Learning and Fourier Analysis

Grigory Yaroslavtsev

http://grigory.us

Penn
University of Pennsylvania

CIS 625: Computational Learning Theory

Fourier Analysis and Learning

- Powerful tool for PAC-style learning under **uniform distribution** over $\{0,1\}^n$
- Sometimes requires **queries** of the form $\boldsymbol{f}(x)$
- Works for learning many classes of functions, e.g:
  - Monotone, DNF, decision trees, low-degree polynomials
  - Small circuits, halfspaces, k-linear, juntas (depend on small # of variables)
  - Submodular functions (analog of convex/concave)
  - …
- Can be extended to **product** distributions over $\{0,1\}^n$, i.e. $D = D_1 \times D_2 \times \cdots \times D_n$ where $\times$ means that draws are independent

Boolean Functions

- **Book: "Analysis of Boolean Functions", Ryan O'Donnell**
  http://analysisofbooleanfunctions.org
- $\boldsymbol{f}(x_1, \ldots, x_n): \{0,1\}^n \to \mathbb{R}$
- Notation switch:
  - $0 \to 1$
  - $1 \to -1$
- $\boldsymbol{f}': \{-1,1\}^n \to \mathbb{R}$
- Example:
  - $\boldsymbol{f}(x_1, \ldots, x_n) = x_2 \oplus x_3 \oplus x_n = Lin(x_2, x_3, x_n)$
  - $\boldsymbol{f}'(x_1, \ldots, x_n) = x_2 \ * x_3 \ * x_n$

Fourier Expansion

- Example:

$$\boldsymbol{min}(x_1, x_2) = -\frac{1}{2} + \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_1x_2$$

- For $\boldsymbol{S} \subseteq [n]$ let **character** $\boldsymbol{\chi_S}(x) = \prod_{i \in \boldsymbol{S}} x_i$

- **Thm**: Every function $\boldsymbol{f}: \{-1,1\} \rightarrow \mathbb{R}$ can be **uniquely** represented as a multilinear polynomial

$$\boldsymbol{f}(x_1, \ldots, x_n) = \sum_{\boldsymbol{S} \subseteq [n]} \hat{\boldsymbol{f}}(\boldsymbol{S}) \boldsymbol{\chi_S}(x)$$

- $\hat{\boldsymbol{f}}(\boldsymbol{S}) \equiv$ Fourier coefficient of $\boldsymbol{f}$ on $\boldsymbol{S}$

Functions = Vectors, Inner Product

- Functions as vectors form a vector space:

$$\boldsymbol{f}: \{-1,1\}^n \to \mathbb{R} \Leftrightarrow \boldsymbol{f} \in \mathbb{R}^{2^n}$$

- Inner product on functions = "correlation":

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = 2^{-n} \sum_{x \in \{-1,1\}^n} \boldsymbol{f}(x)\boldsymbol{g}(x)$$

$$= \mathbb{E}_{x \sim \{-1,1\}^n}[\boldsymbol{f}(x)\boldsymbol{g}(x)]$$

- For $\boldsymbol{f}: \{-1,1\}^n \to \{-1,1\}$:
  $\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \Pr[\boldsymbol{f}(x) = \boldsymbol{g}(x)] - \Pr[\boldsymbol{f}(x) \neq \boldsymbol{g}(x)] = 1 - 2dist(\boldsymbol{f}, \boldsymbol{g})$

where $dist(\boldsymbol{f}, \boldsymbol{g}) = \Pr[\boldsymbol{f}(x) \neq \boldsymbol{g}(x)]$

Orthonormal Basis: Proof

- **Thm:** Characters form an orthonormal basis in $\mathbb{R}^{2^n}$

$$\langle \chi_S, \chi_T \rangle = 1, \qquad if\ \boldsymbol{S} = \boldsymbol{T}$$
$$\langle \chi_S, \chi_T \rangle = 0, \qquad if\ \boldsymbol{S} \neq \boldsymbol{T}$$

