Beyond Histograms: Structure and Distribution Estimation

Ilias Diakonikolas
University of Edinburgh

STOC’14 Workshop
New York, May 2014
Learning (Discrete) Distributions

Probability distributions on $[N] = \{1, \ldots, N\}$

- Learning problem defined by class $\mathcal{C}$ of distributions
- Target distribution $\rho$ in $\mathcal{C}$ unknown to learner
- Learner given sample of i.i.d. draws from $\rho$

**Goal:** w.p. $\geq 9/10$ output $h$ satisfying

$$d_{TV}(h, \rho) := (1/2) \cdot ||h-\rho||_1 \leq \varepsilon$$
Agnostically Learning Distributions

• Learning problem defined by class \( \mathbf{C} \) of distributions

• Target distribution \( p \) unknown to learner and let

\[
\text{OPT} = \inf_{q \in \mathbf{C}} d_{TV}(p, q)
\]

• Learner given sample of i.i.d. draws from \( p \)

**Goal:** w.p. \( \geq 9/10 \) output \( h \) satisfying

\[
d_{TV}(h, p) \leq c \cdot \text{OPT} + \varepsilon
\]

for a constant \( c \geq 1 \).

Sample complexity and running time should depend only on \( \mathbf{C} \).
Analogies with PAC Learning Boolean Functions

- Class $\mathbf{C}$ of distributions
- Unknown target $p$ in $\mathbf{C}$
- Learner gets i.i.d. samples from $p$
- Output approximation $h$ of $p$

- Class $\mathbf{C}$ of Boolean functions
- Unknown target $f$ in $\mathbf{C}$
- Learner gets labeled samples $(x, f(x))$
- Output approximation $f'$ of $f$

Minimize:
- sample size (sample complexity)
- computation time (computational complexity)
Learning Arbitrary Discrete Distributions

Let $\mathbf{C} =$ set of all distributions over $[N]$

\textit{What is the best learning algorithm?}

Simple answer (folklore):

- Algorithm with sample (and time) complexity $O(N/\varepsilon^2)$
- Information theoretic lower bound of $\Omega(N/\varepsilon^2)$
Learning Arbitrary Discrete Distributions: Upper Bound

**Theorem:** Let \( p \) be a distribution over \([N]\). Let \( \hat{p} \) be empirical distribution over \([N]\) obtained by drawing \( m \) samples from \( p \). Then

\[
E[d_{TV}(\hat{p}, p)] \leq \sqrt{N/m}.
\]

**Proof:**

- For each \( i \in [N] \) have \( E[|p(i) - \hat{p}(i)|] \leq \sqrt{p(i)(1 - p(i))/m} \)

- Bound total error \( E[d_{TV}(\hat{p}, p)] \leq \sqrt{N/m} \) (Cauchy-Schwarz)

So can learn to accuracy \( \varepsilon \) from \( O(N/\varepsilon^2) \) samples.
Learning Arbitrary Discrete Distributions: Lower Bound

**Theorem:** There exists a class $H$ of distributions over $[N]$ with the following property: Any algorithm that learns an arbitrary distribution in $H$ to statistical distance $\epsilon$ requires $\Omega(N/\epsilon^2)$ samples.

**Proof:**
Let $H$ be defined as follows: Partition the domain into $N/2$ pairs of points $2i$ and $2i+1$. For each pair, one point has mass $(1+\epsilon)/N$ and another has mass $(1-\epsilon)/N$.

- Need to learn at least half of the pairs.

- Learning each pair requires $\Omega(1/\epsilon^2)$ samples.
Learning Arbitrary Discrete Distributions

Learning an *arbitrary* distribution over \([N]\):

Sample size \(\Theta(N/\varepsilon^2)\)

necessary and sufficient

When can we do better?
Which distributions are easy to learn, which are hard (and why)?
Types of Structured Distributions

- Distributions with “shape restrictions”

- Simple combinations of simple distributions
  
  **Mixtures** of simple distributions

  **Sums** of simple distributions *(talk by Costis)*
Structure and Density Estimation

Main messages of this talk:

• **We can exploit the underlying structure to do statistical estimation more efficiently.**

General recipe:
1. Given a “complex” class $\mathcal{C}$ of distributions: Prove that there exists a “simple” class of distributions $\mathcal{C}'$ such that any distribution $p$ in $\mathcal{C}$ can be *well-approximated* by a distribution in $\mathcal{C}'$.
2. Use samples from $p$ to agnostically learn it using $\mathcal{C}'$.

