

Efficient Distribution Estimation
via
Piecewise Polynomial Approximation

Ilias Diakonikolas
University of Edinburgh

TCS+ Seminar
May 13 2015

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]

[Acharya-D-Li-Schmidt, arxiv'15]

Key Idea:

**Exploit piecewise polynomial approximation
for structured model estimation**

Additional Applications

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]

[D-Kane-Nikishkin, SODA'15]

[D-Kane-Nikishkin, 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
- Future Directions and Concluding Remarks

Outline

- Learning via Piecewise Polynomial Approximation
 - Introduction
 - Framework Overview
 - Statistical Efficiency
 - Computational Efficiency
 - Empirical Results
- Future Directions and Concluding Remarks

Distribution Learning (Density Estimation)

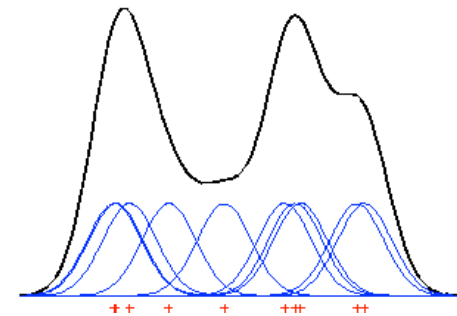
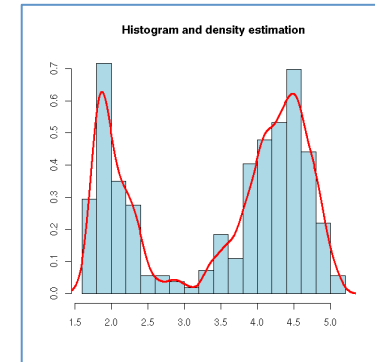
Given samples (observations) from an unknown probability distribution (model), construct an accurate estimate of the distribution.

- 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]

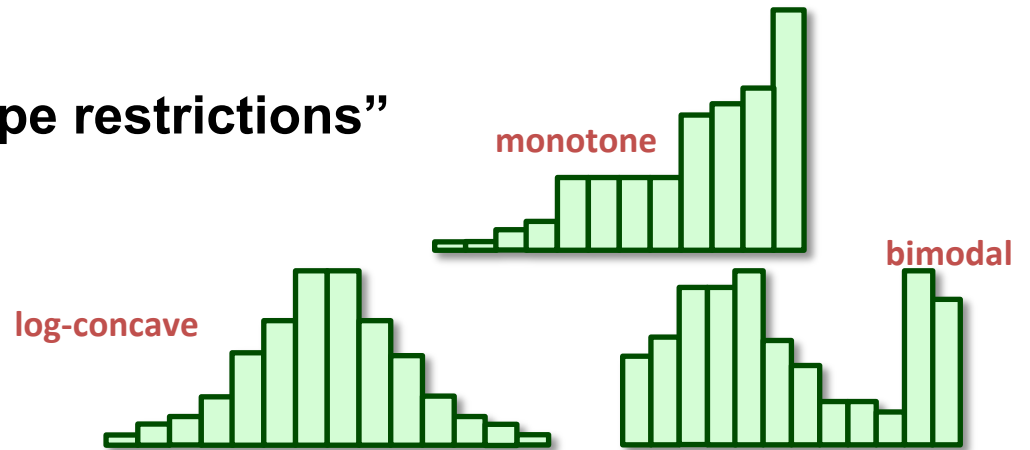


Many others: Nearest Neighbor, Orthogonal Series, ...

Focus traditionally on sample size.

Types of Structured Distributions

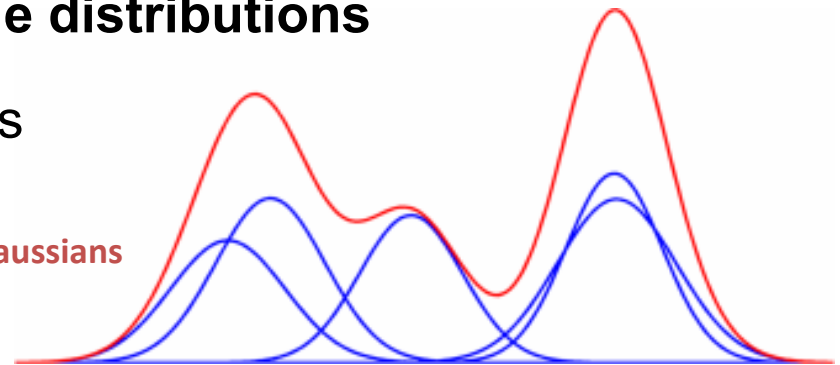
- Distributions with “shape restrictions”



- Simple combinations of simple distributions

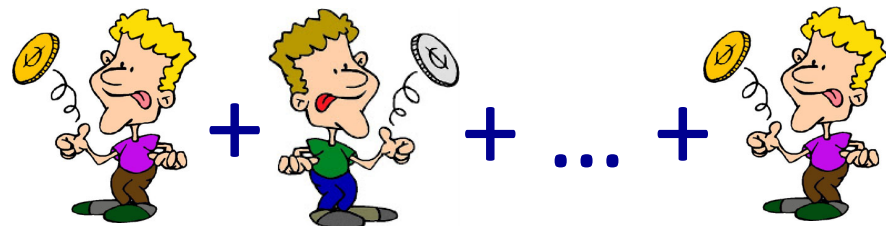
Mixtures of simple distributions

mixtures of Gaussians



Sums of simple distributions

Poisson Binomial Distributions



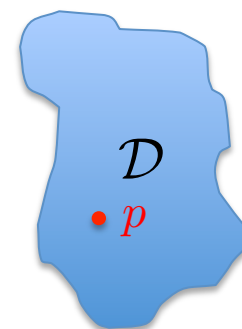
History

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, ...
[Balabdaoui-Wellner'07, Balabdaoui-Rufibach-Wellner'09, Walther'09, Dumbgen-Rufibach'09, Cule-Samworth'10, Koenker-Mizera'10, Guntuboyina-Sen'13, Doss-Wellner'13, Kim-Samworth'14]
- Standard tool in these settings: MLE

