Active Fourier Verifier: PAC Estimation of Model Properties with Influence Functions and Fourier Representations

Estimating all properties at once

Ayoub Ajarra

23 de noviembre de 2023

Joint work with

Bishwamittra Ghosh
National University of Singapore

Debabrota Basu
Équipe Scool, Univ. Lille, Inria

Outline

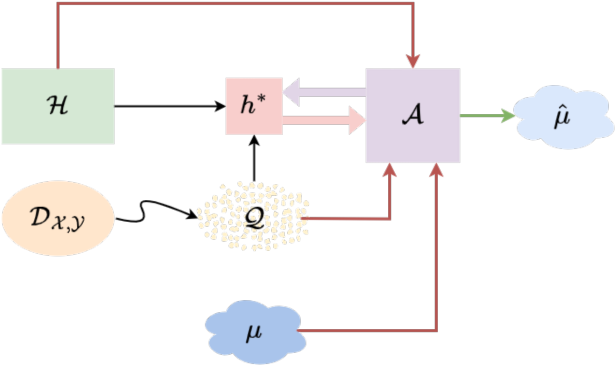
- Our problem: Given (restricted) access to a black box model h, we aim to estimate its properties: robustness and its discrimination toward sensitive subgroups and individuals.
- Our approach: In order to estimate those properties, we choose deterministic or random influence functions:
  - Random influence function for robustness and individual fairness
  - Deterministic influence function for group fairness

Problem definition

Property auditing from ZKP

Our problem setting

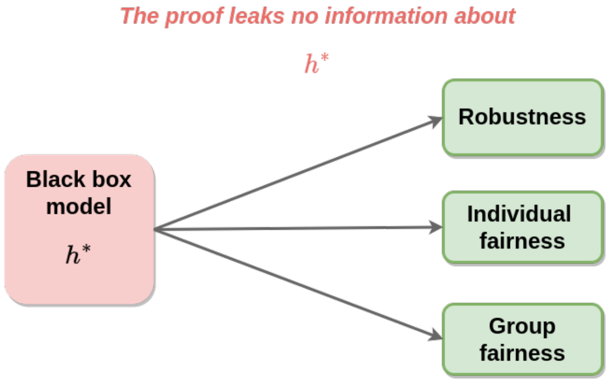
Property auditing: General setting



What's solutions are out there ?

- Reconstruct then audit. [Tom Y. 2022]
- Sequential testing [Ben.C et al.2023]
- Ours: Estimation by embedding the property of interest in the space of Fourier expandable models.

Zero-Knowledge Proof (ZKF) properties estimation: Unrealistic

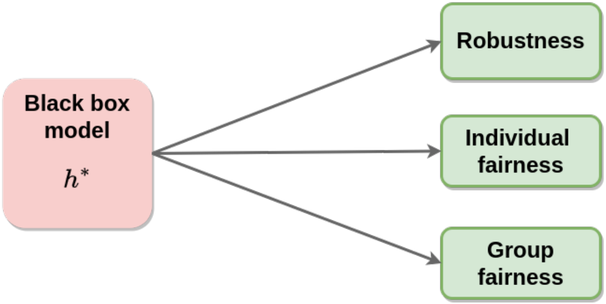


Estimating black-box model's properties in ZKP setting

Friendly Zero-Knowledge Proof (Friendly-ZKF) properties estimation: Our setting

The proof leaks only information of the form

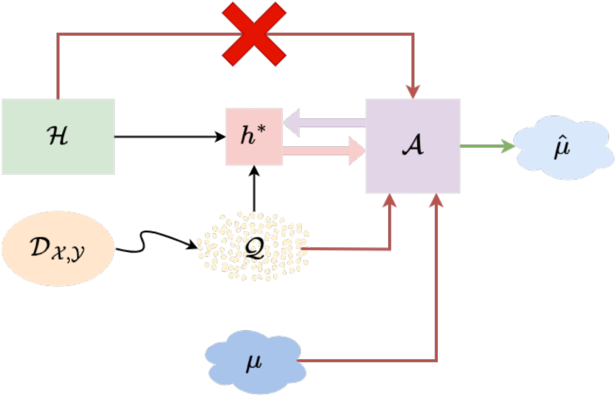
$$\{x, h^*(x)\}$$



Estimating black-box model's properties in friendly ZKP setting

Estimating model's properties from FZKP

Can we reduce the amount of information communicated to $\mathcal{A}$?