• **Histograms are not always sufficient to obtain (sample-) optimal results for statistical estimation problems.**
Statistics and Density Estimation

Classical topic in statistics. Many generic methods:

- Histograms [Pearson, 1900]
- Kernel methods [M. Rosenblatt, 1956]
- Maximum Likelihood [Fischer, 1912]
- Metric Entropy [A.N. Kolmogorov, 1960]

Many others: Nearest Neighbor, Orthogonal Series, ...

Focus traditionally on sample size.
Histograms

- “The oldest and most widely used method” [Silverman ’86]
- Goes back to Karl Pearson (1900).

Main Idea:

Approximation of the unknown density by a piecewise constant distribution
Shape Restricted Density Estimation

- Nonparametric Density 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.

-- Still very active research area: log-concavity, k-monotonicity, …

  Recent survey by Walther:

-- Standard tool in these settings: MLE
References for this Talk

- Learning k-modal distributions via testing
  [Daskalakis-D-Servedio, SODA’12]
- Approximating and Testing k-histogram distributions in sublinear time
  [Indyk-Levi-Rubinfeld, PODS’12]
- Learning Poisson Binomial Distributions
  [Daskalakis-D-Servedio, STOC’12]
- Learning Mixtures of Structured Distributions over Discrete Domains
  [Chan-D-Servedio-Sun, SODA’13]
- Testing $k$-modal Distributions: Optimal Algorithms via Reductions
  [Daskalakis-D-Servedio-Valiant$^2$, SODA’13]
- Learning Sums of Independent Integer Random Variables
  [Daskalakis-D-O’Donnell-Servedio-Tan, FOCS’13]
- Efficient Density Estimation via Piecewise Polynomial Approximation
  [Chan-D-Servedio-Sun, STOC’14, Tuesday morning]
Basic problem: Learning Histograms

Goal: Learn an unknown $k$-flat distribution $p$ over $[N]$.

- Simple setting: Intervals $I_1, \ldots, I_k$ are known:
  - Sample and time complexity $\Theta(k/\varepsilon^2)$

- What if the intervals are unknown?
Learning Histograms: Known Partition (I)

Goal: learn an unknown $k$-flat distribution $p$ over $[N]$.

Known intervals $I_1$, ..., $I_k$

**Definition:** Given a distribution $p$ over $[N]$ and a partition $I = \{I_1, \ldots, I_k\}$, of $[N]$ into $k$ intervals, the flattened distribution $\bar{p}$ is the distribution over $[N]$ that is uniform within each $I_j$ and satisfies $p(I_j) = \bar{p}(I_j)$

**Algorithm:**

- Draw $m = O\left(\frac{k}{\varepsilon^2}\right)$ samples from $p$; let $\hat{p}_m$ be the empirical distribution.
- Output the flattened empirical distribution $\bar{\hat{p}}_m$ over $I_1$, ..., $I_k$. 
Learning Histograms: Known Partition (II)

Known intervals $I_1, \ldots, I_k$

**Algorithm:**
- Draw $m = O\left(\frac{k}{\varepsilon^2}\right)$ samples from $p$; let $\hat{p}_m$ be the empirical distribution.
- Output the flattened empirical distribution $\overline{\hat{p}}_m$ over $I_1, \ldots, I_k$.

