by writing

$$\mathbb{E}\left[\frac{1}{\sqrt{1-X}}\right] \approx \frac{1}{\sqrt{1-\mathbb{E}\left[X\right]}}\,. \tag{33}$$

All that is left to compute will be $\mathbb{E}\left[H_S | S \not\ni j\right]$ and $\mathbb{E}\left[H_S | S \not\ni j, k\right]$. We see in Section 4 that after some combinatoric considerations, it is possible to obtain these expected values as a function of ω_j and ω_k. More precisely, Lemma 3 states that

$$\mathbb{E}\left[H_S | S \not\ni j\right] = \frac{1-\omega_j}{3} + \mathcal{O}\left(\frac{1}{d}\right) \quad \text{and} \quad \mathbb{E}\left[H_S | S \not\ni j, k\right] = \frac{1-\omega_j-\omega_k}{4} + \mathcal{O}\left(\frac{1}{d}\right)\,. \tag{34}$$

When d is large and the ω_ks are small, using Eq. (33), we obtain the following approximations:

$$\mathbb{E}\left[\phi(x)_j\right] \approx \frac{1}{2} \cdot \sqrt{\frac{1}{1-\frac{1}{3}}} \cdot \phi(\xi)_j \approx 0.61 \cdot \phi(\xi)_j\,, \tag{35}$$

and, for any $k \neq j$,

$$\mathbb{E}\left[z_k \phi(x)_j\right] \approx \frac{1}{3} \cdot \sqrt{\frac{1}{1-\frac{1}{4}}} \cdot \phi(\xi)_j \approx 0.38 \cdot \phi(\xi)_j\,. \tag{36}$$

For all practical purposes, we will use Eq. (35) and (36).

Remark 3. One could obtain better approximations than above in two ways. First, it is possible to take into account the dependency in ω_j and ω_k in the expectation of H_S. That is, plugging Eq. (34) into Eq. (33) instead of the numerical values 1/3 and 1/4. This yields more accurate, but more complicated formulas. Without being so precise, it is also possible to consider an arbitrary distribution for the ω_ks (for instance, assuming that the term frequencies follow the Zipf's law (Powers, 1998)). Second, since the mapping $\theta : x \mapsto \frac{1}{\sqrt{1-x}}$ is convex, by Jensen's inequality, we are always *underestimating* by considering $\theta(\mathbb{E}\left[X\right])$ instead of $\mathbb{E}\left[\theta(X)\right]$. Going further in the Taylor expansion of θ is a way to fix this problem, namely using

$$\mathbb{E}\left[\frac{1}{\sqrt{1-X}}\right] \approx \frac{1}{\sqrt{1-\mathbb{E}\left[X\right]}} + \frac{3\mathrm{Var}\left(X\right)}{8\sqrt{1-\mathbb{E}\left[X\right]}}\,,$$

instead of Eq. (33). We found that **it was not useful to do so from an experimental point of view:** our theoretical predictions match the experimental results while remaining simple enough.

2.4 Concentration of $\hat{\Gamma}_n$

We now show that $\hat{\Gamma}_n$ is concentrated around Γ^f. Since the expression of $\hat{\Gamma}_n$ is the same than in the tabular case, and since f is bounded on the unit sphere S^{D-1}, the same reasoning as in the proof of Proposition 24 in Garreau and von Luxburg (2020b) can be applied.

Proposition 12 (Concentration of $\hat{\Gamma}_n$). *Assume that f is bounded by $M > 0$ on S^{D-1}. Then, for any $t > 0$, it holds that*

$$\mathbb{P}\left(\|\hat{\Gamma}_n - \Gamma^f\| \geq t\right) \leq 4d\exp\left(\frac{-nt^2}{32Md^2}\right)\,.$$

Proof. Recall that $\|\phi(x)\| = 1$ almost surely. Since f is bounded by M on S^{D-1}, it holds that $|f(\phi(x))| \leq M$ almost surely. We can then proceed as in the proof of Proposition 24 in Garreau and von Luxburg (2020b). □

3 The study of β^f

In this section, we study the interpretable coefficients. We start with the computation of β^f in Section 3.1. In Section 3.2, we show how $\hat{\beta}_n$ concentrates around β^f.

3.1 Computation of β^f

Recall that, for any model f, we have defined $\beta^f = \Sigma^{-1}\Gamma^f$. Directly multiplying the expressions found for Σ^{-1} (Eq. (17)) and Γ^f (Eq. (26)) obtained in the previous sections, we obtain the expression of β^f in the general case (this is Proposition 2 in the paper).

Proposition 13 (Computation of β^f, general case). *Assume that f is bounded on the unit sphere. Then*

$$\beta_0^f = c_d^{-1}\left\{\sigma_0 \mathbb{E}\left[\pi f(\phi(x))\right] + \sigma_1 \sum_{k=1}^{d} \mathbb{E}\left[\pi z_k f(\phi(x))\right]\right\}, \tag{37}$$

and, for any $1 \leq j \leq d$,

$$\beta_j^f = c_d^{-1}\left\{\sigma_1 \mathbb{E}\left[\pi f(\phi(x))\right] + \sigma_2 \mathbb{E}\left[\pi z_j f(\phi(x))\right] + \sigma_3 \sum_{\substack{k=1 \\ k\neq j}}^{d} \mathbb{E}\left[\pi z_k f(\phi(x))\right]\right\}. \tag{38}$$