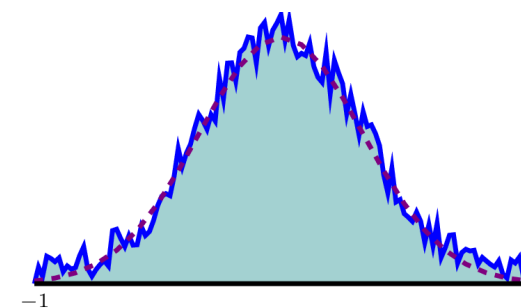
Distribution Learning: Definition

- 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

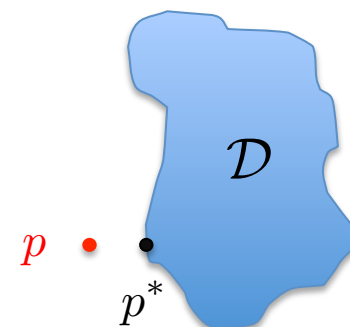
$$\|h - p\|_1 \leq \epsilon.$$



Goal: Sample optimal & computationally efficient algorithms

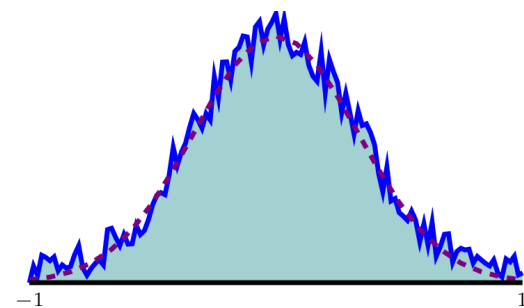
Agnostic Learning: Definition

- 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

$$\|h - p\|_1 \leq C \cdot \text{OPT} + \epsilon.$$

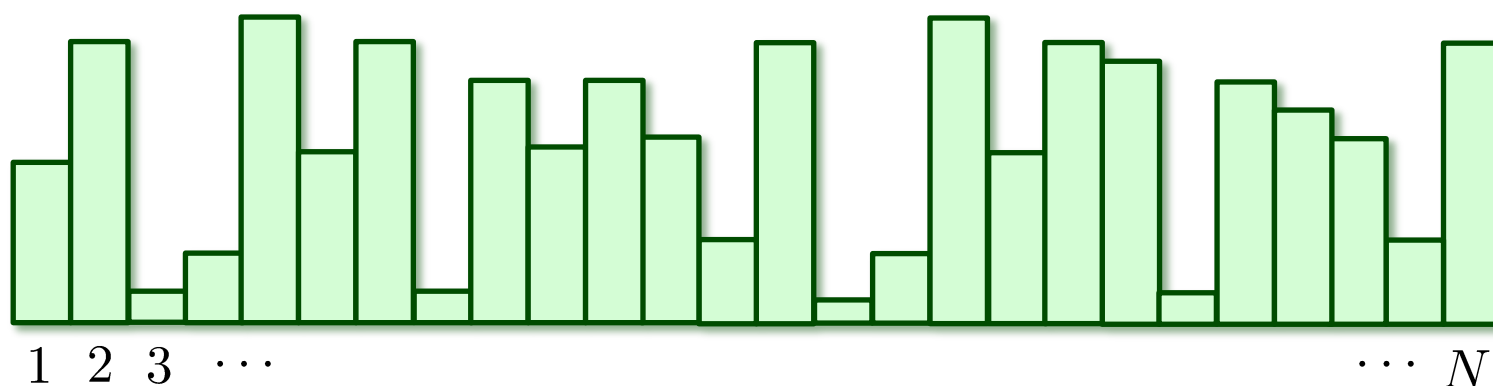


Goal: Sample optimal & computationally efficient algorithms

Learning Arbitrary Discrete Distributions

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

What is the best learning algorithm?

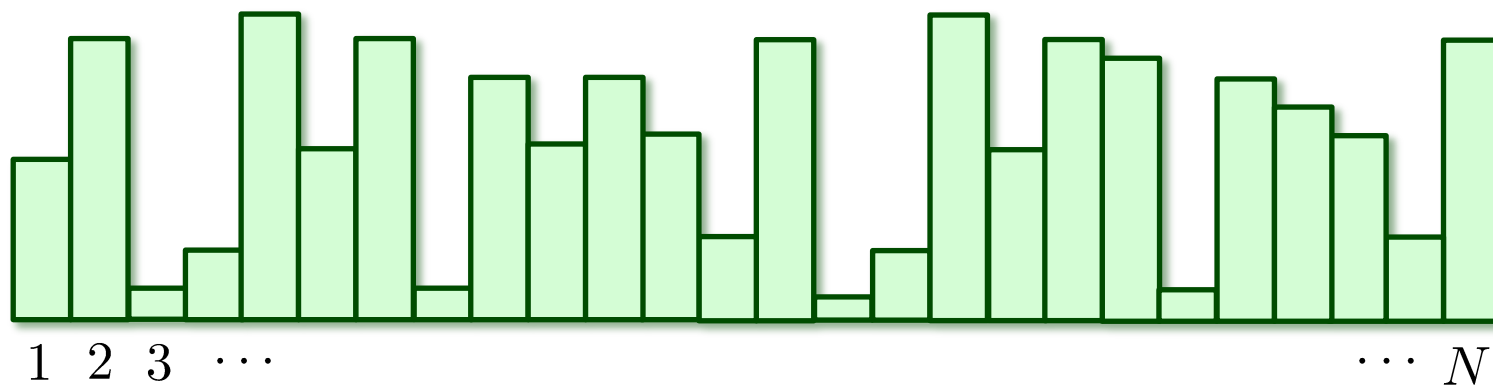


Simple answer (folklore):

- 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)?