Boolean functions & Fourier expansion in the context of learning theory

Motivation

What do we know from Boolean function theory?

Theorem

Any bounded function $h : \{-1, 1\}^n \to \mathcal{Y}$ can be uniquely written as:

$$\forall (X_1, \ldots, X_n) \in \{-1, 1\}^n : h(X_1, \ldots, X_n) = \sum_{S \subseteq [n]} \hat{h}(s) \prod_{i \in S} X_i$$

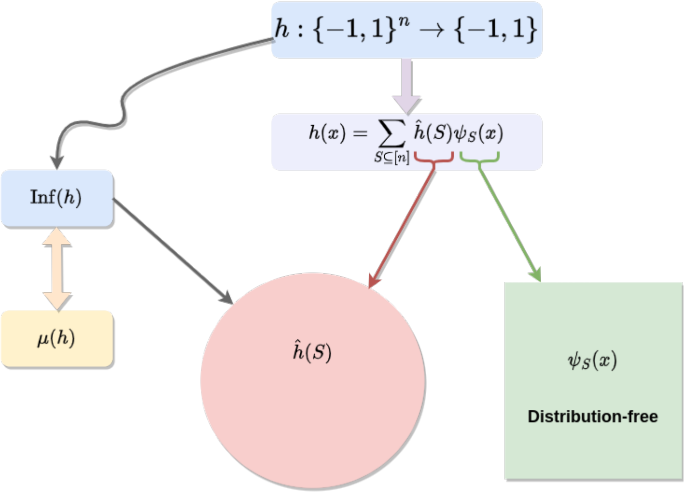
- Assumption 1: We assume that h is bounded. ✔
- Assumption 2: The distribution over binary values is uniform. ✗
- Assumption 3: The feature space is a matrix of binary values. ✗

Distribution free setting

Assumption 2:Generalizing to Agnostic Fourier expansion to the distribution

- Method 1: Instead of expanding the model on the basis of parity functions, we can use Gram-Schmidt-type process.
- Method 2: Instead of expanding the model h in the basis of parity functions, we expand the new model $\frac{\mathrm{h}(\mathrm{x})}{\mathcal{D}_{\mathcal{X}}(\mathrm{x})}$

Summary



$h : \{-1,1\}^n \to \{-1,1\}$
$h(x) = \sum_{S \subseteq [n]} \hat{h}(S)\psi_S(x)$
$\mathrm{Inf}(h)$
$\mu(h)$
$\hat{h}(S)$
$\psi_S(x)$
Distribution-free

Computing target model properties

Target model's properties in terms of Fourier coefficients

Influence functions

Perturbation protocol

$\mathrm{x} \sim \mathcal{D}_{\mathcal{X}},\ \rho \in [-1, 1],\ \mathrm{l} \in [\mathrm{n}]$

- $\mathrm{x}' \sim \mathcal{N}_\rho(\mathrm{x}) \iff \forall \mathrm{i} \in [\mathrm{n}] : \mathrm{x}'_\mathrm{i} = \begin{cases} \mathrm{x}_\mathrm{i} & \text{with probability } \frac{1+\rho}{2} \\ -\mathrm{x}_\mathrm{i} & \text{with probability } \frac{1-\rho}{2} \end{cases}$
- $\mathrm{x}' \sim \mathcal{N}_{\rho,\mathrm{l}}(\mathrm{x}) \iff \forall \mathrm{i} \in [\mathrm{n}]^\mathrm{l} : \mathrm{x}'_\mathrm{i} = \begin{cases} \mathrm{x}_\mathrm{i} & \text{with probability } \frac{1+\rho}{2} \\ -\mathrm{x}_\mathrm{i} & \text{with probability } \frac{1-\rho}{2} \end{cases}$