This is Proposition 2 in the paper, with the additional expression of the intercept β_0^f. Let us see how to obtain an approximate, simple expression when both the bandwidth parameter and the size of the local dictionary are large. When $\nu \to +\infty$, using Corollary 3, we find that

$$\beta_0^f \longrightarrow \left(\beta_\infty^f\right)_0 := \frac{4d-2}{d+1}\mathbb{E}\left[\pi f(\phi(x))\right] - \frac{6}{d+1}\sum_{k=1}^{d}\mathbb{E}\left[\pi z_k f(\phi(x))\right],$$

and, for any $1 \leq j \leq d$,

$$\beta_j^f \longrightarrow \left(\beta_\infty^f\right)_j := \frac{-6}{d+1}\mathbb{E}\left[\pi f(\phi(x))\right] + \frac{6(d^2-2d+3)}{d^2-1}\mathbb{E}\left[\pi z_j f(\phi(x))\right] - \frac{6(d-3)}{d^2-1}\sum_{k\neq j}\mathbb{E}\left[\pi z_k f(\phi(x))\right].$$

For large d, since f is bounded on S^{D-1}, we find that

$$\left(\beta_\infty^f\right)_0 = 4\mathbb{E}\left[\pi f(\phi(x))\right] - \frac{6}{d}\sum_{k=1}^{d}\mathbb{E}\left[\pi z_k f(\phi(x))\right] + \mathcal{O}\left(\frac{1}{d}\right),$$

and, for any $1 \leq j \leq d$,

$$\left(\beta_\infty^f\right)_j = 6\mathbb{E}\left[\pi z_j f(\phi(x))\right] - \frac{6}{d}\sum_{k\neq j}\mathbb{E}\left[\pi z_k f(\phi(x))\right] + \mathcal{O}\left(\frac{1}{d}\right).$$

Now, by definition of the interpretable features, for any $1 \leq j \leq d$,

$$\begin{aligned}\mathbb{E}\left[\pi z_j f(\phi(x))\right] &= \mathbb{E}\left[\pi z_j f(\phi(x)) | w_j \in x\right] \cdot \mathbb{P}\left(w_j \in x\right) + \mathbb{E}\left[\pi z_j f(\phi(x)) | w_j \notin x\right] \cdot \mathbb{P}\left(w_j \notin x\right) \\ &= \mathbb{E}\left[\pi f(\phi(x)) | w_j \in x\right] \cdot \frac{d-1}{2d} + 0\,,\end{aligned}$$

where we used Lemma 5 in the last display. Therefore, we have the following approximations of the interpretable coefficients:

$$\left(\beta_\infty^f\right)_0 = 2\mathbb{E}\left[\pi f(\phi(x))\right] - \frac{3}{d}\sum_{k}\mathbb{E}\left[\pi f(\phi(x)) | w_k \in x\right] + \mathcal{O}\left(\frac{1}{d}\right), \tag{39}$$

and, for any $1 \leq j \leq d$,

$$\left(\beta_\infty^f\right)_j = 3\mathbb{E}\left[\pi f(\phi(x)) | w_j \in x\right] - \frac{3}{d}\sum_{k}\mathbb{E}\left[\pi f(\phi(x)) | w_k \in x\right] + \mathcal{O}\left(\frac{1}{d}\right). \tag{40}$$

The last display is the approximation of Proposition 13 presented in the paper.

Remark 4. In Garreau and von Luxburg (2020b), it is noted that LIME for tabular data provably ignores unused coordinates. In other words, if the model f does not depend on coordinate j, then the explanation β_j^f is 0. We could not prove such a statement in the case of text data, even for simplified expressions such as Eq. (40).

We now show how to compute β^f in specific cases, thus returning to generic ν and d.

Constant model. As a warm up exercise, let us assume that f is a constant, which we set to 1 without loss of generality (by linearity). Recall that, in that case, $\Gamma_0^f = \alpha_0$ and $\Gamma_j^f = \alpha_1$ for any $1 \leq j \leq d$. From the definition of c_d and the σ coefficients (Proposition 6), we find that

$$\begin{cases} \sigma_0\alpha_0 + d\sigma_1\alpha_1 & = c_d\,, \\ \sigma_1\alpha_0 + \sigma_2\alpha_1 + (d-1)\sigma_3\alpha_1 & = 0\,. \end{cases}$$

We deduce from Proposition 13 that $\beta_0^f = 1$ and $\beta_j^f = 0$ for any $1 \leq j \leq d$. This is conform to our intuition: if the model is constant, then no word should receive nonzero weight in the explanation provided by Text LIME.

Indicator functions. We now turn to indicator functions, more precisely *products* of indicator functions. We will prove the following (Proposition 3 in the paper):

Proposition 14 (Computation of β^f, product of indicator functions). *Let $j \subseteq \{1, \dots, d\}$ be a set of p distinct indices and set $f(x) = \prod_{j \in J} \mathbf{1}_{x_j > 0}$. Then*

$$\begin{cases} \beta_0^f & = c_d^{-1}\left(\sigma_0\alpha_p + p\sigma_1\alpha_p + (d-p)\sigma_1\alpha_{p+1}\right)\,, \\ \beta_j^f & = c_d^{-1}\left(\sigma_1\alpha_p + \sigma_2\alpha_p + (d-p)\sigma_3\alpha_{p+1} + (p-1)\sigma_3\alpha_p\right) \;\; \textit{if } j \in J\,, \\ \beta_j^f & = c_d^{-1}\left(\sigma_1\alpha_p + \sigma_2\alpha_{p+1} + (d-p-1)\sigma_3\alpha_{p+1} + p\sigma_3\alpha_p\right) \;\; \textit{otherwise}\,. \end{cases}$$

Proof. The proof is straightforward from Proposition 10 and Proposition 13. □

Linear model. In this last paragraph, we treat the linear case. As noted in Section 2.3, we have to resort to approximate computations: in this paragraph, we assume that $\nu = +\infty$. We start with the simplest linear function: all coefficients are zero except one (this is Proposition 4 in the paper).