- $\chi_S(x)\chi_T(x) = \prod_{i \in S} x_i \prod_{j \in T} x_j = \prod_{i \in S \Delta T} x_i \prod_{j \in S \cap T} x_j^2 = \prod_{i \in S \Delta T} x_i = \chi_{S \Delta T}(x)$
- $\mathbb{E}_x[\chi_S(x)\chi_T(x)] = \mathbb{E}_x[\chi_{S \Delta T}(x)] = \mathbb{E}_x[\prod_{i \in S \Delta T} x_i] = \prod_{i \in S \Delta T} \mathbb{E}_x[x_i]$
- Since $\mathbb{E}_x[x_i] = 0$ we have:
  - $\mathbb{E}_x[\chi_S(x)\chi_T(x)]$ = 0 if $S \Delta T \neq \emptyset$
  - $\mathbb{E}_x[\chi_S(x)\chi_T(x)]$ = 1 if $S \Delta T = \emptyset$
- This proves that $\chi's$ are an orthonormal basis

Fourier Coefficients

- Recall linearity of dot products:
  - $\langle \boldsymbol{f} + \boldsymbol{g}, \boldsymbol{h} \rangle = \langle \boldsymbol{f}, \boldsymbol{h} \rangle + \langle \boldsymbol{g}, \boldsymbol{h} \rangle$
  - $\langle \alpha \boldsymbol{f}, \boldsymbol{g} \rangle = \alpha \langle \boldsymbol{f}, \boldsymbol{g} \rangle$
- **Lemma:** $\hat{\boldsymbol{f}}(\boldsymbol{S}) = \langle \boldsymbol{f}, \boldsymbol{\chi_S} \rangle$

$$\langle \boldsymbol{f}, \boldsymbol{\chi_S} \rangle = \langle \sum_{\boldsymbol{T} \subseteq [n]} \hat{\boldsymbol{f}}_{\boldsymbol{T}} \boldsymbol{\chi_T}, \boldsymbol{\chi_S} \rangle = \sum_{\boldsymbol{T} \subseteq [n]} \langle \hat{\boldsymbol{f}}_{\boldsymbol{T}} \boldsymbol{\chi_T}, \boldsymbol{\chi_S} \rangle =$$

$$= \sum_{\boldsymbol{T} \subseteq [n]} \hat{\boldsymbol{f}}_{\boldsymbol{T}} \langle \boldsymbol{\chi_T}, \boldsymbol{\chi_S} \rangle = \hat{\boldsymbol{f}}(\boldsymbol{S})$$

Parseval's Theorem

- **Parseval's Thm:** For any $\boldsymbol{f}: \{-1,1\}^n \to \mathbb{R}$

$$\langle \boldsymbol{f}, \boldsymbol{f} \rangle = \mathbb{E}_{x \sim \{-1,1\}^n}[\boldsymbol{f}^2(x)] = \sum_{\boldsymbol{S} \subseteq [n]} \hat{\boldsymbol{f}}(\boldsymbol{S})^{\mathbf{2}}$$

If $\boldsymbol{f}: \{-1,1\}^n \to \{-1,1\}$ then $\sum_{\boldsymbol{S} \subseteq [n]} \hat{\boldsymbol{f}}(\boldsymbol{S})^{\mathbf{2}} = 1$

- **Example:**

$$\boldsymbol{f} = \boldsymbol{min}(x_1, x_2) = -\frac{1}{2} + \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_1x_2$$

$$\sum_{\boldsymbol{S} \subseteq [2]} \hat{\boldsymbol{f}}(\boldsymbol{S})^{\mathbf{2}} = 4 \times \frac{1}{4} = 1$$

