Learning and Covering Sums of Independent Random Variables with Unbounded Support

Alkis Kalavasis
NTUA

Konstantinos Stavropoulos
UT Austin

Manolis Zampetakis
UC Berkeley

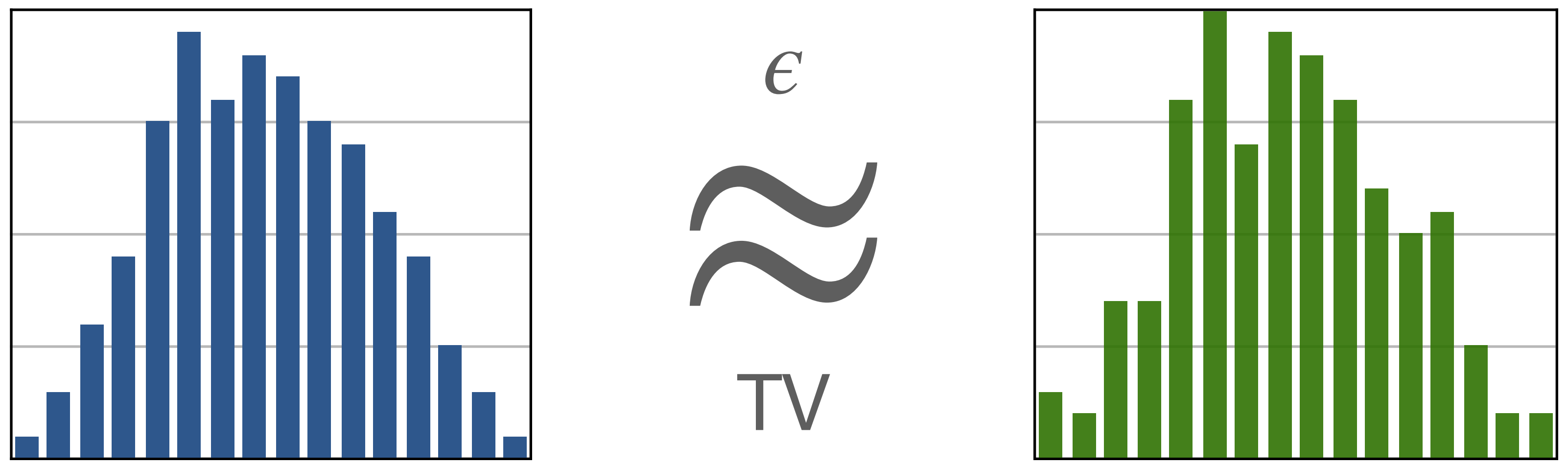
Setup

Sums of Independent Integer Random Variables (SIIRVs)

We focus on a fundamental specific type of integer random variables:

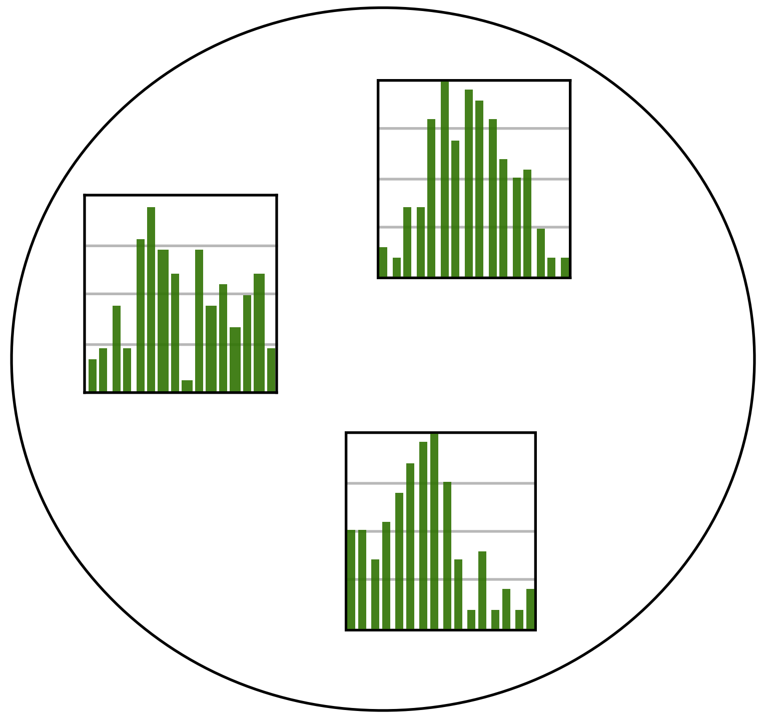
$$\sum_{i=1}^{n} X_i \text{ with independent, integer valued terms}$$