Proposition 15 (Computation of β^f, linear case). *Let $1 \leq j \leq d$ and assume that $f(\phi(x)) = \phi(x)_j$. Recall that we set $E_j = \mathbb{E}\left[(1 - H_S)^{-1/2} \middle| S \not\ni j\right]$ and for any $k \neq j$, $E_{j,k} = \mathbb{E}\left[(1 - H_S)^{-1/2} \middle| S \not\ni j, k\right]$. Then*

$$\left(\beta_\infty^f\right)_0 = \left\{5E_j - \frac{2}{d}\sum_{k \neq j} E_{j,k}\right\}\phi(\xi)_j + \mathcal{O}\left(\frac{1}{d}\right)$$

for any $k \neq j$,

$$\left(\beta_\infty^f\right)_k = \left\{2E_{j,1} - \frac{2}{d}\sum_{\ell \neq k,j} E_{j,\ell}\right\}\phi(\xi)_j + \mathcal{O}\left(\frac{1}{d}\right)\,,$$

and

$$\left(\beta_\infty^f\right)_j = \left\{3E_j - \frac{2}{d}\sum_{k \neq j} E_{j,k}\right\}\phi(\xi)_j + \mathcal{O}\left(\frac{1}{d}\right)\,.$$

Proof. Straightforward from Eqs. (23) and (32). □

Assuming that the ω_k are small, we deduce from Eqs. (35) and (36) that $E_j \approx 1.22$ and $E_{j,k} \approx 1.15$. In particular, *they do not depend on j and k.* Thus we can drastically simplify the statement of Proposition 15:

$$\forall k \neq j, \quad \left(\beta_\infty^f\right)_k \approx 0 \quad \text{and} \quad \left(\beta_\infty^f\right)_j \approx 1.36\phi(\xi)_j\,. \tag{41}$$

We can now go back to our original goal: $f(x) = \sum_{j=1}^d \lambda_j x_j$. By linearity, we deduce from Eq. (41) that

$$\forall 1 \leq j \leq d, \quad \left(\beta_\infty^f\right)_j \approx 1.36 \cdot \lambda_j \cdot \phi(\xi)_j\,. \tag{42}$$

In other words, as noted in the paper, **the explanation for a linear f is the TF-IDF of the word multiplied by the coefficient of the linear model,** up to a numerical constant and small error terms depending on d.

3.2 Concentration of $\hat{\beta}$

In this section, we state and prove our main result: the concentration of $\hat{\beta}_n$ around β^f with high probability (this is Theorem 1 in the paper).

Theorem 2 (Concentration of $\hat{\beta}_n$). *Suppose that f is bounded by $M > 0$ on S^{D-1}. Let $\epsilon > 0$ be a small constant, at least smaller than M. Let $\eta \in (0,1)$. Then, for every*

$$n \geq \max\left\{2^9 \cdot 70^4 M^2 d^9 \mathrm{e}^{\frac{10}{\nu^2}}, 2^9 \cdot 70^2 M d^5 \mathrm{e}^{\frac{5}{\nu^2}}\right\} \frac{\log \frac{8d}{\eta}}{\epsilon^2},$$

we have $\mathbb{P}\left(\|\hat{\beta}_n - \beta^f\| \geq \epsilon\right) \leq \eta$.

Proof. We follow the proof scheme of Theorem 28 in Garreau and von Luxburg (2020b). The key point is to notice that

$$\|\hat{\beta}_n - \beta^f\| \leq 2\left\|\Sigma^{-1}\right\|_{\mathrm{op}} \|\hat{\Gamma} - \Gamma^f\| + 2\left\|\Sigma^{-1}\right\|_{\mathrm{op}}^2 \left\|\Gamma^f\right\| \|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \tag{43}$$

provided that $\|\Sigma^{-1}(\hat{\Sigma} - \Sigma)\|_{\mathrm{op}} \leq 0.32$ (this is Lemma 27 in Garreau and von Luxburg (2020b). Therefore, in order to show that $\|\hat{\beta}_n - \beta^f\| \leq \epsilon$, it suffices to show that each term in Eq. (43) is smaller than $\epsilon/4$ and that $\|\Sigma^{-1}(\hat{\Sigma} - \Sigma)\|_{\mathrm{op}} \leq 0.32$. The concentration results obtained in Section 1 and 2 guarantee that both $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}$ and $\|\hat{\Gamma} - \Gamma^f\|$ are small if n is large enough, with high probability. This, combined with the upper bound on $\|\Sigma^{-1}\|_{\mathrm{op}}$ given by Proposition 9, concludes the proof.

Let us give a bit more details. We start with the control of $\|\Sigma^{-1}(\hat{\Sigma} - \Sigma)\|_{\mathrm{op}}$. Set $t_1 := (220 d^{3/2} \mathrm{e}^{\frac{5}{2\nu^2}})^{-1}$ and $n_1 := 32 d^2 \log \frac{8d}{\eta} / t_1^2$. Then, according to Proposition 8, for any $n \geq n_1$,

$$\mathbb{P}\left(\|\hat{\Sigma}_n - \Sigma\|_{\mathrm{op}} \geq t_1\right) \leq 4d \exp\left(\frac{-n t_1^2}{32 d^2}\right) \leq \frac{\eta}{2}.$$

Since $\|\Sigma^{-1}\|_{\mathrm{op}} \leq 70 d^{3/2} \mathrm{e}^{\frac{5}{2\nu^2}}$ (according to Proposition 9), by sub-multiplicativity of the operator norm, it holds that

$$\|\Sigma^{-1}(\hat{\Sigma} - \Sigma)\|_{\mathrm{op}} \leq \|\Sigma^{-1}\|_{\mathrm{op}} \|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \leq 70/220 < 0.32, \tag{44}$$

with probability greater than $1 - \eta/2$.