Structure and Statistical Estimation

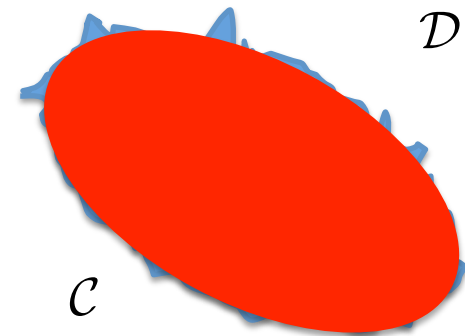
General Recipe for Statistical Estimation:

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 .



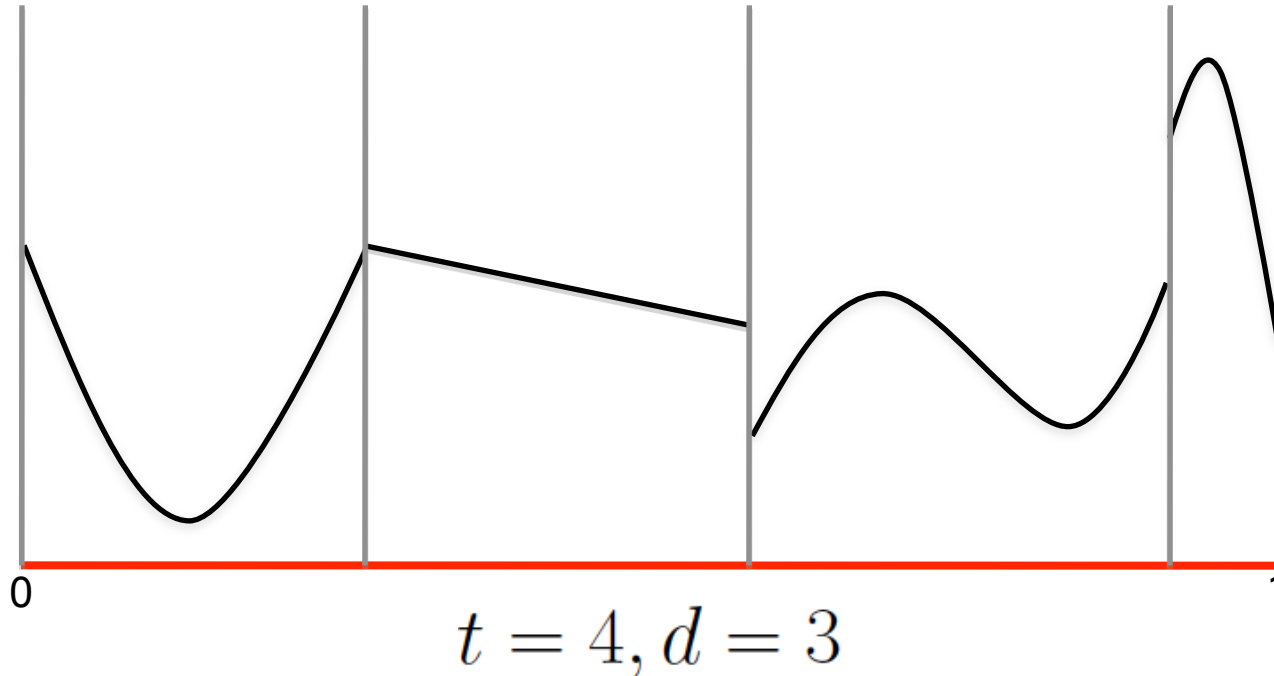
Reduction-based approach.

Main difficulty: Algorithm for \mathcal{C} should be **robust to error** in the data.

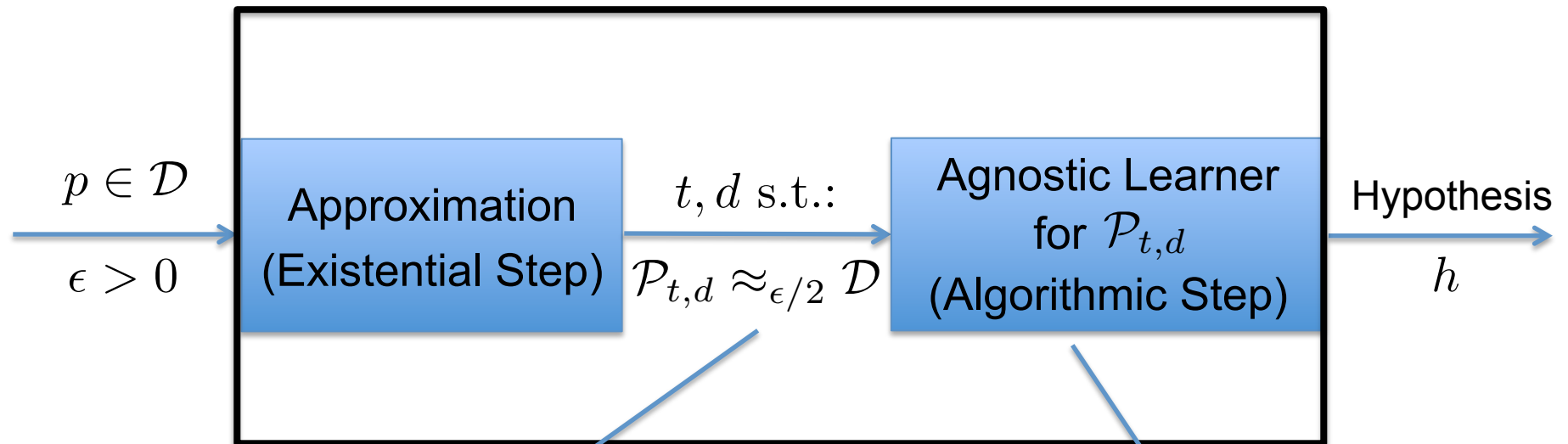
Question: Which “canonical” class should we use?