Plancharel's Theorem

- **Plancharel's Thm**: For any $\boldsymbol{f}, \boldsymbol{g}: \{-1,1\} \to \mathbb{R}$
$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \mathbb{E}_{x \sim \{-1,1\}^n}[\boldsymbol{f}(x)\boldsymbol{g}(x)] = \sum_{\boldsymbol{S} \subseteq [n]} \hat{\boldsymbol{f}}(\boldsymbol{S})\hat{\boldsymbol{g}}(\boldsymbol{S})$$
- **Proof:**
$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \langle \sum_{\boldsymbol{S} \subseteq [n]} \hat{\boldsymbol{f}}(\boldsymbol{S})\boldsymbol{\chi}_{\boldsymbol{S}}, \sum_{\boldsymbol{T} \subseteq [n]} \hat{\boldsymbol{g}}(\boldsymbol{T})\boldsymbol{\chi}_{\boldsymbol{T}} \rangle =$$
$$= \sum_{\boldsymbol{S},\boldsymbol{T} \subseteq [n]} \hat{\boldsymbol{f}}(\boldsymbol{S})\hat{\boldsymbol{g}}(\boldsymbol{T})\langle \boldsymbol{\chi}_{\boldsymbol{S}}, \boldsymbol{\chi}_{\boldsymbol{T}} \rangle = \sum_{\boldsymbol{S} \subseteq [n]} \hat{\boldsymbol{f}}(\boldsymbol{S})\, \hat{\boldsymbol{g}}(\boldsymbol{S})$$

Basic Fourier Analysis

- **Mean** $\equiv \mathbb{E}_x[\boldsymbol{f}(x)] = \langle \boldsymbol{f}, \mathbf{1} \rangle = \hat{\boldsymbol{f}}(\emptyset)$
  - For $\boldsymbol{f}: \{-1,1\}^n \to \{-1,1\}$ we have
    $\mathbb{E}_x[\boldsymbol{f}(x)] = \Pr[\boldsymbol{f}(x) = 1] - \Pr[\boldsymbol{f}(x) = -1]$
- **Variance** $\equiv \boldsymbol{Var}(f) \equiv \mathbb{E}_x[(\boldsymbol{f}(x) - \mathbb{E}_x[\boldsymbol{f}(x)])^2]$

$$\boldsymbol{Var}(f) = \mathbb{E}_x[(\boldsymbol{f}(x) - \mathbb{E}_x[\boldsymbol{f}(x)])^2]$$
$$= \mathbb{E}_x[\boldsymbol{f}^2(x)] - \mathbb{E}_x[\boldsymbol{f}(x)]^{\mathbf{2}}$$
$$= \sum_{\boldsymbol{S} \subseteq [\boldsymbol{n}]} \hat{\boldsymbol{f}}^2(\boldsymbol{S}) - \hat{\boldsymbol{f}}^2(\emptyset) \qquad \text{(Parseval's Thm)}$$
$$= \sum_{\boldsymbol{S} \subseteq [\boldsymbol{n}], \boldsymbol{S} \neq \emptyset} \hat{\boldsymbol{f}}^2(\boldsymbol{S})$$

Convolution

- **Def.:** For $x, y \in \{-1,1\}^n$ define $x \odot y \in \{-1,1\}^n$:
$$(x \odot y)_i = x_i y_i$$
  - **Def.:** For $\boldsymbol{f}, \boldsymbol{g}: \{-1,1\}^n \to \mathbb{R}$ their **convolution**
$$\boldsymbol{f} * \boldsymbol{g}: \{-1,1\}^n \to \mathbb{R}:$$
$$\boldsymbol{f} * \boldsymbol{g}(x) \equiv \mathbb{E}_{y \sim \{-1,1\}^n}[\boldsymbol{f}(y)\boldsymbol{g}(x \odot y)]$$

- Properties:
  1. $\boldsymbol{f} * \boldsymbol{g} = \boldsymbol{g} * \boldsymbol{f}$
  2. $\boldsymbol{f} * (\boldsymbol{g} * \boldsymbol{h}) = (\boldsymbol{f} * \boldsymbol{g}) * \boldsymbol{h}$
  3. For all $\boldsymbol{S} \subseteq [n]$: $\widehat{\boldsymbol{f} * \boldsymbol{g}}(\boldsymbol{S}) = \hat{\boldsymbol{f}}(\boldsymbol{S})\hat{\boldsymbol{g}}(\boldsymbol{S})$