Influence functions

Deterministic & random influence functions

Random influence functions

$$\mathrm{Inf}_\rho(\mathrm{h}) \triangleq \mathop{\mathbb{P}}_{\substack{\mathrm{x}\sim\mathcal{D}_\mathcal{X} \\ \mathrm{x}'\sim\mathcal{N}_\rho}} [\mathrm{h}(\mathrm{x}) \neq \mathrm{h}(\mathrm{x}')] \text{ measures stability.}$$

$$\mathrm{Inf}_{\rho,1}(\mathrm{h}) \triangleq \mathop{\mathbb{P}}_{\substack{\mathrm{x}\sim\mathcal{D}_\mathcal{X} \\ \mathrm{x}'\sim\mathcal{N}_{\rho,1}}} [\mathrm{h}(\mathrm{x}) \neq \mathrm{h}(\mathrm{x}')] \text{ measure individual fairness.}$$

Deterministic influence functions

Let A be a sensitive attribute

$$\mathrm{Inf}_\mathrm{A}(\mathrm{h}) \triangleq \mathop{\mathbb{P}}_{\mathrm{x}\sim\mathcal{D}_\mathcal{X}} [\mathrm{h}(\mathrm{x}) \neq \mathrm{h}(\mathrm{x}^{\sim\mathrm{A}})]$$

$$\mu_{\mathrm{GF}_\mathrm{A}}(\mathrm{h}) = \mathop{\mathbb{P}}_{\mathrm{x}\sim\mathcal{D}}[\mathrm{h}(\mathrm{x}) = \mathrm{y}|\mathrm{x} \in \mathrm{A}^+] - \mathop{\mathbb{P}}_{\mathrm{x}\sim\mathcal{D}}[\mathrm{h}(\mathrm{x}) = \mathrm{y}|\mathrm{x} \in \mathrm{A}^-]$$

Model's properties in terms of its Fourier coefficients

Robustness, Individual fairness, Group fairness

Robustness

$$\mu_{\mathrm{RB}}(\mathrm{h}^*) = \sum_{\mathrm{S}\subseteq[\mathrm{n}]} \hat{\mathrm{h}^*}(\mathrm{S})^2 \rho^{|\mathrm{S}|}$$