**Analysis:** We have that

$$d_{TV}(p, \overline{\hat{p}}_m) = \sum_{j=1}^{k} |p(I_j) - \hat{p}_m(I_j)|$$

Problem reduces to that of learning a distribution over the $k$ intervals. ■

**Note:** Algorithm is agnostic with constant $c = 2$, i.e., if $\text{OPT} = \min_{q \in (k-\text{flat})} d_{TV}(p, q)$ then

$$d_{TV}(p, \overline{\hat{p}}_m) \leq 2 \cdot \text{OPT} + \varepsilon$$
Application: Learning Monotone Distributions (I)

**Informal Structural Lemma:** Monotone distributions are well-approximated by “oblivious” histograms with “few” pieces.

- Consider class of non-increasing distributions over $[N]$.
- Decompose $[N]$ into $\ell = O((1/\varepsilon) \cdot \log N)$ intervals whose “widths” increase as powers of $(1 + \varepsilon)$. Call these the *oblivious buckets*.
Application: Learning Monotone Distributions (II)

**Lemma:** [Birge’87] For any monotone distribution $\rho$, we have

$$d_{TV}(\rho, \bar{\rho}) \leq \varepsilon$$

**Corollary:** The class of monotone distributions over $[N]$ can be efficiently learned to error $\varepsilon$ using $O\left(\frac{1}{\varepsilon} \cdot \log N\right)$ samples.

[Birge’85] Information-theoretic lower bound of $\Omega\left(\frac{1}{\varepsilon^3} \cdot \log N\right)$
Learning Histograms: Unknown Partition

Goal: learn an unknown $k$-flat distribution $p$ over $[N]$.

Easy if we know the $k$ intervals $I_1, \ldots, I_k$:
- Sample and time complexity $\Theta(k/\varepsilon^2)$.

What if the intervals are unknown?

Naïve approaches:
- Guessing them exactly: very inefficient $N^k$
- Guessing them approximately: not too great either $(1/\varepsilon)^k$
Unknown Partition: A first approach

Break up \([N]\) into \(\ell \gg k\) many intervals:

\[ p_{I_j} \] is not constant for at most \(k\) of the intervals \(I_1, \ldots, I_\ell\)

So, outputting uniform (sub-)distribution on each interval will usually give a good answer.
First approach in more detail

1. Divide \([N]\) into \(\ell = \frac{10k}{\varepsilon}\) intervals \(I_1, \ldots, I_\ell\) such that
   \[
p(I_j) \equiv \frac{\varepsilon}{10k}
   \]

2. Draw \(m = O\left(\frac{\ell}{\varepsilon^2}\right)\) samples from \(p\) and output the flattened empirical distribution over the intervals \(I_1, \ldots, I_\ell\)
First approach: Sketch of Analysis

1. Divide \([N]\) into \(\ell = 10k/\varepsilon\) intervals \(I_1, \ldots, I_\ell\) such that
   \[
   p(I_j) \equiv \varepsilon/10k
   \]
2. Draw \(m = O(\ell/\varepsilon^2)\) samples from \(p\) and output the flattened empirical distribution over the intervals \(I_1, \ldots, I_\ell\)

Analysis:

- The unknown \(p\) is not constant in at most \(k\) of the intervals \(I_1, \ldots, I_\ell\)
- Call such intervals “bad”. The total mass of those intervals is at most
  \[
  k \cdot \frac{\varepsilon}{10k} = \frac{\varepsilon}{10}
  \]
- The flattened empirical distribution gives \(\varepsilon\)-accuracy on the remaining intervals.
Improving the sample complexity?

- Sample complexity of \( O\left(\frac{k}{\epsilon^3}\right) \) came from the fact that we partitioned the domain into \( O\left(\frac{k}{\epsilon}\right) \) intervals, instead of just \( k \).

- Not clear whether sample size of \( O\left(\frac{k}{\epsilon^2}\right) \) suffices information-theoretically…

Alternate approach? Metric Entropy

**Definition:** For a class \( \mathcal{C} \) the \( \epsilon \)-metric entropy (Kolmogorov entropy) is:

\[
\text{Ent}(\mathcal{C}) = \inf \left\{ \log_2 (|\mathcal{M}|) \right\}, \text{where } \mathcal{M} \text{ is an } \epsilon \text{-cover of } \mathcal{C} \]

**Theorem:** [Devroye-Lugosi’ 01] For any class \( \mathcal{C} \) of distributions suppose there exists an \( \epsilon \)-cover for \( \mathcal{C} \) of size \( M \). There is an algorithm that learns an arbitrary distribution from \( \mathcal{C} \) to accuracy \( \epsilon \) using

\[
O\left(\frac{1}{\epsilon^2} \cdot \log M\right)
\]

draws from the distribution. (The running time of the algorithm is \( \Omega(M) \).)
Improving the sample complexity: Metric Entropy Bounds

**Theorem:** [DL’01] For any class $C$ suppose there exists an $\varepsilon$-cover of size $M$. There is an algorithm that learns an arbitrary distribution from $C$ to error $\varepsilon$ using $O\left(\left(1/\varepsilon^2\right) \cdot \log M\right)$ draws from the distribution.