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.



Overview of Framework



$$\min t \cdot (d + 1) \text{ s.t.}$$

for each $p \in \mathcal{D}$ there is
 $q \in \mathcal{P}_{t,d}$ with

$$\|q - p\|_1 \leq \epsilon/2$$

$$\|h - p\|_1 \leq \text{OPT} + \epsilon/2$$

$$\leq \epsilon/2 + \epsilon/2$$

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, Willett-Nowak'07]
- Gives sample optimal and computationally efficient estimators for wide range of distribution families

Results: Learning Structured Families

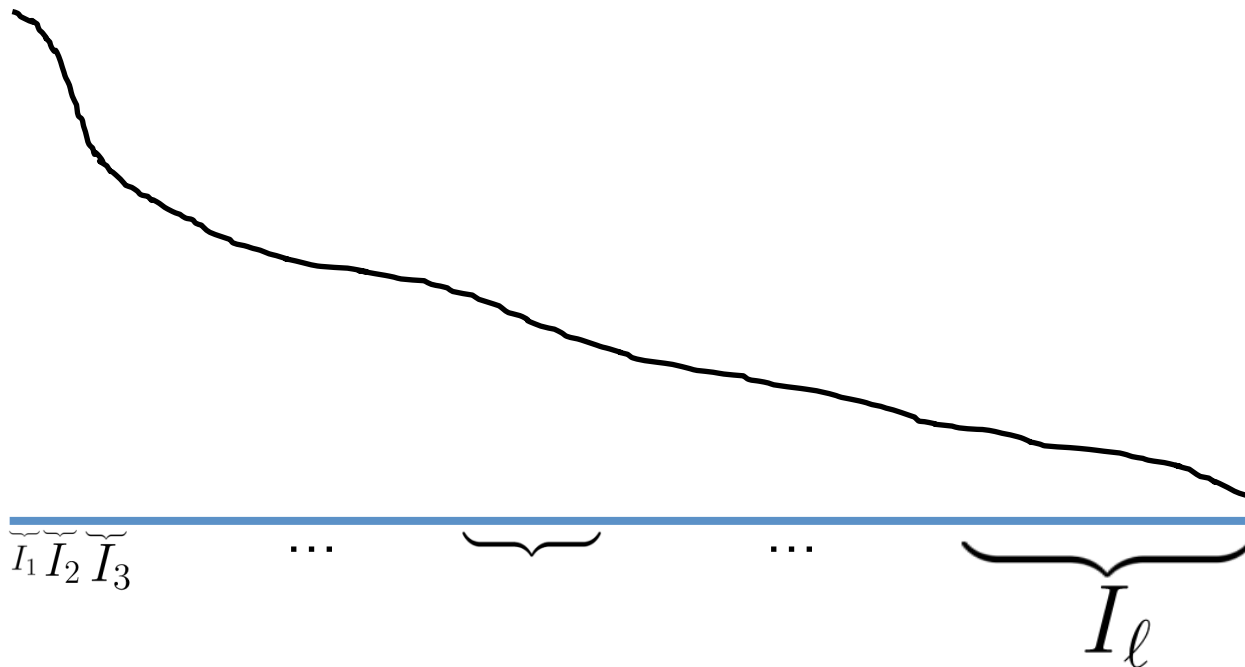
Distribution	Sample Size	Parameters	Reference (struc. result)
Exponential	$O(k/\epsilon^3)$	$t = \log n/\epsilon, d = 0$	Birgé'87
hazard rate	$O(k/\epsilon^3)$	$t = k \log n/\epsilon, d = 0$	Daskalakis-D-Servedio'12
log-concave k -mixture	$O(\log(n/\epsilon)/\epsilon^3)$	$t = \log(n/\epsilon), d = 0$	Chan-D-Servedio-Sun'13
Gaussian k -mixture	$O(k/\epsilon^{5/2})$	$t = k/\sqrt{\epsilon}, d = 1$	Chan-D-Servedio-Sun'14, D-Kane'15
Poisson/Binomial k -mixture	$\tilde{O}(k/\epsilon^2)$	$t = k, d = \log(1/\epsilon)$	Timan'63
Besov spaces	$\tilde{O}(k/\epsilon^2)$	$t = k, d = \log(1/\epsilon)$	Daskalakis-D-Stewart'15
k -monotone	$O(1/\epsilon^{2+1/\alpha})$	$t = \epsilon^{-1/\alpha}, d = \lceil \alpha \rceil$	Devore'98
	$O(k/\epsilon^{2+1/k})$	$t = k, d = \epsilon^{-1/k}$	Konovalov-Leviatan'07

Previous work
(parameter estimation):
Moitra-Valiant'10
 $(1/\epsilon)^{\Omega(k)}$

Example 1: Monotone Distributions (I)