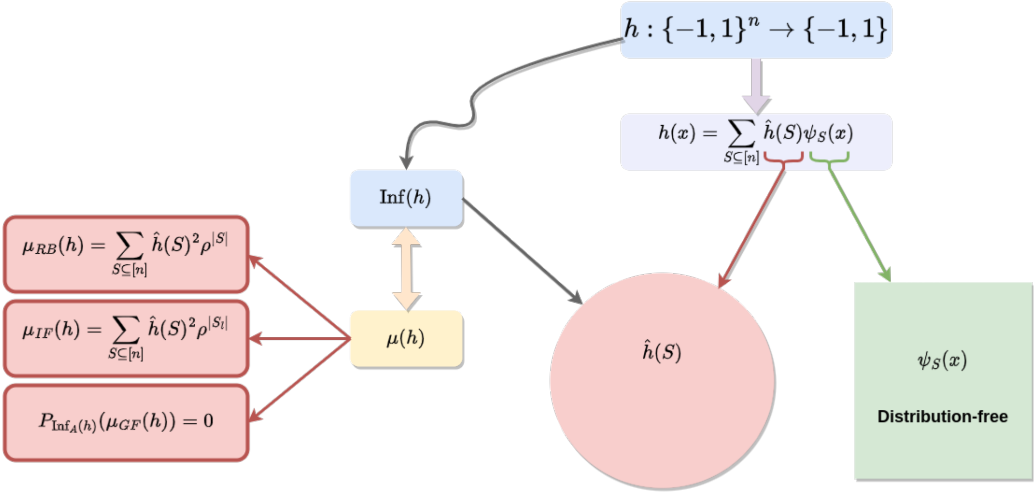
Individual Fairness

$$\mu_{\mathrm{IF}}(\mathrm{h}^*) = \sum_{\mathrm{S}\subseteq[\mathrm{n}]} \hat{\mathrm{h}^*}(\mathrm{S})^2 \rho^{|\mathrm{S}_\mathrm{I}|}$$

Group Fairness (Informal)

Given the assumption that the marginal distribution is invariant under the flip membership action, $\mathrm{Inf}_\mathrm{A}(\mathrm{h}^*)$ is polynomial in $\mu_{\mathrm{GF}}(\mathrm{h}^*)$.

Summary



$h : \{-1,1\}^n \to \{-1,1\}$
$h(x) = \sum_{S \subseteq [n]} \hat{h}(S)\psi_S(x)$
$\mathrm{Inf}(h)$
$\mu(h)$
$\mu_{RB}(h) = \sum_{S \subseteq [n]} \hat{h}(S)^2 \rho^{|S|}$
$\mu_{IF}(h) = \sum_{S \subseteq [n]} \hat{h}(S)^2 \rho^{|S_l|}$
$P_{\mathrm{Inf}_A(h)}(\mu_{GF}(h)) = 0$
$\hat{h}(S)$
$\psi_S(x)$
Distribution-free

Hardness of computing target model properties & Universal lower bounds

Computing target model's properties can be NP-Hard

Finding significant Fourier coefficients is Hard

Theorem

Given a threshold $\tau \in \mathbb{R}$, the problem of testing significant Fourier coefficients with respect to the threshold τ is NP-complete.

Low degree model

The degree of a boolean function is the degree of the polynomial of its Fourier representation.

Example

we can split every function $h : \{-1, 1\}^n \rightarrow \{-1, 1\}$ to its different spectral levels, $h = \sum_{i=1}^{n} h^{(i)}$ where $h^{(i)} = \sum_{\substack{S \subseteq [n] \\ |S| = i}} \hat{h}(S)\psi_S$

Lower bounds

Theorem

Let $\mathrm{d} \in \mathbb{N}$, $\mathrm{h} \in \mathcal{H}_\mathrm{d}$, where $\mathcal{H}_\mathrm{d}$ is the concept class of Boolean functions of degree at most d. In other words, h is ϵ-concentrated in some subset of size at most d. Algorithm $\mathcal{A}_\mu$ outputs an (ϵ, δ)-PAC estimate of $\mu(\mathrm{h})$ with

$$\Omega\left((1-\epsilon)2^{\mathrm{d}-2}\log_2 \mathrm{n} - (\mathrm{d}+1)2^{\mathrm{d}-2}\log_2(1-\delta)\right)$$

queries.

Upper bounds

PAC Fourier Auditor

Algorithm

Estimating all properties at once

Algorithm Property estimation in the Boolean domain

Input: Confidence parameter δ, target error ϵ, Sensitive attribute A, perturbation parameters: ρ, l, q queries

$\{x_k, h(x_k)\}_{k \in [q]} \leftarrow MQ(h, q)$

$L_h \leftarrow CFS(\{x_k, h(x_k)\}_{k \in [q]}, \tau)$

for $S \in L_h$ do

$\quad \hat{h}(S) \leftarrow \frac{1}{q} \sum_{k=1}^{q} h(x_k) \psi_S(x_k)$

end for

Output: $\{\hat{\mu}_{\mathrm{Rob}_\rho}, \hat{\mu}_{\mathrm{IF}}, \hat{\mu}_{\mathrm{GF}_A}\}$

