PAC Learning [Valiant 1984]

$x_1, \ldots, x_m$ from $D$ over $\{-1, 1\}^n$
PAC Learning [Valiant 1984]

\[ x_1, \ldots, x_m \text{ from } D \text{ over } \{-1, 1\}^n \]

Labels \( y_i = f(x_i) \)

\( f \) in some class \( \mathcal{F} \), e.g. linear separators, DNF
\( x_1, \ldots, x_m \) from \( D \) over \( \{-1, 1\}^n \)

Labels \( y_i = f(x_i) \)

\( f \) in some class \( \mathcal{F} \), e.g. linear separators, DNF

Goal: Find hypothesis:
\[ h : \{-1, 1\}^n \rightarrow \{-1, 1\}, \text{ s.t.} \]
\[ \text{err}(h) = \mathbb{P}_{x \sim D}[h(x) \neq f(x)] \leq \epsilon \]
PAC Learning [Valiant 1984]

\(\mathbf{x}_1, \ldots, \mathbf{x}_m\) from \(D\) over \(\{-1, 1\}^n\)

Labels \(y_i = f(\mathbf{x}_i)\)

\(f\) in some class \(\mathcal{F}\), e.g. linear separators, DNF

Goal: Find hypothesis:
\(h : \{-1, 1\}^n \rightarrow \{-1, 1\}\), s.t.
\[
\text{err}(h) = \mathbb{P}_{\mathbf{x} \sim D}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon
\]

Want learning algorithm to succeed for all distributions \(D\)
Agnostic Learning [Vapnik '70s, Haussler '92, Kearns, Schapire, Sellie '94]

Generalization of Valiant’s PAC framework
Agnostic Learning [Vapnik ’70s, Haussler ’92, Kearns, Schapire, Sellie ’94]

Generalization of Valiant’s PAC framework

$$(x_1, y_1), \ldots, (x_m, y_m) \text{ from } D \text{ over } \{-1, 1\}^n \times \{-1, 1\}$$
Agnostic Learning [Vapnik '70s, Haussler '92, Kearns, Schapire, Sellie '94]

Generalization of Valiant’s PAC framework

\[(x_1, y_1), \ldots, (x_m, y_m)\] from \(D\) over \(\{-1, 1\}^n \times \{-1, 1\}\)

For some class \(\mathcal{F}\),

\[
\text{opt} = \min_{f \in \mathcal{F}} P_{(x,y) \sim D}[f(x) \neq y]
\]
Agnostic Learning [Vapnik '70s, Haussler '92, Kearns, Schapire, Sellie '94]

Generalization of Valiant’s PAC framework

$((x_1, y_1), \ldots, (x_m, y_m))$ from $D$ over $\{-1, 1\}^n \times \{-1, 1\}$

For some class $\mathcal{F}$,

$$\text{opt} = \min_{f \in \mathcal{F}} \mathbb{P}_{(x,y) \sim D}[f(x) \neq y]$$

**Goal:** find $h : \{-1, 1\}^n \to \{-1, 1\}$, s.t.

$$\text{err}(h) \leq \text{opt} + \epsilon$$

Want learning algorithm to succeed for **all distributions** $D$
Challenge

- Focus on zero-one loss, or classification error
- (Typically) ERM is computationally intractable
- Focus on computational and statistical efficiency
Uniform Distribution Learning

Breakthrough: Learning $AC^0$ circuits [LMN '91]

Query Model: Decision Trees [KM '93], DNF [Jackson '95]

Many Others: [BMOS '03], [KKMS '05], [OS '06], [GKK '08], [JW '09] etc.

Links to analysis of boolean functions, cryptography, complexity

Use sophisticated results: hypercontractivity, invariance, etc.

