Learning Sums of Independent Integer Random Variables

Costis Daskalakis (MIT)
Ilias Diakonikolas (Edinburgh)
Ryan O’Donnell (CMU)
Rocco Servedio (Columbia)
Li-Yang Tan (Columbia)

FOCS 2013, Berkeley CA
learning discrete distributions

Probability distributions on $[N] = \{0, 1, \ldots, N\}$

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

Goal: w.p. $\geq \frac{9}{10}$ output $\mathcal{D}'$ satisfying

$$d_{TV}(\mathcal{D}, \mathcal{D}') := \frac{1}{2} \|\mathcal{D} - \mathcal{D}'\|_1 \leq \varepsilon$$
analyses with PAC learning Boolean functions

- Class $\mathcal{C}$ of distributions
- Unknown target $\mathcal{D} \in \mathcal{C}$
- Learner gets i.i.d. samples from $\mathcal{D}$
- Output approximation $\mathcal{D}'$ of $\mathcal{D}$

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

Explicit emphasis on computational efficiency
learning distributions: an easy upper bound

Learning *arbitrary* distributions:
\[ \Theta\left(\frac{N}{\varepsilon^2}\right) \] samples necessary and sufficient

When can we do better?
Which distributions are easy to learn, which are hard?
two types of structured distributions

- Distributions with “shape restrictions”
  - log-concave
  - monotone
- Simple combinations of simple distributions
  - Mixtures of simple distributions
  - Mixtures of Gaussians

This work: Sums of independent, simple random variables
One piece of terminology

$k$-IRV: Integer-valued Random Variable supported on $\{0, 1, \ldots, k - 1\}$

$k$-SIIRV: Sum of $n$ Independent (not necessarily identical) $k$-IRVs
starting small

Simplest imaginable learning problem:
Learning 2-IRVs

\( \Theta(1/\varepsilon^2) \) samples necessary and sufficient

Learning 2-SIIRVs:
Sums of \( n \) independent coin flips with distinct biases?

\( \tilde{O}(1/\varepsilon^3) \) samples, independent of \( n \! \)!

Daskalakis, Diakonikolas, Servedio [STOC 2012]

[Defined by Poisson in 1837]
Learning $k$-IRVs: $\Theta\left(\frac{k}{\varepsilon^2}\right)$ samples necessary and sufficient.

Learning $k$-SIIRVs: 
Sum of $n$ independent die rolls, each with distinct biases, in $o(n)$ time? 

Our main result: Yes!

$\text{poly}(k, 1/\varepsilon)$ time and sample complexity, independent of $n$. 

more ambitious
from 2 to $k$ : a whole new ball game

Even just 3-SIIRVs have significantly richer structure than 2-SIIRVs

2-SIIRVs : unimodal, log-concave, close to Binomial

3-SIIRVs :

$\Omega(1)$-far from unimodal
$\Omega(1)$-far from log-concave
$\Omega(1)$-far from Binomial

Prior to our work nothing known, even about sample complexity, even for 3-SIIRVs.
our main theorem

**Theorem.** Let $C$ be the class of $k$-SIIRVs, i.e. all distributions

$$S = X_1 + \ldots + X_n$$

where $X_i$'s are independent, distinct r.v.'s supported on $\{0, 1, \ldots, k - 1\}$. There is an algorithm that learns $C$ with time and sample complexity $\text{poly}(k, 1/\varepsilon)$, independent of $n$.