Convolution: Proof of Property 3

- **Property 3:** For all $\boldsymbol{S} \subseteq [n]$: $\widehat{\boldsymbol{f} * \boldsymbol{g}}(\boldsymbol{S}) = \hat{\boldsymbol{f}}(\boldsymbol{S})\hat{\boldsymbol{g}}(\boldsymbol{S})$
- **Proof:**

$$\widehat{\boldsymbol{f} * \boldsymbol{g}}(\boldsymbol{S}) = \mathbb{E}_{x \sim \{-1,1\}^n}[(\boldsymbol{f} * \boldsymbol{g})(x)\boldsymbol{\chi}_{\boldsymbol{S}}(x)]$$

$$= \mathbb{E}_x\left[\mathbb{E}_y[\boldsymbol{f}(y)\boldsymbol{g}(x \oplus y)]\boldsymbol{\chi}_{\boldsymbol{S}}(x)\right]$$ (def. convolution)

$$= \mathbb{E}_y[\boldsymbol{f}(y)\mathbb{E}_x[\boldsymbol{g}(x \oplus y)\boldsymbol{\chi}_{\boldsymbol{S}}(x)]]$$

$$= \mathbb{E}_y[\boldsymbol{f}(y)\mathbb{E}_x[\boldsymbol{g}(x \oplus y)\boldsymbol{\chi}_{\boldsymbol{S}}(x \oplus y)\boldsymbol{\chi}_{\boldsymbol{S}}(x \oplus y)\boldsymbol{\chi}_{\boldsymbol{S}}(x)]]$$

$$= \mathbb{E}_y[\boldsymbol{f}(y)\boldsymbol{\chi}_{\boldsymbol{S}}(y)\mathbb{E}_x[\boldsymbol{g}(x \oplus y)\boldsymbol{\chi}_{\boldsymbol{S}}(x \oplus y)]]$$

$$= \mathbb{E}_y[\boldsymbol{f}(y)\boldsymbol{\chi}_{\boldsymbol{S}}(y)\hat{\boldsymbol{g}}(\boldsymbol{S})]$$

$$= \mathbb{E}_y[\boldsymbol{f}(y)\boldsymbol{\chi}_{\boldsymbol{S}}(y)]\hat{\boldsymbol{g}}(\boldsymbol{S})$$

$$= \hat{\boldsymbol{f}}(\boldsymbol{S})\hat{\boldsymbol{g}}(\boldsymbol{S})$$

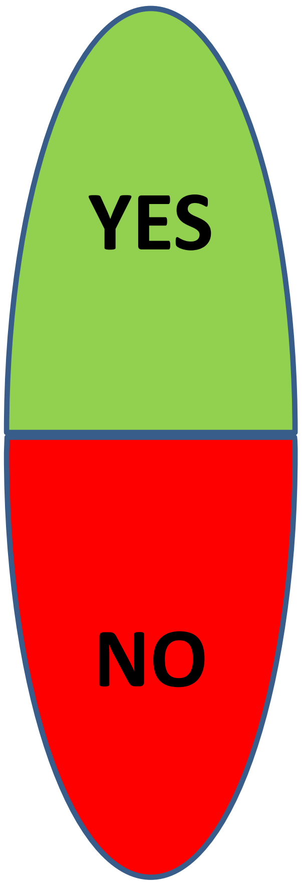
Approximate Linearity

- **Def:** $\boldsymbol{f}: \{0,1\}^n \to \{0,1\}$ is **linear** if
  - $\boldsymbol{f}(x \oplus y) = \boldsymbol{f}(x) \oplus \boldsymbol{f}(y)$ for all $x, y \in \{0,1\}^n$
  - $\exists \boldsymbol{S} \subseteq [\boldsymbol{n}]$ such that $\boldsymbol{f}(x) = \bigoplus_{i \in \boldsymbol{S}} x_i$ for all x
- **Def:** $\boldsymbol{f}: \{0,1\}^n \to \{0,1\}$ is **approximately linear** if
  1. $\boldsymbol{f}(x \oplus y) = \boldsymbol{f}(x) \oplus \boldsymbol{f}(y)$ for **most** $x, y \in \{0,1\}^n$
  2. $\exists \boldsymbol{S} \subseteq [\boldsymbol{n}]$ such that $\boldsymbol{f}(x) = \bigoplus_{i \in \boldsymbol{S}} x_i$ for **most** x
- **Q:** Does ***1.*** imply *2.*?