Main Tool: Discrete Fourier Analysis

$$f(x) = \sum_{S \subseteq \{1, \ldots, n\}} \hat{f}(S) \chi_S(x); \quad \chi_S(x) = \prod_{i \in S} x_i$$
Uniform Distribution Learning

<table>
<thead>
<tr>
<th>Breakthrough: Learning $AC^0$ circuits [LMN ’91]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Model: Decision Trees [KM ’93], DNF [Jackson ’95]</td>
</tr>
<tr>
<td>Many Others: [BMOS ’03], [KKMS ’05], [OS ’06], [GKK ’08], [JW ’09] etc.</td>
</tr>
</tbody>
</table>

| Links to analysis of boolean functions, cryptography, complexity |
| Use sophisticated results: hypercontractivity, invariance, etc. |
Uniform Distribution Learning

Breakthrough: Learning $AC^0$ circuits [LMN ’91]
Query Model: Decision Trees [KM ’93], DNF [Jackson ’95]
Many Others: [BMOS ’03], [KKMS ’05], [OS ’06], [GKK ’08], [JW ’09] etc.

Links to analysis of boolean functions, cryptography, complexity
Use sophisticated results: hypercontractivity, invariance, etc.

Main Tool: Discrete Fourier Analysis

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x); \quad \chi_S(x) = \prod_{i \in S} x_i$$
But . . .

Uniform distribution: too much of an idealization to be practically relevant

In “real-life distributions”, variables are correlated
But . . .

Uniform distribution: too much of an idealization to be practically relevant

In “real-life distributions”, variables are correlated

Can we salvage some of the mathematical theory for more general distributions?
But ...

Uniform distribution: too much of an idealization to be practically relevant

In “real-life distributions”, variables are correlated

Can we salvage some of the mathematical theory for more general distributions?

Yes – we present the first steps ...
Markov Random Fields

- Graph $G = ([n], E)$.
- Each node takes value in finite set $A$ (say $\{-1, 1\}$).
- Distribution over $A^n$: for $\phi_C$ non-negative, $Z$ normalization constant
  \[
  \mathbb{P}((\sigma_v)_{v \in [n]}) = \frac{1}{Z} \prod_{\text{clique } c} \phi_C((\sigma_v)_{v \in c})
  \]
- Uniform distribution: no edges
Markov Random Fields

MRFs widely used in vision, computational biology, biostatistics etc.

Extensive Algorithmic Theory for sampling from MRFs, recovering parameters and structures

Question: For unknown target function $f: \mathbb{R} \rightarrow \{-1, 1\}$:
- How can we learn with respect to MRF distribution?
- Can we utilize the structure of the MRF to aid in learning?
Markov Random Fields

MRFs widely used in vision, computational biology, biostatistics etc.

Extensive Algorithmic Theory for sampling from MRFs, recovering parameters and structures

Learning Question: For unknown target function $f : A^n \rightarrow \{-1, 1\}$

- (How) Can we learn with respect to MRF distribution?
- Can we utilize the structure of the MRF to aid in learning?
Learning Model (PAC/Agnostic)

- $M$ MRF with distribution $\pi$; $f : A^n \rightarrow \{-1, 1\}$ target function
- Access to i.i.d. examples $(x, f(x))$ where $x \sim \pi$
- Learning algorithm “knows” MRF
Gibbs Sampling (MCMC Algorithm)

Gibbs Sampler

Input: \( x^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)}) \in A^n \)

For \( t = 1, 2, 3, \ldots \)

(a) Pick \( i \in [n] \) uniformly at random

(b) Pick \( x_i^{(t)} \sim p(x_i \mid x_1^{(t-1)}, \ldots, x_{i-1}^{(t-1)}, x_{i+1}^{(t-1)}, \ldots, x_n^{(t-1)}) \)

(c) Set \( x_j^{(t+1)} = x_j^{(t)} \) for \( j \neq i \).

- Stationary distribution is MRF distribution
- For constant degree MRF graphs, conditional distribution has constant number of parameters
- We are interested in cases when Gibbs MC is rapidly mixing
Let $G = ([n], E)$ be some degree-$\Delta$ graph

For each $(i, j) \in E$, $\beta_{ij}$ (bounded) interaction energy

Configuration $\sigma \in \{-1, 1\}^n$; Hamiltonian

$$H(\sigma) = - \sum_{(i,j) \in E} \beta_{ij} \sigma_i \sigma_j - B \sum_{i \in [n]} \sigma_i$$