Now let us set $t_2 := (4 \cdot 70^2 M d^{7/2} \mathrm{e}^{\frac{5}{\nu^2}})^{-1} \epsilon$ and $n_2 := 32 d^2 \log \frac{8d}{\eta} / t_2^2$. According to Proposition 8, for any $n \geq n_2$, it holds that

$$\|\hat{\Sigma}_n - \Sigma\|_{\mathrm{op}} \leq \frac{\epsilon}{4 M d^{1/2}} \cdot (70^2 d^3 \mathrm{e}^{5/\nu^2})^{-1},$$

with probability greater than $\eta/2$. Since $\|\Gamma^f\| \leq M \cdot d^{1/2}$ and $\|\Sigma^{-1}\|_{\mathrm{op}}^2 \leq 70^2 d^3 \mathrm{e}^{5/\nu^2}$,

$$\left\|\Sigma^{-1}\right\|_{\mathrm{op}} \|\hat{\Gamma} - \Gamma^f\| \leq \frac{\epsilon}{4}$$

with probability grater than $1 - \eta/2$. Notice that, since we assumed $\epsilon < M$, $t_2 < t_1$, and thus Eq. (44) also holds.

Finally, let us set $t_3 := \epsilon / (4 \cdot 70 d^{3/2} \mathrm{e}^{\frac{5}{2\nu^2}})$ and $n_3 := 32 M d^2 \log \frac{8d}{\eta} / t_3^2$. According to Proposition 12, for any $n \geq n_3$,

$$\mathbb{P}\left(\|\hat{\Gamma}_n - \Gamma^f\| \geq t_3\right) \leq 4d \exp\left(\frac{-n t_3^2}{32 M d^2}\right) \leq \frac{\eta}{2}.$$

Since $\|\Sigma^{-1}\|_{\mathrm{op}} \leq 70 d^{3/2} \mathrm{e}^{\frac{5}{2\nu^2}}$, we deduce that

$$\left\|\Sigma^{-1}\right\|_{\mathrm{op}}^2 \left\|\Gamma^f\right\| \|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \leq \frac{\epsilon}{2},$$

with probability greater than $1 - \eta/2$. We conclude by a union bound argument. □

4 Sums over subsets

In this section, independent from the rest, we collect technical facts about sums over subsets. More particularly, we now consider arbitrary, fixed positive real numbers $\omega_1, \dots, \omega_d$ such that $\sum_k \omega_k = 1$. We are interested in subsets S of $\{1, \dots, d\}$. For any such S, we define $H_S := \sum_{k \in S} \omega_k$ the sum of the ω_k coefficients over S. Our main goal in this section is to compute the expectation of H_S conditionally to S not containing a given index (or two given indices), which is the key quantity appearing in Proposition 15.

Lemma 2 (First order subset sums). *Let $1 \leq s \leq d$ and $1 \leq j, k \leq d$ with $j \neq k$. Then*

$$\sum_{\substack{\#S=s \\ S \not\ni j}} H_S = \binom{d-2}{s-1}(1-\omega_j)\,,$$

and

$$\sum_{\substack{\#S=s \\ S \not\ni j,k}} H_S = \binom{d-3}{s-1}(1-\omega_j-\omega_k)\,.$$

Proof. The main idea of the proof is to rearrange the sum, summing over all indices and then counting how many subsets satisfy the condition. That is,

$$\begin{aligned}\sum_{\substack{\#S=s \\ S \ni j}} H_S &= \sum_{k=1}^{d} \omega_k \cdot \#\{S \text{ s.t. } j,k \in S\} \\ &= \sum_{k \neq j} \omega_k \cdot \binom{d-2}{s-2} + \omega_j \cdot \binom{d-1}{s-1} \\ &= \binom{d-2}{s-2} + \left[\binom{d-1}{s-1} - \binom{d-2}{s-2}\right]\omega_j\,.\end{aligned}$$

We conclude by using the binomial identity

$$\binom{d-1}{s-1} - \binom{d-2}{s-2} = \binom{d-2}{s-1}.$$

Notice that, in the previous derivation, we had to split the sum to account for the case $j = k$. The proof of the second formula is similar. □

Let us turn to expectation computation that are important to derive approximation in Section 2.3. We now see S and H_S as random variables. We will denote by $\mathbb{E}_s[\cdot]$ the expectation conditionally to the event $\{\#S = s\}$.

Lemma 3 (Expectation computation). *Let j, k be distinct elements of $\{1, \dots, d\}$. Then*

$$\mathbb{E}[H_S | S \not\ni j] = \frac{(1-\omega_j)(d+1)}{3(d-1)} = \frac{1-\omega_j}{3} + \mathcal{O}\left(\frac{1}{d}\right), \tag{45}$$

and

$$\mathbb{E}[H_S | S \not\ni j,k] = \frac{(1-\omega_j-\omega_k)(d+1)}{4(d-2)} = \frac{1-\omega_j-\omega_k}{4} + \mathcal{O}\left(\frac{1}{d}\right) \tag{46}$$

Proof. By the law of total expectation, we know that

$$\mathbb{E}[H_S | S \not\ni j] = \sum_{s=1}^{d} \mathbb{E}_s[H_S | S \not\ni j] \cdot \mathbb{P}(\#S = s | S \not\ni j)\,.$$

We first notice that, for any $s < d$,

$$\begin{aligned}\mathbb{P}\left(\#S = s | S \not\ni j\right) &= \frac{\mathbb{P}\left(S \not\ni j | \#S = s\right)\mathbb{P}\left(\#S = s\right)}{\mathbb{P}\left(j \notin S\right)} \\ &= \frac{\binom{d-1}{s}/\binom{d}{s} \cdot \frac{1}{d}}{\frac{d-1}{2d}} \\ \mathbb{P}\left(\#S = s | S \not\ni j\right) &= \frac{2(d-s)}{d(d-1)}\,.\end{aligned}$$