Property Testing

[Goldreich, Goldwasser, Ron; Rubinfeld, Sudan]

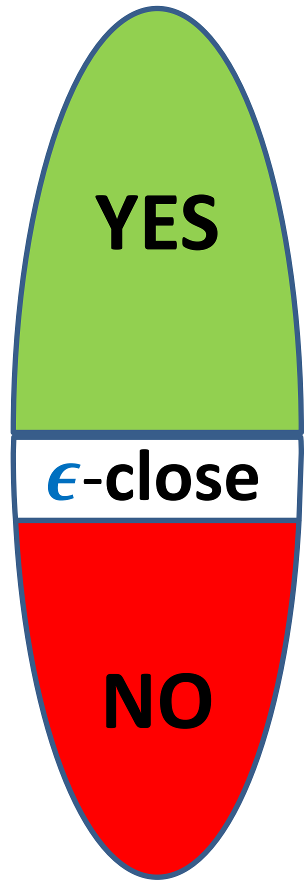
Randomized Algorithm



YES $\Rightarrow$ **Accept with probability** $\geq \frac{2}{3}$

NO $\Rightarrow$ **Reject with probability** $\geq \frac{2}{3}$

Property Tester



YES $\Rightarrow$ **Accept with probability** $\geq \frac{2}{3}$

ϵ-close $\Rightarrow$ **Don't care**

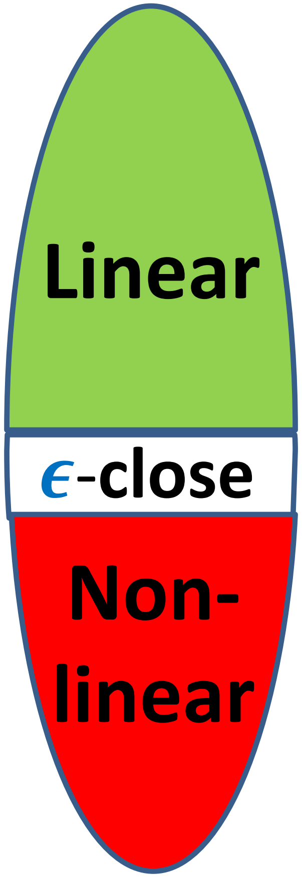
NO $\Rightarrow$ **Reject with probability** $\geq \frac{2}{3}$

ϵ-close : $\leq \epsilon$ fraction has to be changed to become **YES**

Linearity Testing

- $\boldsymbol{f}: \{0,1\}^n \rightarrow \{0,1\}$
- $\boldsymbol{P}$ = class of linear functions
- $dist(\boldsymbol{f}, \boldsymbol{P}) = \min_{\boldsymbol{g} \in \boldsymbol{P}} dist(f, g)$
- $\boldsymbol{\epsilon}$-close: $dist(\boldsymbol{f}, \boldsymbol{P}) \leq \boldsymbol{\epsilon}$

Linearity Tester



$\Rightarrow$ **Accept with probability $= \mathbf{1}$**

$\Rightarrow$ **Don't care**

$\Rightarrow$ **Reject with probability $\geq \frac{\mathbf{2}}{\mathbf{3}}$**

Linearity Testing [Blum, Luby, Rubinfeld]

- **BLR Test** (given access to $\boldsymbol{f}: \{0,1\}^n \to \{0,1\}$):
  - Choose $x \sim \{0,1\}^n$ and $y \sim \{0,1\}^n$ independently
  - Accept if $\boldsymbol{f}(x) \oplus \boldsymbol{f}(y) = \boldsymbol{f}(x \oplus y)$
- **Thm:** If BLR Test accepts with prob. $1 - \epsilon$ then $\boldsymbol{f}$ is ϵ-close to being linear
- **Proof**:
  - Apply notation switch $0 \to -1, -1 \to 1$
  - BLR accepts if $\boldsymbol{f}(x)\boldsymbol{f}(y) = \boldsymbol{f}(x \odot y)$
- $\frac{1}{2} + \frac{1}{2}\boldsymbol{f}(x)\boldsymbol{f}(y)\boldsymbol{f}(x \odot y) = 1$ if $\boldsymbol{f}(x)\boldsymbol{f}(y) = \boldsymbol{f}(x \odot y)$
- $\frac{1}{2} + \frac{1}{2}\boldsymbol{f}(x)\boldsymbol{f}(y)\boldsymbol{f}(x \odot y) = 0$ if $\boldsymbol{f}(x)\boldsymbol{f}(y) \neq \boldsymbol{f}(x \odot y)$