Recall: $\Omega(k/\varepsilon^2)$ samples necessary even for a single $k$-IRV
our main technical contribution

A new limit theorem for $k$-SIIRVs:

“Every $k$-SIIRV is close to sum of two simple random variables”
**Limit Theorem.** Let $S$ be a $k$-SIIRV with $\text{Var}[S] \geq \text{poly}(k/\varepsilon)$. Then $S$ is $\varepsilon$-close to $cZ + Y$, where

- $c \in \{1, 2, \ldots, k - 1\}$
- $Z = \text{discretized normal}$
- $Y = c\text{-IRV}$

$Y, Z$ independent.
Previous limit theorems

Existing \( k \)-SIIRV limit theorems:

Certain highly \textit{structured} \( k \)-SIIRVs close to discretized normals

\[ S = X_1 + \ldots + X_n \approx Z \]

structure = “shift-invariance” of \( X'_i S \)

But general \( k \)-SIIRVs can be far from any disc. norm. \( Z \)

\textbf{Goal:} limit theorem for \textit{arbitrary} \( k \)-SIIRVs
$k$-SIIRVs can be far from $\mathbb{Z}$

Trivial but illustrative example:

$$S = X_1 + \ldots + X_n, \quad \text{all } X_i \text{ uniform over } \{0, 2, 4, \ldots, k\}$$

Our main contribution:

Build on and generalize existing limit theorems to characterize structure of all $k$-SIIRVs

$$d_{TV}(S, Z) \geq \frac{1}{2}$$
for all disc. norm. $Z$

Cause for optimism?

$$d_{TV}(S, 2Z) \leq \epsilon$$
two kinds of numbers

Heavy numbers: \( \sum_{i=1}^{n} \Pr[X_i = b] \) large

Light numbers: \( \sum_{i=1}^{n} \Pr[X_i = b] \) small

S

X_1

X_2

X_n

structured global component

C Z

Y

arbitrary local component
a useful special case: *all* numbers heavy

\[ S \]

\[
\begin{align*}
X_1 & = 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \\
X_2 & = 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \\
X_n & = 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \\
\end{align*}
\]

**Intuition:** No “mod structure” in \( S \)

E.g. \( S \) equally likely to be 0 or 1 mod 2

Use [Chen-Goldstein-Shao 2011] limit theorem to establish closeness to discretized normal

\[ d_{TV}(S, Z) \leq \varepsilon \]
a sampling procedure for $k$-SIIRVs

$\{3, 5\}$ heavy, $\{0,1,2,4\}$ light

1. Decide independently for each $X_i$ whether outcome will be heavy or light.

2. Draw either $X_i^h$ or $X_i^l$ according to respective conditional distributions.
Every outcome $\mathcal{O}$ of Stage 1 induces distribution

$$S_{\mathcal{O}} = \sum_{i \in \text{heavy}(\mathcal{O})} X_i^h + \sum_{j \in \text{light}(\mathcal{O})} X_j^l$$

$S = \text{mixture of } 2^n \text{ many } S_{\mathcal{O}'}$s

Key technical lemma:
With high probability over outcomes $\mathcal{O}$

$$\sum_{i \in \text{heavy}(\mathcal{O})} X_i^h \approx c Z$$

where $Z = \text{disc. norm. independent of } \mathcal{O}$.

- Proof uses “all numbers heavy” special case
- $c = \gcd(\text{heavy numbers})$
using the limit theorem to learn

**Limit Theorem.** Let $S$ be a $k$-SIIRV with $\text{Var}[S] \geq \text{poly}(k/\varepsilon)$. Then $S$ is $\varepsilon$-close to $cZ + Y$, where

- $c \in \{1, 2, \ldots, k - 1\}$
- $Z = \text{discretized normal}$
- $Y = c\text{-IRV}$

$Y, Z$ independent

- If $\text{Var}[S] \leq \text{poly}(k/\varepsilon)$, $S$ is close to sparse. Easily learn by “brute force”.
- Else guess $c \in \{1, 2, \ldots, k - 1\}$
- For each $c$, learn $Y$ and $Z$ separately.
- Do hypothesis testing over all $k$ possibilities.
summary of contributions

1. A limit theorem for $k$-SIIRVs

$$S \approx c Z + Y$$

structured

global component

2. Efficient algorithm for learning $k$-SIIRVs

$\text{poly}(k, 1/\varepsilon)$ time and samples.
thank you!