According to Lemma 2, for any $1 \leq s < d$,

$$\sum_{\substack{\#S=s \\ S \not\ni j}} H_S = \binom{d-2}{s-1}(1-\omega_j)\,.$$

Moreover, there are $\binom{d-1}{s}$ such subsets. Since $\binom{d-1}{s-1}^{-1}\binom{d-2}{s} = \frac{s}{d-1}$, we deduce that

$$\mathbb{E}_s\left[H_S | S \not\ni j\right] = \frac{s}{d-1}(1-\omega_j)\,.$$

Finally, we write

$$\begin{aligned}\mathbb{E}\left[H_S | S \not\ni j\right] &= \sum_{s=1}^{d-1} \frac{s}{d-1}(1-\omega_j) \cdot \frac{2(d-s)}{d(d-1)} \\ &= (1-\omega_j) \cdot \frac{2}{d(d-1)^2} \sum_{s=1}^{d-1} s(d-s) \\ \mathbb{E}\left[H_S | S \not\ni j\right] &= \frac{(d+1)(1-\omega_j)}{3(d-1)}\,.\end{aligned}$$

The second case is similar. One just has to note that

$$\begin{aligned}\mathbb{P}\left(\#S = s | S \not\ni j, k\right) &= \frac{\mathbb{P}\left(S \not\ni j, k | \#S = s\right)}{\mathbb{P}\left(j, k \notin S\right)} \\ &= \frac{3(d-s)(d-s-1)}{d(d-1)(d-2)}\,. \qquad \text{(Lemma 5)}\end{aligned}$$

Then we can conclude since

$$\sum_{s=1}^{d-2} s(d-s)(d-s-1) = \frac{(d-2)(d-1)d(d+1)}{12}\,.$$

□

5 Technical results

In this section, we collect small probability computations that are ubiquitous in our derivations. We start with the probability for a given word to be present in the new sample x, conditionally to $\#S = s$.

Lemma 4 (Conditional probability to contain given words). *Let $w_1, \ldots, w_p$ be p distinct words of D_ℓ. Then, for any $1 \leq s \leq d$,*

$$\mathbb{P}_s\left(w_1 \in x, \ldots, w_p \in x\right) = \frac{(d-s)(d-s-1)\cdots(d-s-p+1)}{d(d-1)\cdots(d-p+1)} = \frac{(d-s)!}{(d-s-p)!} \cdot \frac{(d-p)!}{d!}\,.$$

In the proofs, we use extensively Lemma 4 for $p = 1$ and $p = 2$, that is,

$$\mathbb{P}_s\left(w_j \in x\right) = \frac{d-s}{d} \quad \text{and} \quad \mathbb{P}_s\left(w_j \in x, w_k \in x\right) = \frac{(d-s)(d-s-1)}{d(d-1)}\,,$$

for any $1 \leq j, k \leq d$ with $j \neq k$.

Proof. We prove the more general statement. Conditionally to $\#S = s$, the choice of S is uniform among all subsets of $\{1, \dots, d\}$ of cardinality s. There are $\binom{d}{s}$ such subsets, and only $\binom{d-p}{s}$ of them do not contain the indices corresponding to $w_1, \dots, w_p$. □

We have the following result, without conditioning on the cardinality of S:

Lemma 5 (Probability to contain given words). *Let $w_1, \dots, w_p$ be p distinct words of D_ℓ. Then*

$$\mathbb{P}\left(w_1, \dots, w_p \in x\right) = \frac{d-p}{(p+1)d}\,.$$

Proof. By the law of total expectation,

$$\begin{aligned}\mathbb{P}\left(w_1, \dots, w_p \in x\right) &= \frac{1}{d}\sum_{s=1}^{d} \mathbb{P}\left(w_1, \dots, w_p \in x | s\right) \\ &= \frac{1}{d}\sum_{s=1}^{d} \frac{(d-s)!}{(d-s-p)!}\cdot\frac{(d-p)!}{d!}\,,\end{aligned}$$

where we used Lemma 4 in the last display. By the hockey-stick identity (Ross, 1997), we have

$$\sum_{s=1}^{d}\binom{d-s}{p} = \sum_{s=p}^{d-1}\binom{s}{p} = \binom{d}{p+1}\,.$$

We deduce that

$$\sum_{s=1}^{d}\frac{(d-s)!}{(d-s-p)!} = \frac{d!}{(p+1)\cdot(d-p-1)!}\,. \tag{47}$$

We deduce that

$$\begin{aligned}\mathbb{P}\left(w_1, \dots, w_p \in x\right) &= \frac{1}{d}\frac{(d-p)!}{d!}\sum_{s=1}^{d}\frac{(d-s)!}{(d-s-p)!} \\ &= \frac{1}{d}\frac{(d-p)!}{d!}\frac{d!}{(p+1)\cdot(d-p-1)!} && \text{(by Eq. (47))} \\ \mathbb{P}\left(w_1, \dots, w_p \in x\right) &= \frac{d-p}{(p+1)d}\,.\end{aligned}$$

□

6 Additional experiments

In this section, we present additional experiments. We collect the experiments related to decision trees in Section 6.1 and those related to linear models in Section 6.2.

Setting. All the experiments presented here and in the paper are done on Yelp reviews (the data are publicly available at `https://www.kaggle.com/omkarsabnis/yelp-reviews-dataset`). For a given model f, the general mechanism of our experiments is the following. For a given document ξ containing d distinct words, we set a bandwidth parameter ν and a number of new samples n. Then we run LIME n_{exp} times on ξ, with no feature selection procedure (that is, all words belonging to the local dictionary receive an explanation). We want to emphasize again that this is the only difference with the default implementation. Unless otherwise specified, the parameters of LIME are chosen by default, that is, $\nu = 0.25$ and $n = 5000$. The number of experiments n_{exp} is set to 100. The whisker boxes are obtained by collecting the empirical values of the n_{exp} runs of LIME: they give an indication as to the variability in explanations due to the sampling of new examples. Generally, we report a subset of the interpretable coefficients, the other having near zero values.