Probability distribution: $p(\sigma) \propto \exp(-H(\sigma))$

If $0 \leq \beta_{ij} \leq \beta(\Delta)$, Gibbs MC is rapidly mixing
Harmonic Analysis Using Eigenvectors

Let $\Omega = A^n$ be the statespace (MRF graph $G = ([n], E)$)

Gibbs Markov Chain over $\Omega$ is reversible
  ► $P$ transition matrix; $\pi$ stationary distribution
  ► Reversibility: $\pi_i P_{ij} = \pi_j P_{ji}$

An eigenvector of $P$ is a function $\nu: \Omega \to \mathbb{R}$

Eigenvectors form orthogonal basis with respect to stationary distribution $\pi$

Can we perform Fourier analysis using this basis?

Parity functions are eigenvectors for the uniform distribution
Harmonic Analysis Using Eigenvectors

Let $\Omega = A^n$ be the statespace (MRF graph $G = ([n], E)$)

Gibbs Markov Chain over $\Omega$ is reversible

- $P$ transition matrix; $\pi$ stationary distribution
- Reversibility: $\pi_i P_{ij} = \pi_j P_{ji}$

An eigenvector of $P$ is a function $\nu : \Omega \to \mathbb{R}$

Eigenvectors form orthogonal basis with respect to stationary distribution $\pi$

Can we perform Fourier analysis using this basis?
Harmonic Analysis Using Eigenvectors

Let $\Omega = A^n$ be the statespace (MRF graph $G = ([n], E)$)

Gibbs Markov Chain over $\Omega$ is reversible

- $P$ transition matrix; $\pi$ stationary distribution

- Reversibility: $\pi_i P_{ij} = \pi_j P_{ji}$

An eigenvector of $P$ is a function $\nu : \Omega \rightarrow \mathbb{R}$

Eigenvectors form orthogonal basis with respect to stationary distribution $\pi$

Can we perform Fourier analysis using this basis?

Parity functions are eigenvectors for the uniform distribution
The approach seems naïve:

- Each eigenvector is of size $|A|^n$
- How do we find these eigenvectors?
- How do we find the expansion of an arbitrary function using eigenvectors?
Power Iteration to Find Eigenvectors

Let $g : \Omega \to [-1, 1]$ be some function:

$$g = \alpha_1 \nu_1 + \alpha_2 \nu_2 + \cdots + \alpha_k \nu_k + \cdots.$$

$\nu_i$ is eigenvector with eigenvalue $\lambda_i$ and $\lambda_1 > \lambda_2 > \cdots$

$$P^t g = \alpha_1 \lambda_1^t \nu_1 + \alpha_2 \lambda_2^t \nu_2 + \cdots$$

$$\mathbb{1}_x^t P^t g = \alpha_1 \lambda_1^t \nu_1(x) + \alpha_2 \lambda_2^t \nu_2(x) + \cdots$$

$$\alpha_1^{-1} \lambda_1^{-t} \mathbb{1}_x^t P^t g = \nu_1(x) + \alpha_1^{-1} \alpha_2^{-1} (\lambda_1^{-1} \lambda_2)^t \nu_2(x) + \cdots$$
So, we have

$$1^\dagger_x P^t g = \alpha_1 \lambda_1^t \nu_1(x) + \alpha_2 \lambda_2^t \nu_2(x) + \cdots$$

$1^\dagger_x P^t$ distribution after running Gibbs MC for $t$ steps starting from $x$

$$1^\dagger_x P^t g = \mathbb{E}_{x' \sim 1^\dagger_x P^t} [g(x')]$$

Value of $\nu_1(x)$ can be estimated by sampling from Gibbs MC

In fact $\nu_i$ can be written as linear combinations of $P^t_1 g, P^t_2 g, \ldots$
When can we do this?

- Want eigenvalues of $P$ to show sharp drops
  - Discrete Spectrum
  - Weakly-correlated variables, high-temperature regimes
When can we do this?

