This Talk

Algorithmic Framework for Distribution Estimation: Leads to *fast & robust* estimators for a *wide variety* of statistical models.

[Chan-D-Servedio-Sun, STOC’14]
[Chan-D-Servedio-Sun, NIPS’14]
[Acharya-D-Hegde-Li-Schmidt, PODS’15]

[Acharya-D-Li-Schmidt, manuscript’15]

**Key Idea:**

Exploit piecewise polynomial approximation for structured model estimation
This Talk

A family of optimal estimators for hypothesis testing for a wide variety of structured models.

“Given samples from a statistical model does it satisfy a given property?”

[Daskalakis-D-Servedio-Valiant-Valiant, SODA’13]
[Chan-D-Valiant-Valiant, SODA’14]
[D-Kane-Nikishkin, SODA’15]

[D-Kane-Nikishkin, manuscript’15]
[Canonne-D-Gouleakis-Rubinfeld, manuscript’15]
Main Message of the Talk

We can algorithmically exploit the underlying structure to perform statistical estimation efficiently.
Outline

• Learning via Piecewise Polynomial Approximation
  ▪ Introduction
  ▪ Framework Overview
  ▪ Statistical Efficiency
  ▪ Computational Efficiency
  ▪ Empirical Results

• Applications to other Inference Tasks

• Future Directions and Concluding Remarks
Outline

• Learning via Piecewise Polynomial Approximation
  ▪ Introduction
  ▪ Framework Overview
  ▪ Statistical Efficiency
  ▪ Computational Efficiency
  ▪ Empirical Results

• Applications to other Inference Tasks

• Future Directions and Concluding Remarks
Distribution Learning (Density Estimation)

Given samples (observations) from an unknown probability distribution (model), construct an accurate estimate of the distribution.

- Classical Problem in Statistics
- Introduced by Karl Pearson (1891).
- Last fifteen years (TCS): computational aspects
Distribution Learning: History

- Histograms [Pearson, 1895]
- Kernel methods [M. Rosenblatt, 1956]
- Metric Entropy [A.N. Kolmogorov, 1960]
- Wavelets
  [Donoho, Johnstone, Kerkyacharian, Picard, ’90’s]

Many others: Nearest Neighbor, Orthogonal Series, …

Focus traditionally on sample size.
Types of Structured Distributions

• Distributions with “shape restrictions”

• Simple combinations of simple distributions

  Mixtures of simple distributions

  Sums of simple distributions

  Poisson Binomial Distributions
History

Nonparametric Estimation under “shape restrictions”

• Long line of work in statistics since the 1950’s
  [Gre’56, Rao69, Weg70, Gro85, Bir87,…]

• Shape restrictions studied in early work: monotonicity, unimodality, concavity, convexity, Lipschitz continuity, …

• Very active research area: log-concavity, $k$-monotonicity, …

• Standard tool in these settings: MLE
Distribution Learning: Definition

- Learning problem defined by family $\mathcal{D}$ of distributions
- Target distribution $p \in \mathcal{D}$ unknown to learner.
- Learner given sample of IID draws from $p$.

Output: with probability $\geq 9/10$ output $h$ satisfying

$$\|h - p\|_1 \leq \epsilon.$$ 

Goal: Sample optimal & computationally efficient algorithms
Agnostic Learning: Definition

- Learning problem defined by class $\mathcal{D}$ of distributions
- Target distribution $p$ unknown to learner and let
  \[ \text{OPT} = \inf_{q \in \mathcal{D}} \|p - q\|_1 \]
- Learner given sample of IID draws from $p$

**Output:** with probability $\geq 9/10$ output $h$ satisfying
\[ \|h - p\|_1 \leq \text{OPT} + \epsilon. \]

**Goal:** Sample optimal & computationally efficient algorithms
Learning Arbitrary Discrete Distributions

Let $\mathcal{D}$ be the set of all distributions over $[N]$.

What is the best learning algorithm?

Simple answer (folklore):

• Algorithm with sample (and time) complexity $O(N/\epsilon^2)$.
• Information theoretic lower bound of $\Omega(N/\epsilon^2)$. 
Learning Arbitrary Discrete Distributions

Learning an *arbitrary* distribution over $[N]$:
Sample size $\Theta(N/\epsilon^2)$
necessary and sufficient

When can we do better?
Which distributions are easy to learn, which are hard (and why)?
General Recipe for Statistical Estimation:

Given a “complex” distribution family $\mathcal{D}$.

1. Find a “canonical” class of distributions $\mathcal{C}$ that approximates $\mathcal{D}$ well.
   (For every $p \in \mathcal{D}$ there is $q \in \mathcal{C}$ such that $p \approx q$.)

2. Use samples from $p$ to estimate it as if it was $q$.

Reduction-based approach.
Main difficulty: Algorithm for $\mathcal{C}$ should be robust to error in the data.

Question: Which “canonical” class should we use?
Piecewise polynomial distributions

- Distribution $p$ is **$t$-piecewise degree-$d$** if there exists a partition of the domain into $t$ intervals such that within each interval, the density of $p$ is a degree-$d$ polynomial.
- Let $\mathcal{P}_{t,d}$ be the family of all such distributions.