**Claim:** There exists an $\varepsilon$-cover for $k$-flat distributions of size $\left(k/\varepsilon\right)^{O(k)}$

**Corollary:** The class of $k$-flat distributions is learnable to accuracy $\varepsilon$ with sample size $\tilde{O}\left(k/\varepsilon^2\right)$

**Main Caveat:** Not a computationally efficient algorithm.

**Can we obtain a computationally efficient algorithm with optimal sample complexity?**
Towards a computationally efficient sample-optimal algorithm

Proposed Algorithm:

• Make \( m = \tilde{O}(k/\varepsilon^2) \) draws from \( \rho \) and let \( \hat{P}_m \) be the empirical distribution
• Find a hypothesis \( h \) that minimizes the variation distance from \( \hat{P}_m \)

Fails badly...

Also fails if we additionally require the hypothesis \( h \) to be \( k \)-flat
The VC-inequality

Recall the definition of statistical distance. For two distributions $p$, $q$ over $[N]$ we have that

$$d_{TV}(p, q) \equiv \max_{A \subseteq [N]} |p(A) - q(A)|$$

The VC inequality relates the empirical and the true distribution under a weaker metric.

**Definition**: Let $\mathcal{A}_k$ be the collection of unions of at most $k$ intervals in $[N]$. We define the $\mathcal{A}_k$-distance between $p$ and $q$ by

$$d_{\mathcal{A}_k}(p, q) \equiv \max_{A \in \mathcal{A}_k} |p(A) - q(A)|$$

**Theorem (VC inequality)**: Let $p$ be an arbitrary distribution over $[N]$. We have that

$$E[d_{\mathcal{A}_k}(p, \hat{p}_m)] = O\left(\frac{\sqrt{k}}{m}\right)$$
Theorem (VC inequality): Let $p$ be an arbitrary distribution over $[N]$. We have that $E[d_A^k(p, \hat{p}_m)] = O(\sqrt{k/m})$

Corollary: After $m = O\left(\frac{k}{\epsilon^2}\right)$ samples with probability at least 9/10, we have

$$d_A^k(p, \hat{p}_m) \leq \epsilon/2$$

Note that $d_{TV}(p, \hat{p}_m) \approx 1$!

How to proceed?

- Compute a **k-flat** distribution $h$ that minimizes $d_A^k(h, \hat{p}_m)$
- Output $h$

Why does this work?
Optimally Learning k-histograms: Upper Bound (II)

Corollary: After \( m = O\left(\frac{k}{\varepsilon^2}\right) \) samples with probability at least 9/10, we have

\[
d_{\mathcal{A}_k}(p, \hat{p}_m) \leq \varepsilon/2
\]

Algorithm:
- Compute a \textbf{k-flat} distribution \( h \) that minimizes \( d_{\mathcal{A}_k}(h, \hat{p}_m) \)
- Output \( h \)

Analysis: Note that \( d_{\mathcal{A}_k}(h, \hat{p}_m) \leq \varepsilon/2 \), hence \( d_{\mathcal{A}_k}(h, p) \leq \varepsilon \)

But since \( h \) and \( p \) are both \( k \)-flat \( d_{TV}(h, p) = d_{\mathcal{A}_k}(h, p) \). ■
Essentially same argument works for agnostic case. Let \( p \) be an arbitrary distribution over \([N]\) and let

\[
\text{OPT} = \inf_{q \in (k-\text{flat})} d_{TV}(p,q)
\]

“Non-constructive” algorithm:

- Draw \( m = O\left( \frac{k}{\epsilon^2} \right) \) samples from \( p \).
- Compute a \textbf{k-flat} distribution \( h \) that minimizes \( d_{A_k}(h,\hat{p}_m) \)
- Output \( h \)