Algorithm

Concentrated Fourier Spectral Estimator

Algorithm CFS

Input: $\tau, \delta \in (0,1)$

$\varepsilon \leftarrow \tau^2/4$

Instantiate set L

for $\mathrm{k} = 0, \ldots, \mathrm{n}$ do

  for $\mathrm{S} \subseteq [\mathrm{k}]$ do

    $\widetilde{\mathrm{W}}^{\mathrm{S,k}} \leftarrow \mathrm{W}^{\mathrm{S,k}}$ (estimate sum up to accuracy ε w.p. $\geq 1-\delta$)

    if $\widetilde{\mathrm{W}}^{\mathrm{S,k}} \leq \frac{\tau^2}{2}$ then

      Discard $\mathrm{W}^{\mathrm{S,k}}$

    end if

  end for

  Output: L (A list of single elements)

Upper bounds for robustness estimation

$$\mathcal{O}\left(\left(\frac{(\mathrm{C}+1)\rho^{\min(\mathrm{S}_{\mathcal{F}},\mathrm{S}_{\mathcal{F}})}|\mathcal{F}|}{\epsilon}\right)^2 \log\frac{1}{\delta}\right)$$

$$|\mathcal{F}| = \mathcal{O}\left(\frac{\mathrm{n}}{\tau^2}\right)$$

$$\mathrm{S}_{\mathcal{F}} = \min\{|\mathrm{S}| : \mathrm{S} \in \mathcal{F}\}$$

$$\mathrm{S}_{\overline{\mathcal{F}}} = \min\{|\mathrm{S}| : \mathrm{S} \notin \mathcal{F}\}$$

Upper bounds
Individual fairness

Upper bounds for Individual Fairness estimation

$$\mathcal{O}\left(\left(\frac{(\mathrm{C}+1)\rho^{\min(\mathrm{S}_{1,\mathcal{F}},\mathrm{S}_{1,\mathcal{F}})}|\mathcal{F}|}{\epsilon}\right)^2 \log\frac{1}{\delta}\right)$$

.

$$|\mathcal{F}| = \mathcal{O}\left(\frac{\mathrm{n}}{\tau^2}\right)$$

.

$$\mathrm{S}_{1,\mathcal{F}} = \min\{|\mathrm{S}_1| : \mathrm{S}_1 \in \mathcal{F}\}$$

$$\mathrm{S}_{1,\overline{\mathcal{F}}} = \min\{|\mathrm{S}_1| : \mathrm{S}_1 \notin \mathcal{F}\}$$

Upper bounds

Group fairness

Upper bounds for group fairness

$$\mathcal{O}\left(\left(\frac{(\mathrm{C}+1)|\mathcal{F}|+\frac{1}{4}}{\epsilon^2}\right)^2 \log\frac{1}{\delta}\right)$$

$$|\mathcal{F}| = \mathcal{O}\left(\frac{\mathrm{n}}{\tau^2}\right)$$

A comment on the adversarial reconstruction

Adversarial reconstruction

Our algorithm guarantees an adversarial reconstruction of the target model only on 2ϵ-concentrated region over the Fourier space.

Extension to categorical & continuous domains

Group characters & Lováz extension

Extension to general domain

Function extension

Categorical domain:

- $h(x) = \sum_{\zeta} \hat{h}(\zeta)\omega_p(\langle \zeta, x \rangle)$

Continuous domain:

- $h(x) = \sum_{S \subseteq [n]} \hat{h}(S) \min_{i \in S} x_i$

Extension to general domain

Computing properties

Categorical domain:

- Robustness:

$$(\frac{1}{1-p})^{n} \sum_{\zeta:\zeta_j \neq 0} |\hat{f}(\zeta)|^2 + \sum_{\zeta:\zeta_j = 0} |\hat{f}(\zeta)|^2$$

Continuous domain:

- Robustness

$$\frac{1-\rho}{2} \sum_{S \subseteq [n]} \hat{h}(S)^2 (1 + |S| \zeta \frac{1+\rho}{1-\rho})$$

Extension to general domain

Computing properties

Categorical domain:

- Robustness:

$$(\frac{1}{1-p})^n \sum_{\zeta:\zeta_j\neq 0} |\hat{f}(\zeta)|^2 + \sum_{\zeta:\zeta_j=0} |\hat{f}(\zeta)|^2$$

- Individual fairness:

$$\frac{1}{p}\sum_{\zeta}|\hat{f}(\zeta)|^2\cos(\frac{2\pi}{p}\sum_{\mathcal{T}\in\mathbb{F}_p^n}\langle\ \zeta, \mathcal{T}\rangle)$$

Continuous domain:

- Robustness

$$\frac{1-\rho}{2}\sum_{S\subseteq[n]}\hat{h}(S)^2(1+|S|\zeta\frac{1+\rho}{1-\rho})$$

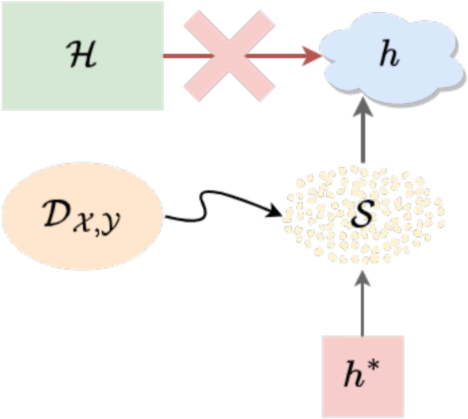
- Individual fairness

$$\frac{1-\rho}{2}\sum_{S\subseteq[n]}\hat{h}(S)^2(1+|S_l|\zeta\frac{1+\rho}{1-\rho})$$

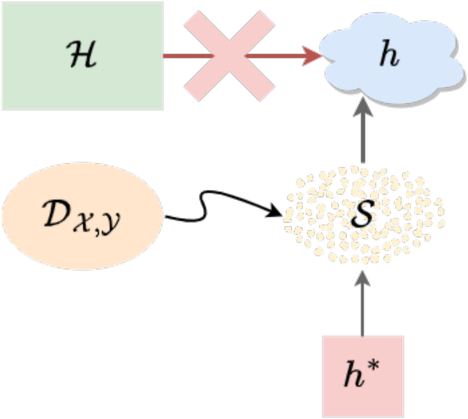
Open Problem

PAC learning with zero inductive bias

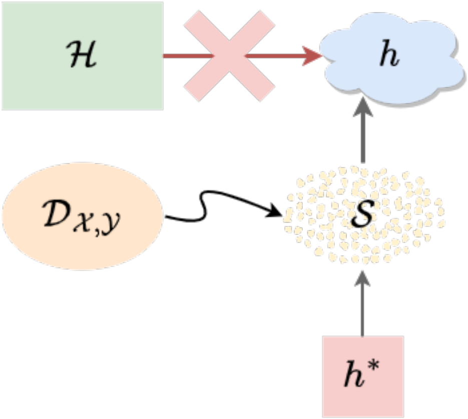
Can we learn with "no inductive bias"?



No-free lunch



Relaxation: Assumptions on the distribution



PAC-Fourier Auditor
Estimating all properties at once

Appendix

Ommited technical details

Gram-Schmidt orthogonalization

Assumption 2:Generalizing to Agnostic Fourier expansion to the distribution

For an unknown distribution $\mathcal{D}$, the set of parity functions are not necessarily orthogonal. Fix the following subsets of$\{1, \ldots, n\}$ in the following order: $\{\emptyset\}, \{1\}, \{2\}\{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \ldots, \{1,2,\ldots,n\}$

- Apply Gram-Schmidt process on the monomials of parity functions (to make them orthogonal) with the above ordering.
- The first element of the basis is trivially: $\psi_\emptyset = 1$
- The jth basis function corresponding to S_j is obtained from the following operation.