$t = 4, \; d = 3$
Overview of Framework

**Approximation (Existential Step)**

\[ p \in \mathcal{D}, \quad \epsilon > 0 \]

\[ \exists t, d \text{ s.t.}: \mathcal{P}_{t,d} \approx \frac{\epsilon}{2} \mathcal{D} \]

**Agnostic Learner for \( \mathcal{P}_{t,d} \) (Algorithmic Step)**

\[ \min t \cdot (d + 1) \text{ s.t. for each } p \in \mathcal{D} \text{ there is } q \in \mathcal{P}_{t,d} \text{ with } \|q - p\|_1 \leq \frac{\epsilon}{2} \]

\[ \|h - p\|_1 \leq \text{OPT} + \frac{\epsilon}{2} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \]

Hypothesis \( h \)
Why Piecewise Polynomials?

- Analogy with PAC learning of Boolean functions
  [Linial-Mansour-Nisan’93]

- Common method in statistics: fitting splines to data
  [Wegman-Wright’83, Stone et al.’90’s, Willet-Nowak’07]

- Gives sample optimal and computationally efficient estimators for wide range of distribution families
# Results: Learning Structured Families

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Sample Size</th>
<th>Parameters</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>monotone</strong></td>
<td>$O(\log(n/\epsilon)/\epsilon^3)$</td>
<td>$t = \log n/\epsilon$, $d = 0$</td>
<td>Birgé’87</td>
</tr>
<tr>
<td><strong>$k$-modal</strong></td>
<td>$O(k/\epsilon^5/2)$</td>
<td>$t = k/\sqrt{\epsilon}$, $d = 1$</td>
<td>Daskalakis-D-Servedio’12</td>
</tr>
<tr>
<td><strong>log-concave</strong></td>
<td>$O(k/\epsilon^2)$</td>
<td>$t = k$, $d = \log(k/\epsilon)$</td>
<td>Chan-D-Servedio-Sun’13</td>
</tr>
<tr>
<td><strong>Gaussian</strong></td>
<td>$\widetilde{O}(k/\epsilon^2)$</td>
<td>$t = k$, $d = \log(k/\epsilon)$</td>
<td>Chan-D-Servedio-Sun’14, D-Kane’15</td>
</tr>
<tr>
<td><strong>$k$-mixture</strong></td>
<td>$\widetilde{O}(k/\epsilon^2)$</td>
<td>$t = k$, $d = \log(k/\epsilon)$</td>
<td></td>
</tr>
<tr>
<td><strong>Poisson/Binomial</strong></td>
<td>$\widetilde{O}(k/\epsilon^2)$</td>
<td>$t = k$, $d = \log(k/\epsilon)$</td>
<td>Daskalakis-D-Stewart’15</td>
</tr>
<tr>
<td><strong>Besov spaces</strong></td>
<td>$O(1/\epsilon^{2+1/\alpha})$</td>
<td>$t = \epsilon^{-1/\alpha}$, $d = \lceil \alpha \rceil$</td>
<td>Devore’ 98</td>
</tr>
<tr>
<td><strong>$k$-monotone</strong></td>
<td>$O(k/\epsilon^{2+1/k})$</td>
<td>$t = k$, $d = \epsilon^{-1/k}$</td>
<td>Konovalov-Leviatan’07</td>
</tr>
</tbody>
</table>

---

Previous work (parameter estimation): Moitra-Valiant’10

$\frac{1}{\epsilon} \Omega(k)$

hazard rate

$t = \log n/\epsilon, d = 0$
Question: Let $p$, $q$ be probability density functions. How many samples are required to distinguish between them?

Partial Answer: If $p$, $q$ have a few “crossings”, distinguishing is easy.
Question: Let $p$, $q$ be probability density functions. How many samples are required to distinguish between them?

Partial Answer: If $p$, $q$ have a few “crossings”, distinguishing is easy.

Typically, unbounded many crossings, but only a few are important.
“Complexity measure” for learning a distribution family

**Definition.** For $p, q : \mathbb{R} \to \mathbb{R}_+$ and $k \geq 1$, we define the $A_k$-distance between $p, q$ as follows:

$$
\|p - q\|_{A_k} = \sup_{\mathcal{I}=(I_i)_{i=1}^k} \sum_{i=1}^k |p(I_i) - q(I_i)|
$$

**Upper Bound on Sample Complexity:** For a family of one-dimensional distributions $\mathcal{D}$ and $\epsilon > 0$, let $k = k(\mathcal{D}, \epsilon)$ be the smallest integer such that for any $p, q \in \mathcal{D}$ it holds

$$
\|p - q\|_1 \approx_\epsilon \|p - q\|_{A_k}.
$$

Then, the parameter $k$ is an upper bound on the sample complexity of agnostic learning for $\mathcal{D}$. 
**Statistical Estimator**

**Lemma.** For any $\mathcal{D}$ and $\epsilon > 0$, let $k = k(\mathcal{D}, \epsilon)$ be such that for any $p, q \in \mathcal{D}$ it holds $\|p - q\|_1 \leq \|p - q\|_{A_k} + \epsilon$. Then there exists an agnostic learning algorithm for $\mathcal{D}$ using $O(k/\epsilon^2)$ samples.

