My Title

Lev Stambler
levstamb@umd.edu
University of Maryland, College Park,
NeverLocal Ltd.

September 10, 2025

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri.

1 Introduction

TODO: intro here

2 Core Idea: Datasets and Basic In-Distribution Testing

Ideally, we'd like to take the useful tools of Fourier Analysis (which assumes a product space) and generalize them to any distribution. Though, we do not have a formal method of reasoning about this, we do not think that it is quite possible.

Rather, we will attempt to think about analysis as over a *dataset:* i.e. we will think of our distribution as being defined by a finite-sized dataset, $\mathcal{D}$, and the probability vector will be defined as

$$p(x) = \begin{cases} \frac{1}{|\mathcal{D}|} & \text{if } \ x \in \mathcal{D} \\ 0 \text{ otherwise} \end{cases}.$$

More formally, we will be over a space $\mathcal{T}^n$ (think of $\mathcal{T}$ as either $\mathbb{R}$ or the space of tokens etc.) and $\mathcal{D} \subset \mathcal{T}^n$. Then, we will be focusing on functions $f \in \mathcal{T}^n \to \mathbb{R}$. f can either be a trained model, ideal labeling function on the dataset, or some other efficiently computable function.

Then, we want to reason about Fourier coefficients as

$$\hat{f}(S) = \mathbb{E}_{x \sim \mathcal{D}}[f(x)\chi_S^{-1}(x)].$$

Further, we will abuse notation to denote $\mathcal{D} : \mathcal{T}^n \to \{0, 1\}$ as the indicator function for the inclusion within the dataset (i.e. an in-distribution tester). We will write $\hat{f}_{\mathrm{OG}}(S)$ to denote the normal ("original") Fourier coefficient:

$$\hat{f}_{\mathrm{OG}}(S) = \mathbb{E}_{x \sim \mathcal{T}^n}[f(x)\chi_S^{-1}(x)].$$

Importantly, notice that

$$\hat{f}(S) = \frac{|\mathcal{D}|}{|\mathcal{T}|^n}^{-1} \hat{f}_{\mathrm{OG}}(\mathcal{D} \circ f) = \frac{|\mathcal{D}|}{|\mathcal{T}|^n}^{-1} \cdot \frac{1}{|\mathcal{T}|^n} \sum_{x \in \mathcal{D}} f(x)\chi_S(x).$$

where $\circ$ is the element wise composition. TODO: I think we need to use χ^{-1}?

We conveniently have our first lemma.

Lemma 2.1: For all $x \in \mathcal{D}$,

$$f(x) = \frac{|\mathcal{D}|}{|\mathcal{T}|^n} \sum_S \hat{f}(S) \chi_S(x)$$

Proof: First, note that $f(x) = \mathcal{D}(x) \cdot f(x)$ for $x \in \mathcal{D}$ and then $\widehat{\mathcal{D} \circ f}_{\mathrm{OG}}(S) = \frac{|\mathcal{D}|}{|\mathcal{T}|^n} \cdot \hat{f}(S)$ and as such, using the standard Fourier identity

$$\mathcal{D}(x) \cdot f(x) = \sum_S \widehat{\mathcal{D} \circ f}(S) \chi_S(x) = \frac{|\mathcal{D}|}{|\mathcal{T}|^n} \sum_S \hat{f}(S).$$

■

2.1 In Distribution Testing and Related Notions

Unfortunately, we were not able to find a good way to define property testing, *soley* over the distribution. Rather, we will need to test both in-distribution and out-of-distribution samples. Still, we will only need to test out-of-distribution samples in a very limited way: samples which are only hamming distance 1 away from in-distribution samples for most of our properties.

To denote the need for in-distribution testing, we will append our notation with $\mathrm{In}\mathcal{D}$ to denote the in-distribution case.

Definition 2.1 *(Distribution-Testing Coordinate Averaging)*: Let $E^i_{\mathrm{In}\mathcal{D}}$ for $i \in n$ be the i -th in distribution operator for $x \in \mathcal{D}$:

$$E^i_{\mathrm{In}\mathcal{D}}[f](x) = \mathbb{E}_{a \in \mathcal{T}}[\mathcal{D}(x_i \mapsto a) \circ f(x^{x_i \mapsto a})].$$

Note that Definition 2.1 *zeros out* any coordinate setting which does not remain within the dataset. Intuitively, we can think about $E^i_{\mathrm{In}\mathcal{D}}$ as the generic coordinate averaging operator for a function which will test whether an input is in the dataset and output 0 if it is not.