$$\tilde{\psi}_{S_j} = \chi_{S_j} - \sum_{l=1}^{j-1} \langle \psi_{S_l}, \chi_{S_j} \rangle_{\mathcal{D}} \psi_{S_l}$$

$$\psi_{S_j} = \begin{cases} \frac{\tilde{\psi}_{S_j}}{||\tilde{\psi}_{S_j}||_{2,\mathcal{D}}} & \text{if} ||\tilde{\psi}_{S_j}||_{2,\mathcal{D}} > 0, \\ 0 & \text{Otherwise} \end{cases}$$

Fourier expansion of $h(x)\mathcal{D}_{\mathcal{X}}(x)$

Assumption 2:Generalizing to Agnostic Fourier expansion to the distribution

- The new Fourier expansion is given by the formula :

$$\frac{h(x)}{\mathcal{D}_{\mathcal{X}}(x)} = \sum_{S \subseteq [n]} \hat{f}_{\mathcal{D}}(S)\chi_S(x)$$

- The Fourier coefficients are given by:

$$\forall S \subseteq [n] : \hat{f}(S) = \langle\, f, \chi_S \rangle$$

- All the results obtained from uniform distribution remain the same considering this approach. However, in practice we don't have access to $\mathcal{D}$, considering an empirical distribution could lead to unsatisfying results.

Extension to finite groups

Group character

Let G be a finite abelian group. We consider the general case: $G = \prod_{i=1}^{k} \mathbb{F}_{p_i}$

Group character

A map $\chi : G \to \mathbb{C} - \{0\}$ is called a character of G if it is a group homomorphism, that is:

$$\chi(0) = 1$$

$$\forall a, b \in G : \chi(a + b) = \chi(a) + \chi(b)$$

The constant map 1 is always a character for any abelian group, and is called the principal character of G.

Extension to finite groups

Construction of Fourier basis

Let G be a finite abelian group. We consider the general case: $G = \prod_{i=1}^{k} \mathbb{F}_{p_i}$

For every observation $a = (a_1, \ldots, a_k) \in G$, define $\chi_a \in \mathcal{L}_2(G)$:

$$\chi_a : x \to \prod_{j=1}^{k} e^{\frac{i2\pi}{p_j} a_j x_j}$$

Theorem

If G is a finite Abelian group, then the characters of G form an orthonormal basis for $\mathcal{L}_2(G)$. Furthermore, we have $G \cong \tilde{G}$.
$\tilde{G}$ is called the Pontryagin dual of G and it is the group of characters of G together with the usual point-wise product of complex-valued functions.

Extension to finite groups

Fourier expansion in finite domain as cartesian product of different cardinals

Let G be a finite abelian group. We consider the general case: $G = \prod_{i=1}^{k} \mathbb{F}_{p_i}$

For every observation $a = (a_1, \ldots, a_k) \in G$, define $\chi_a \in \mathcal{L}_2(G)$:

$$\chi_a : x \to \prod_{j=1}^{k} e^{\frac{i2\pi}{p_j} a_j x_j}$$

Theorem

The Fourier transform of a function $f : G \to \mathbb{C}$ is the unique function $\hat{f} : \hat{G} \to \mathbb{C}$ defined as $\hat{f}(a) =< f, \chi >= \mathbb{E}[f(x)\bar{\chi}(x)]$.
It follows from the fact that the characters from an orthonormal basis for $\mathcal{L}_2(G)$ that $f = \sum_{a \in G} \hat{f}(a)\chi_a$

Relaxation: Assumptions on the distribution

Partial answer

- Distribution dependent rates [Cohen, K., "Local Glivenko-Cantelli", COLT 2023].
- Bayes consistency, No rate [Steve Hanneke et al. Ümiversal Bayes consistency in metric spaces", AOS 2021.]

Estimating all properties at once!