**Proof Sketch.**

Consider the following procedure:
1. Draw $m = \Omega(k/\epsilon^2)$ samples from $p$ and let $\hat{p}_m$ be the empirical distr.
2. Compute $h \in \mathcal{D}$ that minimizes $\|h - \hat{p}_m\|_{A_k}$.

**Analysis:**

Empirical Process Theory (Vapnik, Chervonenkis, Dudley ~70’s)
Difficulties in Implementing Estimator

For any $\mathcal{D}$ and $\epsilon > 0$, let $k = k(\mathcal{D}, \epsilon)$ be such that for any $p, q \in \mathcal{D}$ it holds $\|p - q\|_1 \leq \|p - q\|_{\mathcal{A}_k} + \epsilon$.

**Algorithm:**
1. Draw $m = \Omega(k/\epsilon^2)$ samples from $p$ and let $\hat{p}_m$ be the empirical distr.
2. Compute $h \in \mathcal{D}$ that minimizes $\|h - \hat{p}_m\|_{\mathcal{A}_k}$.

**Main Issues:**

1. How do we bound the value of $k = k(\mathcal{D}, \epsilon)$?

2. How do we efficiently perform the “projection” step? (Non-convex optimization problem)

**Solution:** Replace $\mathcal{D}$ by $\mathcal{P}_{t,d}$ such that $\mathcal{D} \approx_{\epsilon/2} \mathcal{P}_{t,d}$.
Agnostically Learning Piecewise Polynomials

Application of general framework for $C = \mathcal{P}_{t,d}$ and $k = O(t(d + 1))$.

1. Draw $m = \Omega(t(d + 1)/\epsilon^2)$ samples from $p$.
2. Compute $h \in \mathcal{P}_{t,d}$ that minimizes $\|h - \hat{p}_m\|_{\mathcal{A}_k}$.

Still non-convex optimization problem…

Main Algorithmic Contribution:

Polynomial time algorithm for Step 2.
Agnostically Learning Piecewise Polynomials

**Theorem** [Chan-D-Servedio-Sun, STOC’14]

There exists an agnostic learning algorithm for $\mathcal{P}_{t,d}$ that uses

$$\tilde{O}(t(d + 1)/\epsilon^2)$$

samples and runs in time

$$\text{poly}(t, d + 1, 1/\epsilon).$$

Moreover, $\Omega(t(d + 1)/\epsilon^2)$ samples are information-theoretically necessary.

**Recent Progress:**

- Piecewise constant: near-linear time [Chan-D-Servedio-Sun, NIPS’14]
- **General Case:** $O(t(d + 1)/\epsilon^2)$ samples and $(t/\epsilon^2) \cdot \text{poly}(d)$ time.
  [Acharya-D-Li-Schmidt’15]
Overview of Techniques

- Empirical Process Theory
- Non-Convex Optimization Problem
- Convex Optimization, Dynamic Programming
- Information Theory
- Approximation Theory
- Piecewise Polynomial Approximation
Predictive performance of straightforward implementation: speed-up over recent implementations of the MLE.
Application in Databases: Succinct Representation of Data

[Acharya-D-Hegde-Li-Schmidt, PODS’15]: **Approximating Data Distributions by Histograms**

Classical problem in databases

**Goal:** Given data distribution, construct a succinct approximation (histogram). Minimize *computation time, approximation error.*

**Our Result:** Sample optimal, sample-linear time algorithm with optimal error (up to small constant factor).

**Experimental Evaluation:** Outperforms all previous algorithms for the problem by one to two orders of magnitude.
Empirical Results (I)
[Acharya-D-Hegde-Li-Schmidt, PODS’15]

- Two synthetic and one real-word data set (same as [Guha-Koudas-Shim’06])
Outline

• Learning via Piecewise Polynomial Approximation
  ▪ Introduction
  ▪ Framework Overview
  ▪ Statistical Efficiency
  ▪ Computational Efficiency
  ▪ Empirical Results

• Applications to other Inference Tasks

• Future Directions and Concluding Remarks
Additional Applications of Framework

Hypothesis Testing (Property Testing)

- Testing Identity of Structured Distributions [D-Kane-Nikishkin’15a, ‘15b]
  “Given samples from a structured distribution, is it uniform?”
  “Given samples from two structured distributions, are they the identical?”

- Testing Shape Restrictions [Canonne-D-Gouleakis-Rubinfeld’15]
  “Given samples from a (potentially arbitrary) distribution, is it structured?”
Outline

• Learning via Piecewise Polynomial Approximation
  ▪ Introduction
  ▪ Framework Overview
  ▪ Statistical Efficiency
  ▪ Computational Efficiency
  ▪ Empirical Results

• Applications to other Inference Tasks

• Future Directions and Concluding Remarks
Future Directions

Broad Context:

Complexity theory for statistical estimation

Specific Challenges:

• Agnostic proper learning

• “Instance optimal” (adaptive) algorithms

• Tradeoffs between sample size and computational efficiency

Thank you for your attention!