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

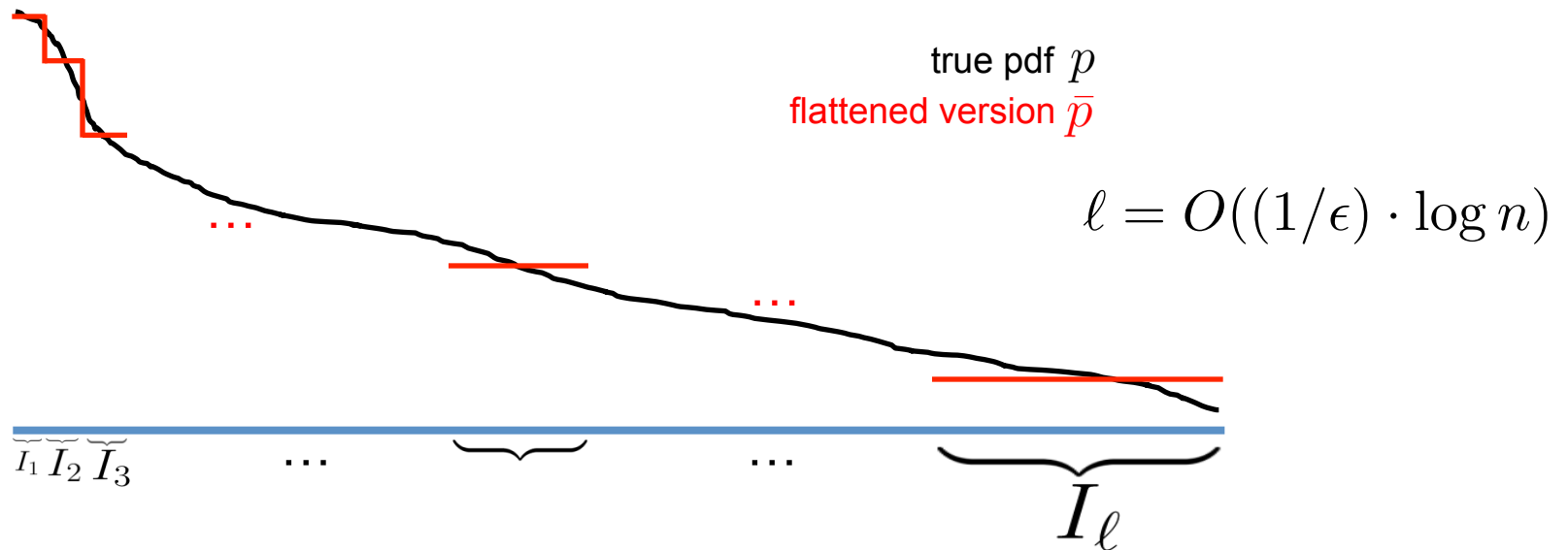
- Consider class of non-increasing distributions over $[n]$.
- Decompose $[n]$ into $\ell = O((1/\epsilon) \cdot \log n)$ intervals whose “widths” increase as powers of $(1 + \epsilon)$. Call these the *oblivious buckets*.



Example 1: Monotone Distributions (II)

Lemma: [Birge'87] For any monotone distribution p , we have

$$\|p - \bar{p}\|_1 \leq \epsilon$$



Corollary: The class of monotone distributions over $[n]$ can be efficiently learned to error ϵ using $O(\ell/\epsilon^2) = O((1/\epsilon^3) \log n)$ samples.

[Birge'85] Matching Information-theoretic lower bound

Case Study: Log-concave (LC) Distributions

Lemma [CDSS, SODA'13]: Every LC distribution can be ϵ - approximated by a piecewise constant distribution with $O(1/\epsilon)$ pieces.

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

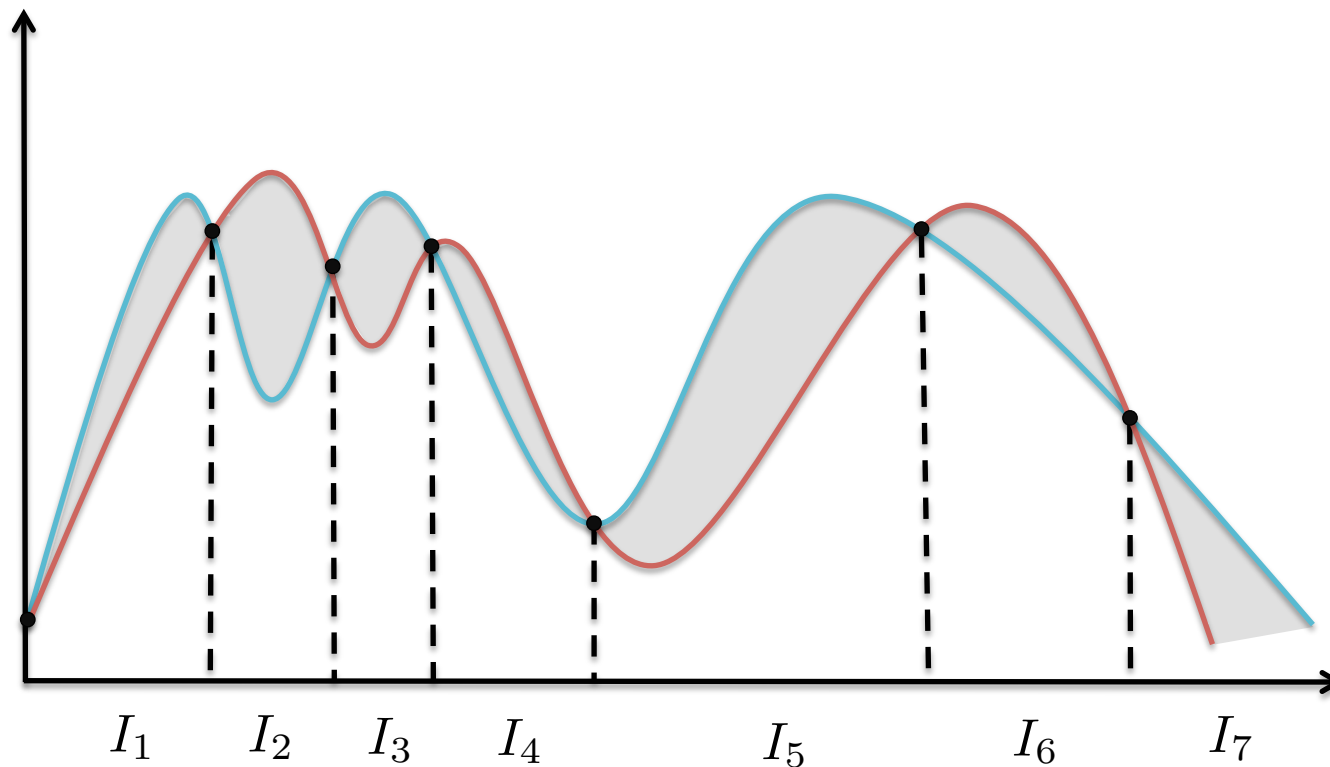
[Devroye-Lugosi'01] Lower bound of $\Omega(1/\epsilon^{5/2})$. Construction uses piecewise *linear* distributions.