Tasks: 1. *Density estimation* 2. *Sparse Covering*



Distribution of Samples
(from the sum)

Output Distribution



Small Number of
Representatives

Setup

Sums of Independent Integer Random Variables (SIIRVs)

We focus on a fundamental specific type of integer random variables:

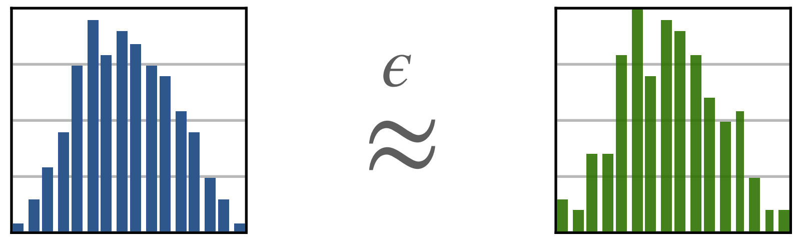
$$\sum_{i=1}^{n} X_i \text{ with independent, integer valued terms}$$

Tasks:

1. *Density estimation*

Given m i.i.d samples from $X = \sum_{i=1}^{n} X_i$, output Y such that

$$d_{TV}(Y, X) \leq \epsilon$$

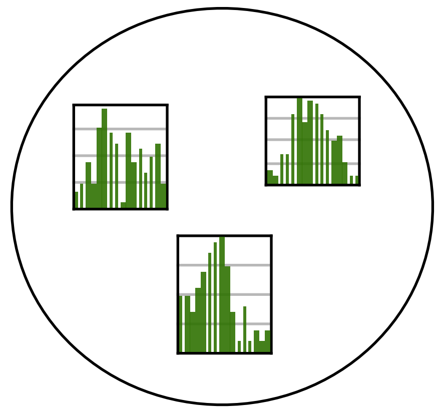


Distribution of Samples (from the sum) $\overset{\epsilon}{\approx}$ Output Distribution

2. *Sparse Covering*

For a family $\mathscr{F}$ of SIIRVs, identify a small set of distributions $\mathscr{C}$ ($|\mathscr{C}| < \infty$) so that for any $X \in \mathscr{F}$, there exists $Y \in \mathscr{C}$ with

$$d_{TV}(Y, X) \leq \epsilon$$



$\mathscr{C}$

Setup

Sums of Independent Integer Random Variables (SIIRVs)

We focus on a fundamental specific type of integer random variables:

$$\sum_{i=1}^{n} X_i \text{ with independent, integer valued terms}$$

Tasks: 1. *Density estimation* 2. *Sparse Covering*

Challenge 1: $m = \Theta_n(1)$
Sample Complexity
independent from n

Challenge 2: $\mathscr{C} \subseteq \mathscr{F}$
Representatives are members of
the considered family of SIIRVs
(proper covering)

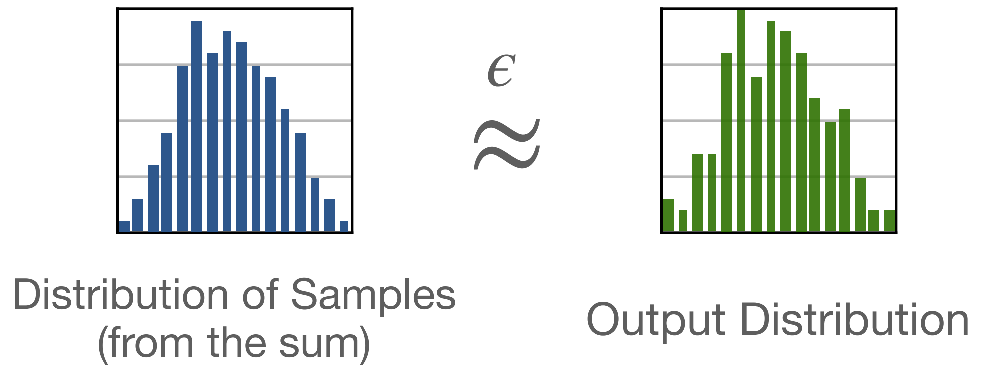
Motivation

Sums of Independent Integer Random Variables (SIIRVs)