**Theorem:** Above algorithm outputs a distribution \( h \) that with probability at least 9/10 satisfies

\[
d_{TV}(h,p) \leq 3 \cdot \text{OPT} + \epsilon
\]

Main Issue: How to efficiently implement the second step?
Optimally Learning k-histograms: Upper Bound (IV)

- Draw $m = O\left(\frac{k}{\varepsilon^2}\right)$ samples from $p$.
- Compute a $\textbf{k-flat}$ distribution $h$ that minimizes
- Output $h$

Second step can be done in time $\tilde{O}\left(\frac{k^3}{\varepsilon^2}\right)$ by an appropriate DP.

Main Idea:

\textbf{Fact: } $d_{\mathcal{A}_k} (p, q)^{(J \cup K)} \leq \max_{0 \leq l \leq k} \left\{ d_{\mathcal{A}_l} (p, q)^{(J)} + d_{\mathcal{A}_{l-k+1}} (p, q)^{(K)} \right\}$

Can we learn $k$-histograms with optimal sample size and in near-linear time?

\textbf{Yes } [Chan-D-Servedio-Sun ’14b]
Application: Learning Structured distributions (I)

Hazard rate of $p$ over $[N]$: $H(i) = \frac{p(i)}{\sum_{j \geq i} p(j)}$

Consider the class of Monotone Hazard Rate (MHR) Distributions. (Important in reliability, economics, etc.)

**Lemma**: Every MHR distribution over $[N]$ is $\varepsilon$-close to being $k$-flat for

$$k = O\left((1/\varepsilon) \cdot \log n\right)$$

**Corollary**: MHR distributions over $[N]$ are efficiently learnable with sample complexity

$$O\left((1/\varepsilon^3) \cdot \log n\right)$$

**Note**: The above bound is best possible: $\Omega\left((1/\varepsilon^3) \cdot \log n\right)$ samples are information-theoretically required to learn MHR Distributions.
Application: Learning Structured distributions (II)

<table>
<thead>
<tr>
<th>Distribution Class</th>
<th>Sample Complexity Upper Bound</th>
<th>Sample Complexity Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monotone</td>
<td>$O\left(\frac{1}{\varepsilon^3} \cdot \log n\right)$</td>
<td>Matching</td>
</tr>
<tr>
<td>$t$-modal</td>
<td>$O\left(\frac{t}{\varepsilon^3} \cdot \log \left(n/t\right)\right)$</td>
<td>Matching</td>
</tr>
<tr>
<td>MHR</td>
<td>$O\left(\frac{1}{\varepsilon^3} \cdot \log n\right)$</td>
<td>Matching</td>
</tr>
<tr>
<td>Log-concave</td>
<td>$O\left(\frac{1}{\varepsilon^3}\right)$</td>
<td>$\Omega\left(\frac{1}{\varepsilon^{5/2}}\right)$</td>
</tr>
</tbody>
</table>

Upper (and lower) bounds immediately generalize to mixtures.

Another application:
Learning Sums of Independent Integer random variables
[Daskalakis-D-O’Donnell-Servedio-Tan, FOCS’13]
Case Study: Log-concave (LC) Distributions

Fact: Every LC distribution can be $\epsilon$-approximated by a piecewise constant distribution with $O(1/\epsilon)$ pieces.

Corollary 1: The class of LC distributions can be efficiently learned with $O(1/\epsilon^3)$ samples.

Above fact is quantitatively tight.

Lower bound of $\Omega\left(\left(\frac{1}{\epsilon^{5/2}}\right)\right)$ considers piecewise linear distributions.

Lemma: Every LC distribution can be $\epsilon$-approximated by a piecewise linear distribution with $O(1/\sqrt{\epsilon})$ pieces.

Can we agnostically learn piecewise linear distributions?
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 PDF of $p$ is a degree-$d$ polynomial.