Lemma [CDSS'14, DK'15]: Every LC distribution can be ϵ - approximated by a piecewise **linear** distribution with $O(1/\sqrt{\epsilon})$ pieces.

Statistical Performance: Intuition (I)

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.

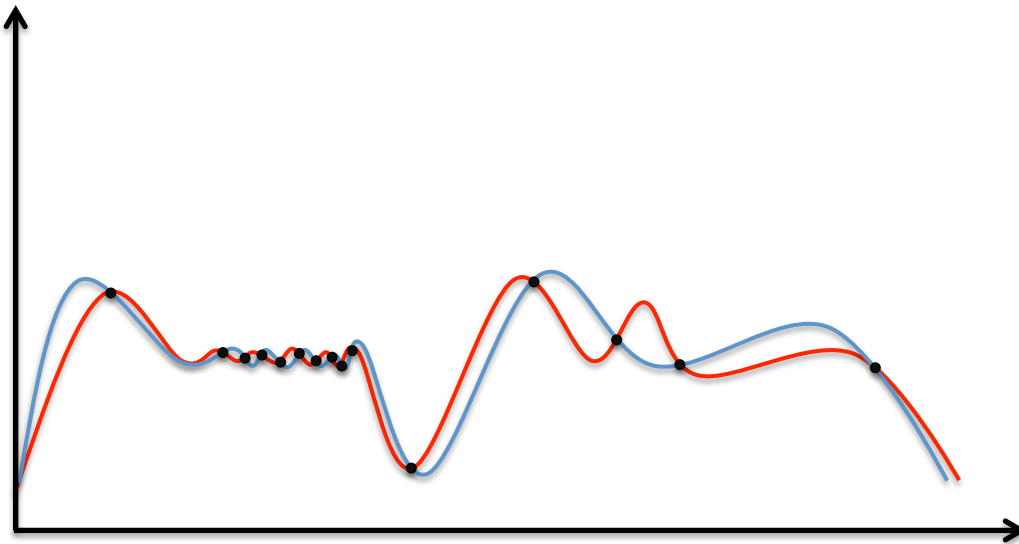


Statistical Performance: Intuition (II)

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 $p, q : \mathbb{R} \rightarrow \mathbb{R}_+$ and $k \geq 1$, we define the \mathcal{A}_k - distance between p, q as follows:

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



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

$$\|p - q\|_1 \approx_\epsilon \|p - q\|_{\mathcal{A}_k}.$$

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

Generic Algorithm for Learning a Distribution Family (I)

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 $\|p - q\|_1 \leq \|p - q\|_{\mathcal{A}_k} + \epsilon$. Then there exists an agnostic learning algorithm for \mathcal{D} using $O(k/\epsilon^2)$ samples.

Proof. Consider the following 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 $\|h - \hat{p}_m\|_{\mathcal{A}_k}$.

Analysis:

We will use the following result from the theory of empirical process:

Theorem (“VC-inequality”) For any density function $p : \mathbb{R} \rightarrow \mathbb{R}_+$ we have:

$$\mathbb{E}[\|p - \hat{p}_m\|_{\mathcal{A}_k}] = O(\sqrt{k/m})$$

Generic Algorithm for Learning a Distribution Family (II)

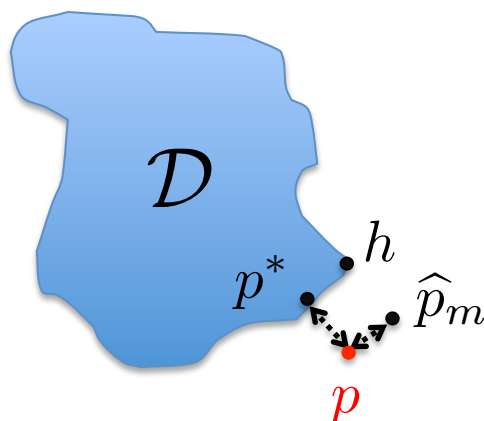
Algorithm:

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 $\|h - \hat{p}_m\|_{\mathcal{A}_k}$.