Let us explain briefly how to read these whisker boxes: to each word corresponds a whisker box containing all the n_{exp} values of interpretable coefficients provided by LIME ($\hat{\beta}_j$ in our notation). The horizontal dark lines

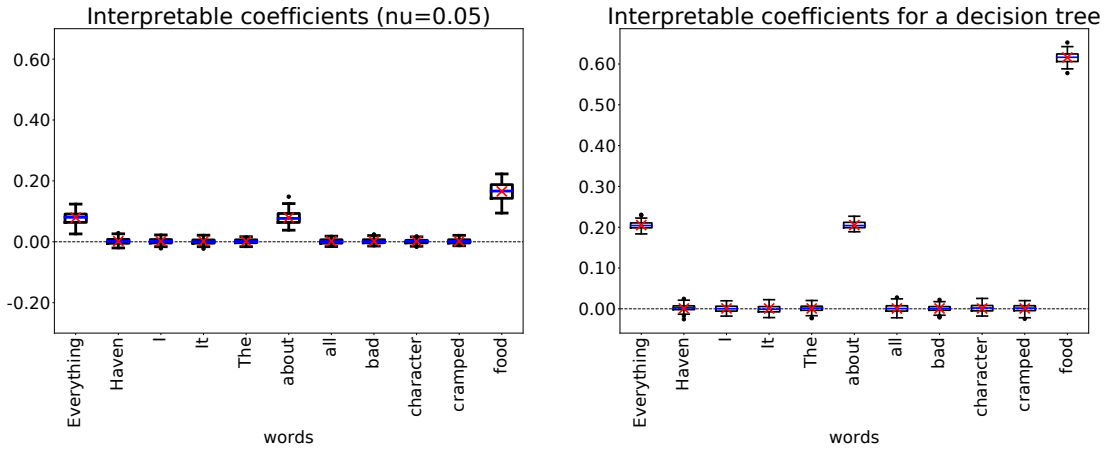

Figure 12: Influence of the bandwidth on the explanation given for a small decision tree on a Yelp review ($n = 5000, n_{\text{exp}} = 100$, $d = 29$). *Left panel:* $\nu = 0.05$, *right panel:* $\nu = 0.35$. Our theoretical predictions remain accurate for non-default bandwidths.

mark the quartiles of these values, and the horizontal blue line is the median. On top of these experimental results, we report with red crosses the values predicted by our analysis (β_j^f in our notation).

The Python code for all experiments is available at `https://github.com/dmardaoui/lime_text_theory`. We encourage the reader to try and run the experiments on other examples of the dataset and with other parameters.

6.1 Decision trees

In this section, we present additional experiments for small decision trees. We begin by investigating the influence of ν and n on the quality of our theoretical predictions.

Influence of the bandwidth. Let us consider the same example ξ and decision tree as in the paper. In particular, the model f is written as

$$\mathbf{1}_{\text{“food”}} + (1 - \mathbf{1}_{\text{“food”}}) \cdot \mathbf{1}_{\text{“about”}} \cdot \mathbf{1}_{\text{“Everything”}} .$$

We now consider non-default bandwidths, that is, bandwidths different than 0.25. We present in Figure 12 the results of these experiments. In the left panel, we took a smaller bandwidth ($\nu = 0.05$) and in the right panel a larger bandwidth ($\nu = 0.35$). We see that while the numerical value of the coefficients changes slightly, their relative order is preserved. Moreover, our theoretical predictions remain accurate in that case, which is to be expected since we did not resort to any approximation in this case. Interestingly, the empirical results for small ν seem more spread out, as hinted by Theorem 2.

Influence of the number of samples. Keeping the same model and example to explain as above, we looked into non-default number of samples n. We present in Figure 13 the results of these experiments. We took a very small n in the left panel ($n = 50$ is two orders of magnitude smaller than the default $n = 5000$) and a larger n in the right panel. As expected, when n is larger, the concentration around our theoretical predictions is even better. To the opposite, for small n, we see that the explanations vary wildly. This is materialized by much wider whisker boxes. Nevertheless, to our surprise, it seems that our theoretical predictions still contain some relevant information in that case.

Influence of depth. Finally, we looked into more complex decision trees. The decision rule used in Figure 14 is given by

$$\mathbf{1}_{\text{“food”}} + (1 - \mathbf{1}_{\text{“food”}})\mathbf{1}_{\text{“about”}}\mathbf{1}_{\text{“Everything”}} + \mathbf{1}_{\text{“bad”}} + \mathbf{1}_{\text{“bad”}}\mathbf{1}_{\text{“character”}} .$$

We see that increasing the depth of the tree is not a problem from a theoretical point of view. It is interesting to see that words used in several nodes for the decision receive more weight (*e.g.*, “bad” in this example).

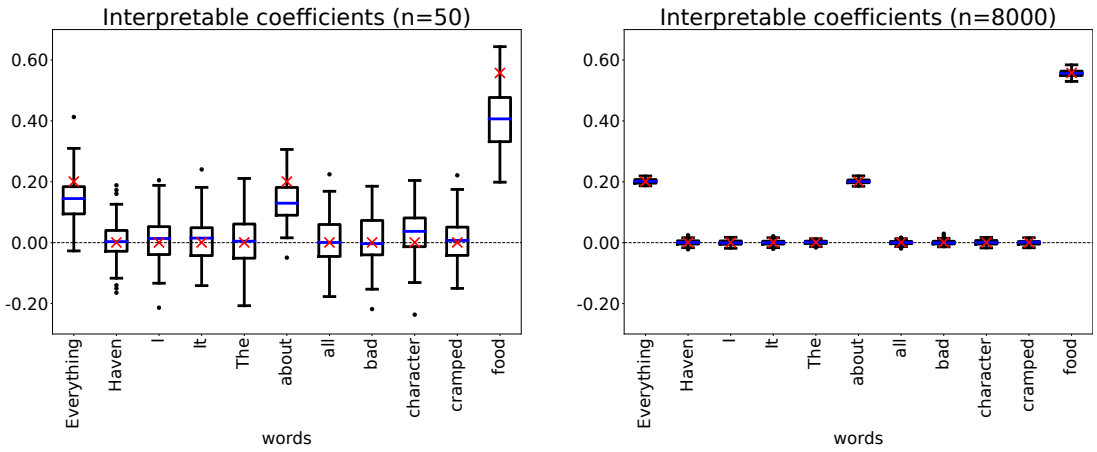

Figure 13: Influence of the number of perturbed samples on the explanation given for a small decision tree on a Yelp review ($\nu = 0.25, n_{\text{exp}} = 100, d = 29$). *Left panel:* $n = 50$, *right panel:* $n = 8000$. Empirical values are less likely to be close to the theoretical predictions for small n.