$t = 4, d = 3$
Learning distributions that are close to t-piecewise degree-d

Informal Theorem:
(with Chan, Servedio, Sun, STOC’14 Tuesday morning)

There is a **computationally efficient** learning algorithm that finds a hypothesis distribution which approximates any unknown distribution \( p \) almost as well as the best \( t \)-piecewise degree-\( d \) distribution does.
Theorem: Let \( p \) be an arbitrary distribution and

\[
\text{OPT} = \inf_{q \in (t - \text{piecewise degree} - d)} d_{TV}(p, q)
\]

There is an algorithm that uses \( \tilde{O}(t \cdot d/\varepsilon^2) \) samples from \( p \), runs in time \( \text{poly}(t, d, 1/\varepsilon) \) and outputs a hypothesis distribution \( h \) such that

\[
d_{TV}(h, p) \leq 3 \cdot \text{OPT} + \varepsilon
\]

Moreover, sample complexity of \( \Omega(t \cdot d/\varepsilon^2) \) is information-theoretically necessary even for \( \text{OPT} = 0 \).
Why Piecewise Polynomials?

Three main justifications:

• Analogy with PAC learning of Boolean functions (Linial-Mansour-Nisan’93)

• Common heuristic: fitting splines to the data

• Gives sample optimal efficient estimators for wide range of distribution classes
Applications: Learning with Piecewise Polynomials

High-level description of Algorithm:
- Linear Programming within Each “piece”
  (Analysis requires polynomial approximation theory)
- Dynamic Programming to “discover” the “correct partition”

Sample optimal bounds for essentially all previously studied shape constrained density estimation problems.

<table>
<thead>
<tr>
<th>Distribution Class</th>
<th>Sample Complexity Upper Bound</th>
<th>Sample Complexity Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-concave</td>
<td>$\tilde{O}(1/\varepsilon^{5/2})$</td>
<td>Matching</td>
</tr>
<tr>
<td>Mixture of k Gaussians</td>
<td>$\tilde{O}(k/\varepsilon^2)$</td>
<td>Matching</td>
</tr>
<tr>
<td>k-monotone</td>
<td>$\tilde{O}(k/\varepsilon^{2+1/k})$</td>
<td></td>
</tr>
</tbody>
</table>
Goals for future work

• Better accuracy? What is the optimal constant $c$ such that
  \[ d_{TV}(h, p) \leq c \cdot \text{OPT} + \epsilon \]
  (using same sample size)?
  -- Our upper bound $c=3$. No better than 2 possible [CDSS’14b].

• Better running time. Can we do near-linear time?
  -- For k-flat distributions, YES [CDSS’14b]. General case? OPEN

• Proper algorithms? e.g., k-GMMs

• Higher dimensions?

• Property Testing? Some preliminary progress [DDSVV’13]
Multi-dimensional histograms

Target distribution over $[0,1]^d$ is specified by $k$ hyper-rectangles that cover $[0,1]^d$; pdf is constant within each rectangle.

**Question:** Can we learn such distributions without incurring the “curse of dimensionality”?  
(Don’t want runtime to be exponential in $d$)
Higher dimensions

• Learning multi-dimensional histograms:

• Sample size well-understood: \( O\left( k \cdot d / \varepsilon^2 \right) \)

• Computational complexity?

  -- At least as hard as learning \( k \)-leaf decision trees over \( d \) variables.

  -- Bottleneck: \( k^{\Omega(\log d)} \)

-- Can we get such an algorithm?
References

- Learning k-modal distributions via testing
  [Daskalakis-D-Servedio, SODA’12]
- Approximating and Testing k-histogram distributions in sublinear time
  [Indyk-Levi-Rubinfeld, PODS’12]
- Learning Poisson Binomial Distributions
  [Daskalakis-D-Servedio, STOC’12]
- Learning Mixtures of Structured Distributions over Discrete Domains
  [Chan-D-Servedio-Sun, SODA’13]
- Testing $k$-modal Distributions: Optimal Algorithms via Reductions
  [Daskalakis-D-Servedio-Valiant$^2$, SODA’13]
- Learning Sums of Independent Integer Random Variables
  [Daskalakis-D-O’Donnell-Servedio-Tan, FOCS’13]
- Efficient Density Estimation via Piecewise Polynomial Approximation
  [Chan-D-Servedio-Sun, STOC’14, Tuesday morning]
Thank you