- Want eigenvalues of $P$ to show sharp drops
  - Discrete Spectrum
  - Weakly-correlated variables, high-temperature regimes

- Want useful functions $g$ for extracting eigenvector
  - If target class has “low-degree” Fourier expansion
  - Expect such a basis of functions may exist
Discrete Spectrum

\[ \beta = 0.00 \quad \beta = 0.02 \]

\[ \beta = 0.1 \quad \beta = 1.00 \]

Spectra of Ising model on cycle for various temperatures
**Learning Algorithm**

<table>
<thead>
<tr>
<th>Learning Algorithm</th>
</tr>
</thead>
</table>

**Input:** Gibbs MC, Basis: $g_1, \ldots, g_m$, $T$, Sample: $\langle (x^i, y^i) \rangle_{i=1}^n$

$$\text{minimize}_{w_{t,j}} \sum_{i=1}^n \left| \sum_{t \leq T,j} w_{t,j} P^t g_j(x^i) - y^i \right|$$

subject to $\sum_{t,j} |w_{t,j}| \leq W$

**Output:**
- $h(x) = \sum_{t,j} w_{t,j} P^t g_j(x^i)$
- Predict $\text{sign}(h(x) - \theta)$ for $\theta \in [-1, 1]$ at random
Main Results

Informal Theorem

If a function class $\mathcal{F}$ is approximated by eigenvectors with high-eigenvalue ("low-degree") and

1. Spectrum of $\mathbf{P}$ is "discrete"
2. Have access to useful basis functions to extract eigenvectors

Can agnostically learn $\mathcal{F}$
Main Results

<table>
<thead>
<tr>
<th>Informal Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>If a function class $\mathcal{F}$ is approximated by eigenvectors with high-eigenvalue (&quot;low-degree&quot;) and</td>
</tr>
<tr>
<td>1. Spectrum of P is “discrete”</td>
</tr>
<tr>
<td>2. Have access to useful basis functions to extract eigenvectors</td>
</tr>
<tr>
<td>Can agnostically learn $\mathcal{F}$</td>
</tr>
</tbody>
</table>

- Uniform Case: Recover LMN’s low degree algorithm if $g_i$ are short conjunctions, disjunctions, etc.
Main Results

Informal Theorem

If a function class $\mathcal{F}$ is approximated by eigenvectors with high-eigenvalue ("low-degree") and

1. Spectrum of $P$ is “discrete”
2. Have access to useful basis functions to extract eigenvectors

Can agnostically learn $\mathcal{F}$

- Uniform Case: Recover LMN’s low degree algorithm if $g_i$ are short conjunctions, disjunctions, etc.
- High-eigenvalue eigenvectors easier to access
- Correspond to stable functions
- Generalize the notion of noise-sensitivity
Open Questions

- For simple MRFs and function classes $\mathcal{F}$, can we show that $\mathcal{F}$ is well-approximated by higher eigenvectors?
  - Yes, for high-temperature Ising models and linear separators
  - Uses generalized notion of noise-sensitivity

- Can access to a labelled random walk from Gibbs MC help?

- Under some conditions can learn $k$-juntas

- Is rapid mixing of Gibbs MC enough for learning $k$-juntas?

- Can we understand which Markov chains would have discrete spectra?
Open Questions

- For simple MRFs and function classes $\mathcal{F}$, can we show that $\mathcal{F}$ is well-approximated by higher eigenvectors?
  
  - Yes, for high-temperature Ising models and linear separators
  
  - Uses generalized notion of noise-sensitivity

- Can access to a labelled random walk from Gibbs MC help?

  \[(x^0, f(x^0)), (x^1, f(x^1)), \ldots,\]

  - Under some conditions can learn $k$-juntas
  
  - Is rapid mixing of Gibbs MC enough for learning $k$-juntas?
Open Questions

► For simple MRFs and function classes $\mathcal{F}$, can we show that $\mathcal{F}$ is well-approximated by higher eigenvectors?
  
  ● Yes, for high-temperature Ising models and linear separators
  
  ● Uses generalized notion of noise-sensitivity

► Can access to a labelled random walk from Gibbs MC help?

  $$(x^0, f(x^0)), (x^1, f(x^1)), \cdots,$$

  ● Under some conditions can learn $k$-juntas
  
  ● Is rapid mixing of Gibbs MC enough for learning $k$-juntas?

► Can we understand which Markov chains would have discrete spectra?