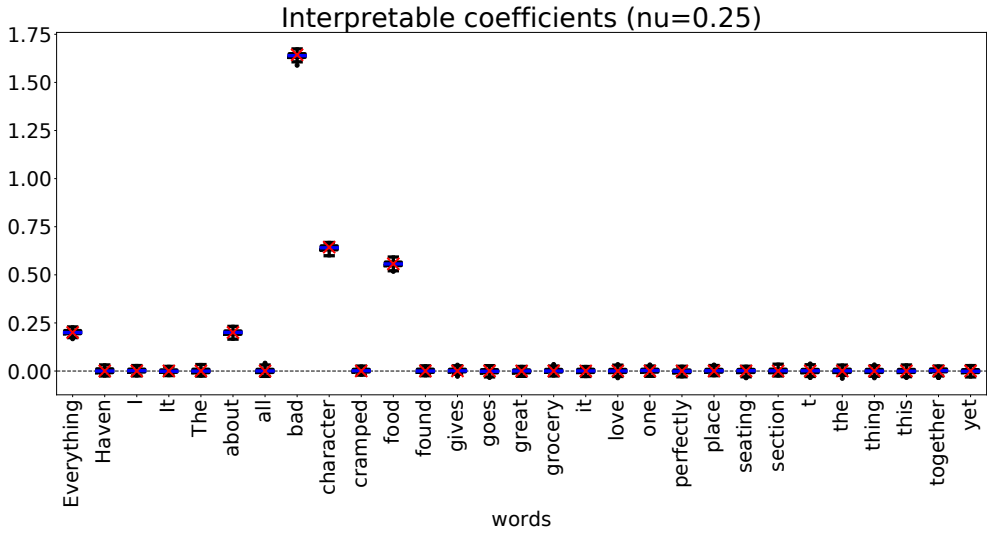

Figure 14: Theory meets practice for a more complex decision tree ($\nu = 0.25, n_{\text{exp}} = 100, n = 5000, d = 29$). Here we report all coefficients. The theory still holds for more complex trees.

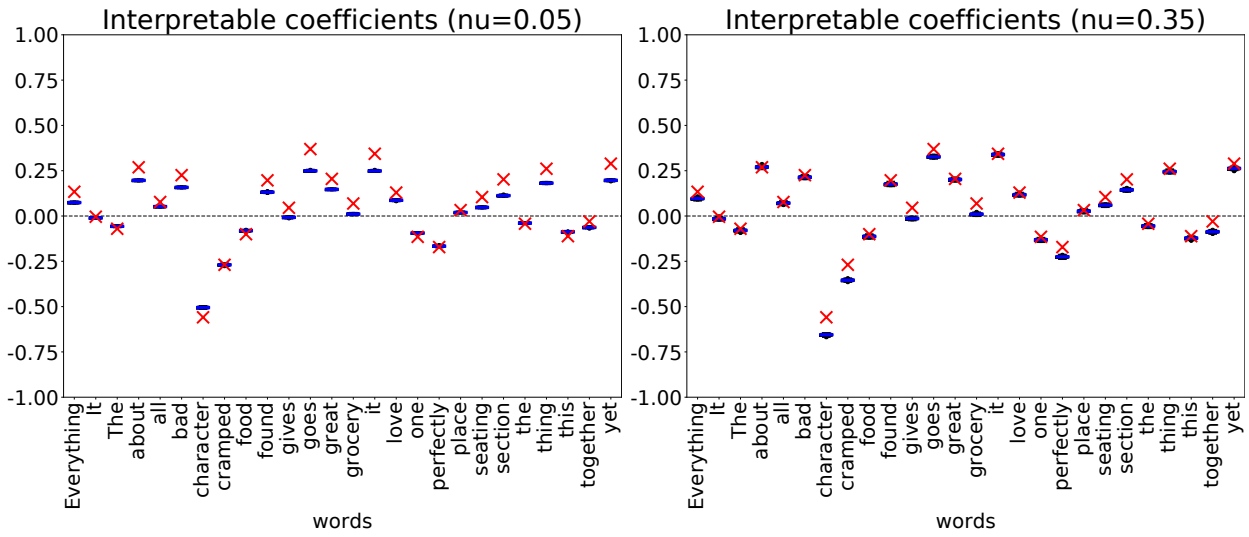

Figure 15: Influence of the bandwidth on the explanation for a linear model on a Yelp review ($n_{\text{exp}} = 100, n = 5000, d = 29$). *Left panel:* $\nu = 0.05$, *right panel:* $\nu = 0.35$. The approximate theoretical values are less accurate for smaller bandwidths.

6.2 Linear models

Let us conclude this section with additional experiments for linear models. As in the paper, we consider an arbitrary linear model

$$f(\phi(x)) = \sum_{j=1}^{d} \lambda_j \phi(x)_j \,.$$

In practice, the coefficients λ_j are drawn i.i.d. according to a Gaussian distribution.

Influence of the bandwidth. As in the previous section, we start by investigating the role of the bandwidth in the accuracy of our theoretical predictions. We see in the right panel of Figure 15 that taking a larger bandwidth does not change much neither the explanations nor the fit between our theoretical predictions and the empirical results. This is expected, since our approximation (Eq. (42)) is based on the large bandwidth approximation. However, the left panel of Figure 15 shows how this approximation becomes dubious when the bandwidth is small. It is interesting to note that in that case, the theory seems to always *overestimate* the empirical results, in absolute value. The large bandwidth approximation is definitely a culprit here, but it could also be the regularization coming into play. Indeed, the discussion at the end of Section 2.4 in the paper that lead us to ignore the regularization is no longer valid for a small ν. In that case, the π_is can be quite small and the first term in Eq. (5) of the paper is of order $\mathrm{e}^{-1/(2\nu^2)}n$ instead of n.