We can now define influence:

Definition 2.2 *(i-th Coordinate Distribution-Testing Influence Operator)*:

$$\mathrm{Inf}^i_{\mathrm{In}\mathcal{D}} f = \mathbb{E}_{x \in \mathcal{D}}\left[\left(f(x) - E^i_{\mathrm{In}\mathcal{D}} f(x)\right)^2\right].$$

Proposition 2.1: We now prove that basic identities still hold as in O'Donnell's Analysis of Boolean Functions [1]:

$$\mathrm{Inf}^i_{\mathrm{In}\mathcal{D}} f = \frac{|\mathcal{D}|}{|\mathcal{T}|^n} \langle f, f - E^i_{\mathrm{In}\mathcal{D}} \rangle,$$

$$E^{i^i_{\mathrm{In}\mathcal{D}}}_{\mathrm{In}\mathcal{D}} f = \frac{|\mathcal{D}|}{|\mathcal{T}|^n} \sum_{S, S_i = 0} \hat{f}(S) \chi_S(x)$$

$$\mathrm{Inf}^i_{\mathrm{In}\mathcal{D}} f = \frac{|\mathcal{D}|}{|\mathcal{T}|^n} \sum_{S, S_i \neq 0} \hat{f}(S)^2.$$

Proof: We start with the second equality. TODO: check constants Note that,

$$
\begin{aligned}
E^{i i_{\mathrm{In}\mathcal{D}}}_{\mathrm{In}\mathcal{D}}(x) &= \mathbb{E}_{a\in\mathcal{T}}[\mathcal{D}(x_i \mapsto a)\circ f(x^{x_i\mapsto a})] \\
&= \sum_{S, S_i=0} \hat{f}_{\mathrm{OG}}(f)(S)\chi_S(x) \\
&= \frac{|\mathcal{D}|}{|\mathcal{T}|^n} \sum_{S,S_i=0} \hat{f}(S)\chi_S(x).
\end{aligned}
$$

The first equality holds as we note that,

$$
\begin{aligned}
\mathrm{Inf}^i_{\mathrm{In}\mathcal{D}} f &= \frac{|\mathcal{D}|}{|\mathcal{T}|^n}^{-1} \left\langle f(x) - E^i_{\mathrm{In}\mathcal{D}}(x), f(x) - E^i_{\mathrm{In}\mathcal{D}}(x)\right\rangle \\
&= \frac{|\mathcal{D}|}{|\mathcal{T}|^n}^{-1} \left(\mathbb{E}[f(x)^2] - 2E_x\left[f(x)E^i_{\mathrm{In}\mathcal{D}}(x)\right] + \mathbb{E}_{\mathcal{D}}\left[E^i_{\mathrm{In}\mathcal{D}}(x)^2\right]\right).
\end{aligned}
$$

TODO: norm constant Then, note that by the second equality and Parseval's theorem, $E_{x\sim\mathcal{T}^n}\left[f(x)E^i_{\mathrm{In}\mathcal{D}}(x)\right] = \sum_{S,S_i=0} \hat{f}_{\mathrm{OG}}(S)^2$ and $\mathbb{E}_{\mathcal{D}}\left[E^i_{\mathrm{In}\mathcal{D}}(x)^2\right] = \sum_{S,S_i=0} \hat{f}(S)^2$ as desired.

Finally, we note that as $\mathrm{Inf}^i_{\mathrm{In}\mathcal{D}} f = \mathbb{E}_{\mathcal{D}}[f(x)^2] - \mathbb{E}_{\mathcal{D}}\left[E^i_{\mathrm{In}\mathcal{D}}(x)^2\right]$, we can see that

$$
\mathbb{E}_{\mathcal{D}}[f(x)^2] - E_x\left[f(x)E^i_{\mathrm{In}\mathcal{D}}(x)\right] = \sum_S \hat{f}(S)^2 - \sum_{S,S_i=0} \hat{f}(S)^2 = \sum_{S,S_i\neq 0} \hat{f}(S)^2.
$$

■

Just as in TODO: cite , we can define total influence and get some convenient corollaries:

Definition 2.3 *(Total Influence, $\boldsymbol{I}_{\mathrm{In}\mathcal{D}}$)*:

$$
\boldsymbol{I}_{\mathrm{In}\mathcal{D}}[f] = \sum_{i\in[n]} \mathrm{Inf}^i_{\mathrm{In}\mathcal{D}}[f].
$$

We immediately get:

Proposition 2.2:

$$
\boldsymbol{I}_{\mathrm{In}\mathcal{D}}[f] = \sum_S \#S\ \hat{f}(S)^2
$$

which, for finite groups, we get

$$
\boldsymbol{I}_{\mathrm{In}\mathcal{D}}[f] = \sum_S k \cdot W^k[f].
$$

as in page 213 of TODO: a where $\#S = |\mathrm{supp}\ (S)| = \mathrm{supp}\ (S) = \{i : S_i \neq 0\}$.

Finally, we introduce one more definition which will capture the closeness of two functions, f, g over $\mathcal{D}$.