Analysis:

- With probability at least $\geq 9/10$ we have $\|p - \hat{p}_m\|_{\mathcal{A}_k} \leq \epsilon$.
- Let $\text{OPT} = \inf_{q \in \mathcal{D}} \|p - q\|_1$. Then there exists $p^* \in \mathcal{D}$:

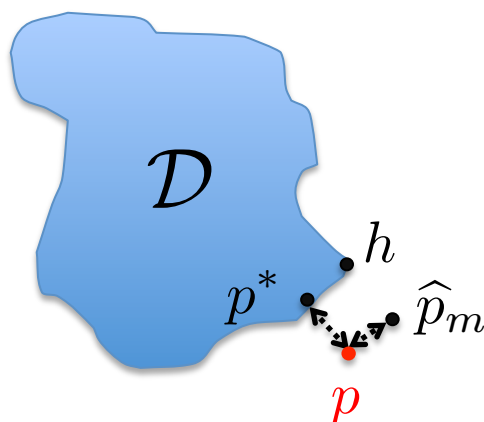
$$\|p - p^*\|_1 \leq \text{OPT}$$



$$\begin{aligned} \|h - p\|_1 &\leq \|h - p^*\|_1 + \|p^* - p\|_1 \\ &\leq \|h - p^*\|_1 + \text{OPT} \end{aligned}$$

Generic Algorithm for Learning a Distribution Family (II)

- With probability at least $\geq 9/10$ we have $\|p - \hat{p}_m\|_{\mathcal{A}_k} \leq \epsilon$.
- There exists $p^* \in \mathcal{D} : \|p - p^*\|_1 \leq \text{OPT}$.



$$\begin{aligned} \|h - p\|_1 &\leq \|h - p^*\|_1 + \|p^* - p\|_1 \\ &\leq \|h - p^*\|_1 + \text{OPT} \end{aligned}$$

We can write:

$$\begin{aligned} \|h - p^*\|_1 &\leq \|h - p^*\|_{\mathcal{A}_k} + \epsilon \\ &\leq \|h - \hat{p}_m\|_{\mathcal{A}_k} + \|\hat{p}_m - p^*\|_{\mathcal{A}_k} + \epsilon \\ &\leq 2\|\hat{p}_m - p^*\|_{\mathcal{A}_k} + \epsilon \end{aligned}$$

$$\begin{aligned} \|p^* - \hat{p}_m\|_{\mathcal{A}_k} &\leq \|p^* - p\|_{\mathcal{A}_k} + \|p - \hat{p}_m\|_{\mathcal{A}_k} \\ &\leq \text{OPT} + \epsilon \end{aligned}$$



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 $\mathcal{C} = \mathcal{P}_{t,d}$ 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.

Main Result of [CDSS14]

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.

- Piecewise constant: near-linear time [Chan-D-Servedio-Sun, NIPS'14]

Recent Progress [ADLS'15]

Theorem [Acharya-D-Li-Schmidt, '15]

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

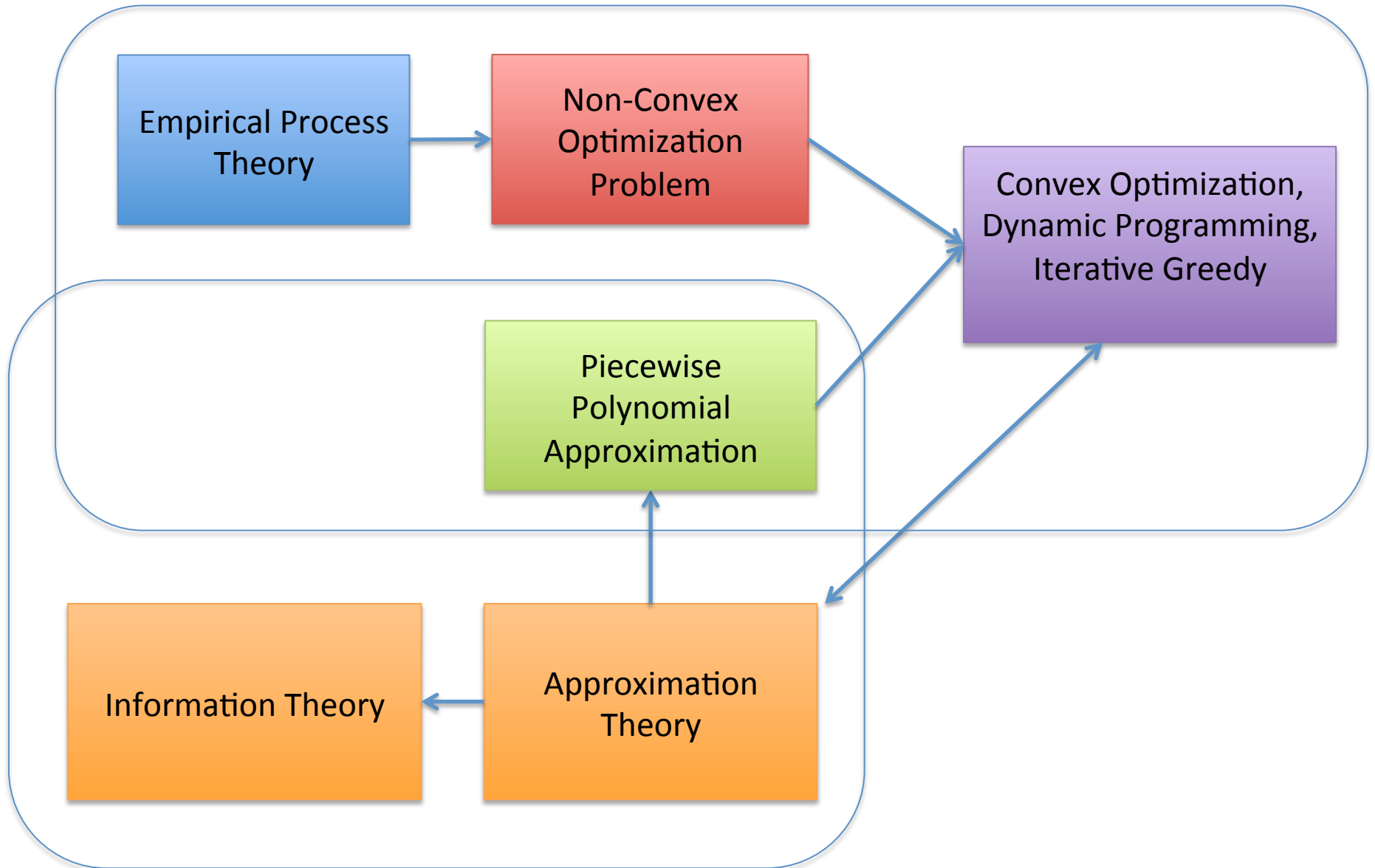
$$O(t(d+1)/\epsilon^2)$$

samples and runs in time

$$(t/\epsilon^2) \cdot \text{poly}(d).$$

Sample Optimal and Near-Linear Time as long as $d = (1/\epsilon)^{o(1)}$.