Influence of the number of samples. Now let us look at the influence of the number of perturbed samples. As in the previous section, we look into very small values of n, *e.g.*, $n = 50$. We see in the left panel of Figure 16 that, as expected, the variability of the explanations increases drastically. The theoretical predictions seem to overestimate the empirical results in absolute value, which could again be due to the regularization beginning to play a role for small n, since the discussion in Section 2.4 of the paper is only valid for large n.

Influence of d. To conclude this section, let us note that d does not seem to be a limiting factor in our analysis. While Theorem 2 hints that the concentration phenomenon may worsen for large d, as noted before in Remark 2, we have reason to suspect that it is not the case. All experiments presented on this section so far consider an example whose local dictionary has size $d = 29$. In Figure 17 we present an experiment on an example that has a local dictionary of size $d = 52$. We observed no visible change in the accuracy of our predictions.

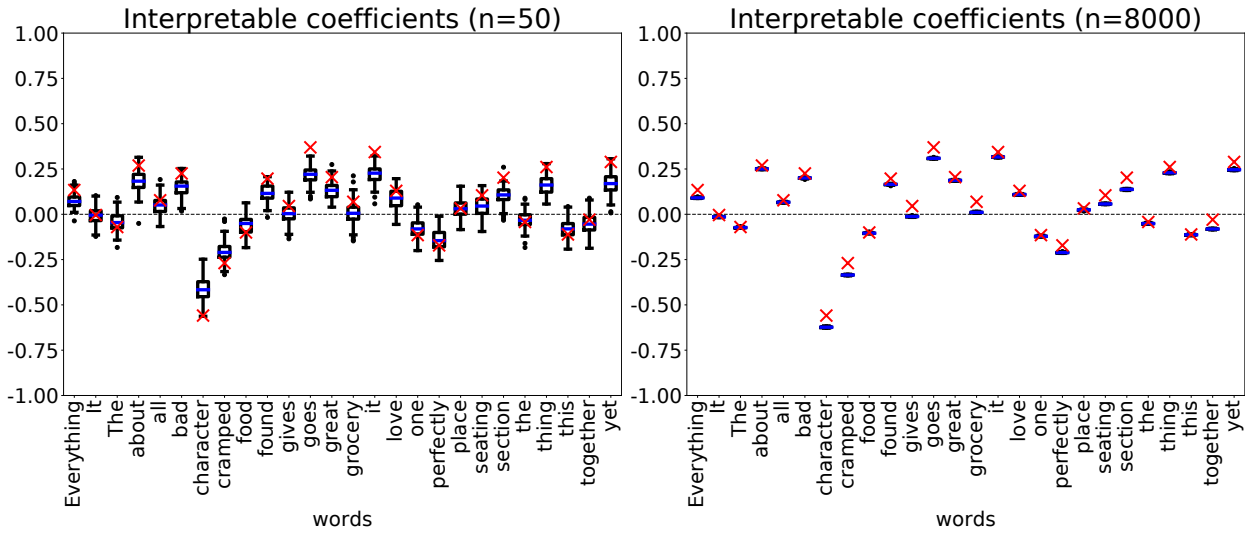

Figure 16: Influence of the number of perturbed samples on the explanation for a linear model on a Yelp review ($\nu = 0.25, n_{\text{exp}} = 100, d = 29$). *Left panel:* $n = 50$, *right panel:* $n = 8000$. The empirical explanations are more spread out for small values of n.

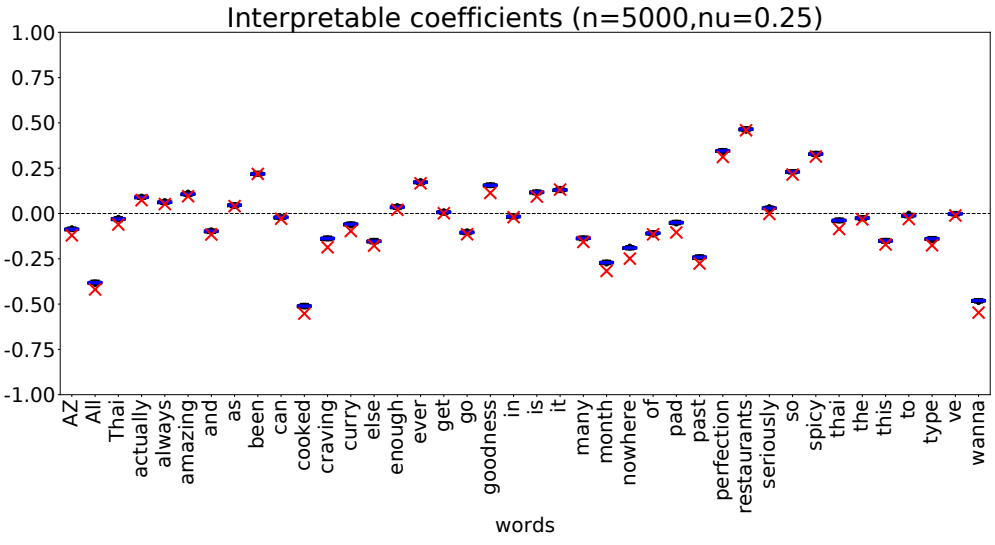

Figure 17: Theory meets practice for an example with a larger vocabulary ($\nu = 0.25, n_{\text{exp}} = 100, n = 5000, d = 52$). Here we report all the interpretable coefficients. Our theoretical predictions seem to hold for larger local dictionaries.