Linearity Testing: Analysis

$$1-\epsilon = \Pr[BLR\ Accepts\ \boldsymbol{f}]$$

$$= E_{x,y}\left[\frac{1}{2}+\frac{1}{2}\boldsymbol{f}(x)\boldsymbol{f}(y)\boldsymbol{f}(x\odot y)\right]$$

$$= \frac{1}{2}+\ \frac{1}{2}E_{x,y}[\boldsymbol{f}(x)\boldsymbol{f}(y)\boldsymbol{f}(x\odot y)]$$

$$= \frac{1}{2}+\ \frac{1}{2}E_x[\boldsymbol{f}(\boldsymbol{x})E_y[\boldsymbol{f}(y)\boldsymbol{f}(x\odot y)]]$$

$= \frac{1}{2}+\frac{1}{2}E_x[\boldsymbol{f}(\boldsymbol{x})\cdot(\boldsymbol{f}*\boldsymbol{f})(\boldsymbol{x})]$ (def. of convolution)

$= \frac{1}{2}+\frac{1}{2}\sum_{\boldsymbol{S}\subseteq[n]}\hat{\boldsymbol{f}}(\boldsymbol{S})\,(\widehat{\boldsymbol{f}*\boldsymbol{f}})(\boldsymbol{S})$ (Plancharel's thm)

$$= \frac{1}{2}+\frac{1}{2}\sum_{\boldsymbol{S}\subseteq[n]}\hat{\boldsymbol{f}}^3(\boldsymbol{S})$$

Linearity Testing: Analysis Continued

- $1 - 2\epsilon = \sum_{S \subseteq [n]} \hat{f}^3(S)$

$$\leq \max_{S \subseteq [n]} \{\hat{f}(S)\} \sum_{S \subseteq [n]} \hat{f}^2(S)$$

$$= \max_{S \subseteq [n]} \{\hat{f}(S)\}$$

- Recall $\hat{f}(S) = \langle f, \chi_S \rangle = 1 - 2dist(f, \chi_S)$
- Let $S^* = arg\max_{S \subseteq [n]} \{\hat{f}(S)\}$: $1 - 2\epsilon \leq 1 - 2dist(f, \chi_{S^*})$
- $dist(f, \chi_{S^*}) \leq \epsilon$

Local Correction

- Learning linear functions takes $\boldsymbol{n}$ queries
- **Lem:** If f is $\boldsymbol{\epsilon}$-close to linear function $\boldsymbol{\chi_S}$ then for every x one can compute $\boldsymbol{\chi_S}(x)$ w.p. $1-2\boldsymbol{\epsilon}$ as:
  - Pick $y \sim \{0,1\}^n$
  - Output $\boldsymbol{f}(y) \oplus \boldsymbol{f}(x \oplus y)$
- **Proof:**

$$Pr[\boldsymbol{f}(y) \neq \boldsymbol{\chi_S}(y)] = Pr[\boldsymbol{f}(x \oplus y) \neq \boldsymbol{\chi_S}(x \oplus y)] = \boldsymbol{\epsilon}$$

By union bound:

$$Pr[\boldsymbol{f}(y) = \boldsymbol{\chi_S}(y), \boldsymbol{f}(x \oplus y) = \boldsymbol{\chi_S}(x \oplus y)] \geq 1 - 2\boldsymbol{\epsilon}$$

Then $\boldsymbol{f}(y) \oplus \boldsymbol{f}(x \oplus y) = \boldsymbol{\chi_S}(y) \oplus \boldsymbol{\chi_S}(x \oplus y) = \boldsymbol{\chi_S}(x)$

Thanks!