Definition 2.4 *(In-Distribution Closeness, $\epsilon_{\mathcal{D}}$-closeness)*: We say that a function g is $\epsilon_{\mathcal{D}}$ close to f if:

$$
\mathop{\mathbb{E}}_{x\in\mathcal{D}}\left[(f(x) - g(x))^2\right] \leq \epsilon
$$

or equivalently,

$$
\frac{|\mathcal{D}|}{|\mathcal{T}|^n}^{-1} \mathop{\mathbb{E}}_{x\in\mathcal{T}^{\otimes n}}\left[(\mathcal{D}\circ f(x) - \mathcal{D}\circ g(x))^2\right] \leq \epsilon.
$$

2.2 Immediate Consequences

Learning from Random Examples

We can already adopt learning low-degree functions from random examples to our setting. Specifically, if the in-distribution Fourier mass is concentrated on low-degree terms, we can learn the function from random examples drawn from $\mathcal{D}$.

TODO: this is auto-gened, check over and make right! CLAUDE CHECK

Theorem 2.1: Let $f : \mathcal{T}^n \to \mathbb{R}$ be a function such that $\sum_{S:\ |S|\ >d} \hat{f}(S)^2 \leq \frac{\epsilon^2}{4}$. Then, there exists an algorithm which, given $O\left(\frac{n^d}{\epsilon^2} \log(n)\right)$ random examples from $\mathcal{D}$, outputs a function g which is $\epsilon_{\mathcal{D}}$ close to f with probability at least $\frac{2}{3}$.

One-Way Property Testing for Computable via Decision Tree

2.3 In Distribution Testing Definitions

Surprisingly, if want to characterize the complexity (or sensitivity or one of many Fourier properties) of our function f over distribution $\mathcal{D}$ and in-distribution testing, we can more or less use standard analysis.

In more detail, we will model our modified function $g : |\mathcal{T}|^n \to {\mathbb{R}^+}^m \cup \{0\}$ as

$$g(x) = \begin{cases} f(x) & \text{if } \ x \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

And, taking the standard inner-product (TODO: we are over sphere/ real numbers, first take the inner-product then the other thing!)

Note that g is required to differentiate between in and out-of distribution.

Influence and Related Operators

We would like coordinate wise influence to be defined in the standard way, but adopted to our setting:

$$\mathrm{Inf}^i_{\mathcal{D}}[g] = \mathop{\mathbb{E}}_{x \in \mathcal{D}}\left[\left(g - E^i_{\mathcal{D}} g\right)^2\right]$$

where $E^i_{\mathcal{D}}$ is the i-th coordinate expectation operator:

$$\left[E^i_{\mathcal{D}} g\right](x) = \mathop{\mathbb{E}}_{x_i \in \mathcal{T}}\left[g\left(x^{i \mapsto x_i}\right)\right].$$

In words, the i-th expectation operator "averages out" the i-th coordinate over the dataset while the Influence measures the difference.

TODO: define distribution coeffs! TODO: This prob dist thing doesn't work??? Just have a subset thingy for now!!! Then, we have the following proposition:

Proposition 2.3:

$$E^i_{\mathcal{D}} g = \sum_{s : s_i \neq 0} p_{\text{scale}} \widehat{\;\cdot\;} f(s) \hat{f}_{\mathcal{D}}$$

Lemma 2.2: We can re-write the influence as an inner-product

$$\mathrm{Inf}^i_{\mathcal{D}}[f] = \left\langle p \cdot \left(g - E^i_{\mathcal{D}} g\right), g\right\rangle = \left\langle p \cdot \left(g - E^i_{\mathcal{D}} g\right), g - E^i_{\mathcal{D}} g\right\rangle$$

Proof: The second equality follows directly from the definition. The first is because

$$\langle p \cdot (g - E_{\mathcal{D}}^i g), g - E_{\mathcal{D}}^i g \rangle = \mathbb{E}_x[g^2] - 2\mathbb{E}_x[g \cdot E_{\mathcal{D}}^i f] + \mathbb{E}_x\left[\left(E_{\mathcal{D}}^i f\right)^2\right]$$

■

3 Conclusion and Future Directions

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet, ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius.

Acknowledgments

AI usage: the author would like to acknowledge the use of language models, Gemini and Claude, in generating the SVG…

Bibliography

[1] Ryan O'Donnell. 2021. Analysis of boolean functions. *arXiv preprint arXiv:2105.10386* (2021).

A Appendix

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet, ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam ex eo est consecutus? – Laudem et caritatem, quae sunt vitae sine metu degendae praesidia firmissima. – Filium morte multavit. – Si sine causa, nollem me ab eo delectari, quod ista Platonis, Aristoteli, Theophrasti orationis ornamenta neglexerit. Nam illud quidem physici, credere aliquid esse minimum, quod profecto numquam putavisset, si a Polyaeno, familiari suo, geometrica discere maluisset quam illum etiam ipsum dedocere. Sol Democrito magnus videtur, quippe homini erudito in geometriaque perfecto, huic pedalis fortasse; tantum enim esse omnino in nostris poetis aut inertissimae segnitiae est aut fastidii delicatissimi. Mihi quidem videtur, inermis ac nudus est. Tollit definitiones, nihil de dividendo ac partiendo docet, non quo ignorare vos arbitrer, sed ut.