Overview of Techniques



Information Theoretic Lower Bound

Assouad's Lemma: Reduction of statistical estimation to a combinatorial construction.

Construct a sequence of polynomials with “large” L_1 distance and “small” Hellinger distance.

High-level Idea: Approximation of piecewise constant functions by low-degree polynomials

Main Technical Ingredient: Construction in [DGJSV, FOCS'09] of low-degree polynomial approximation to the sign function.

Algorithm for Piecewise Polynomial Densities

Application of general framework for $\mathcal{C} = \mathcal{P}_{t,d}$ 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 $\|h - \hat{p}_m\|_{\mathcal{A}_k}$.

Polynomial time algorithm for Step 2.

High-level description:

- Convex Programming within each “piece”

Main Difficulty: Exponential Number of Inequalities.

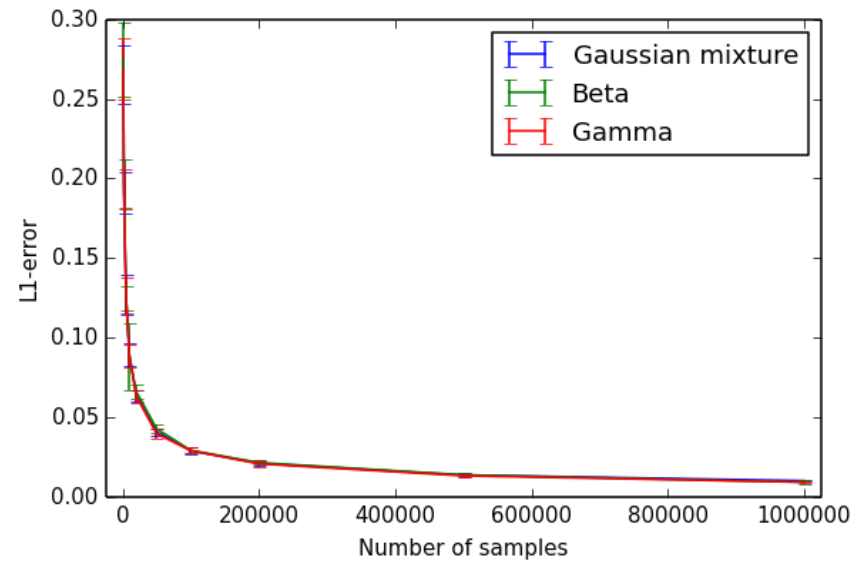
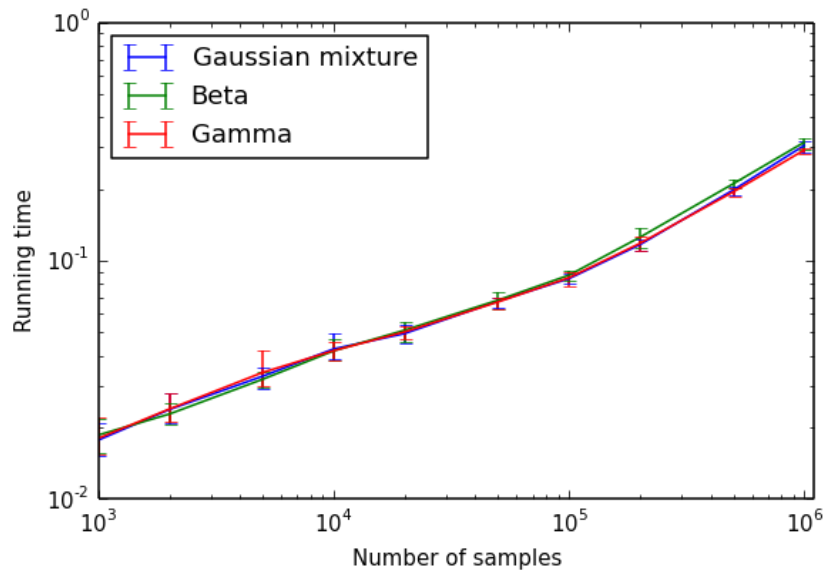
- [CDSS'14]: Polynomial size LP
- [ADLS'15]: Linear Time Separation Oracle

- “Discover” the “correct partition”
- [CDSS'14]: Dynamic Programming
- [ADLS'15]: Iterative Greedy Merging

Illustrative Empirical Results

[Acharya-D-Li-Schmidt '15]

Predictive performance of straightforward implementation:
speed-up over recent implementations of the MLE.



Outline

- Learning via Piecewise Polynomial Approximation
 - Introduction
 - Framework Overview
 - Statistical Efficiency
 - Computational Efficiency
 - Empirical Results
- Future Directions and Concluding Remarks

Future Directions

Broad Context:

Complexity theory for statistical estimation

Specific Challenges:

- Agnostic proper learning
(e.g., [Daskalakis-Kamath'14], [Li-Schmidt'15], [D-Kane-Stewart'15])
- Higher Dimensions
- Tradeoffs between sample size and computational efficiency

Thank you for your attention!