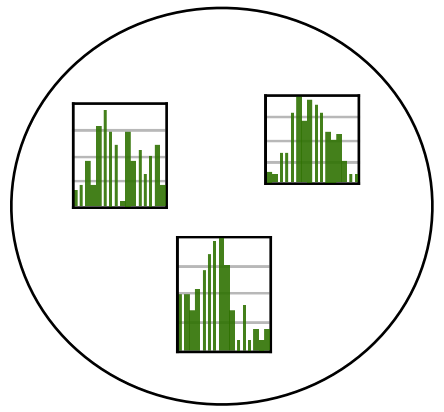
We focus on a fundamental specific type of integer random variables:

$$\sum_{i=1}^{n} X_i \text{ with independent, integer valued terms}$$

Challenge 1.
Sample Complexity
independent from n



Challenge 2.
Representatives are
SIIRVs themselves
(proper covering)



Applications (of challenges 1 *and* 2) in:

Mechanism Design:
Designing Auctions
[GT15]

Game Theory:
Computing Equilibrium in
Anonymous Games
[DDKT16],[DKS16],
[GT17],[CDS17]

Stochastic Optimization
[De18]

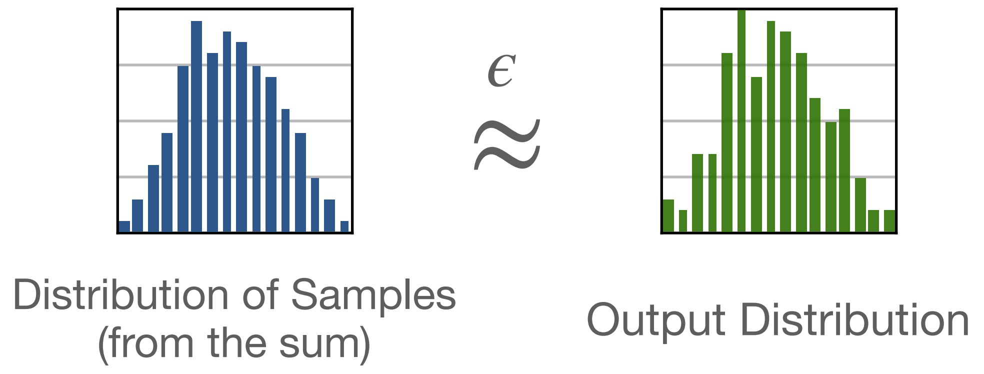
Previous results

Sums of Independent Integer Random Variables (SIIRVs)

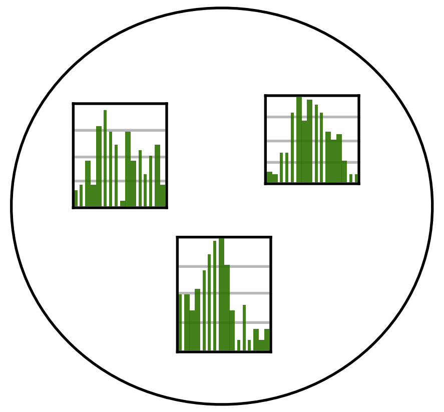
We focus on a fundamental specific type of integer random variables:

$$\sum_{i=1}^{n} X_i \text{ with independent, integer valued terms}$$

Challenge 1.
Sample Complexity
independent from n



Challenge 2.
Representatives are
SIIRVs themselves
(proper covering)



Restrict the distributions of the terms X_i to be

Bernoulli:
Learning and Covering
Poisson Binomial Distributions
[DP15, DDS15, DKS16]

poly$(1/\epsilon)$ samples & *proper* sparse covers

Supported on $\{0,\ldots,m-1\}$:
Bounded Support
[DDO+13, DKS16]

poly(m/ϵ) samples & sparse covers

High dimensional (not integer), again with bounded support:
[DKT15, DDKT16, DKS16]

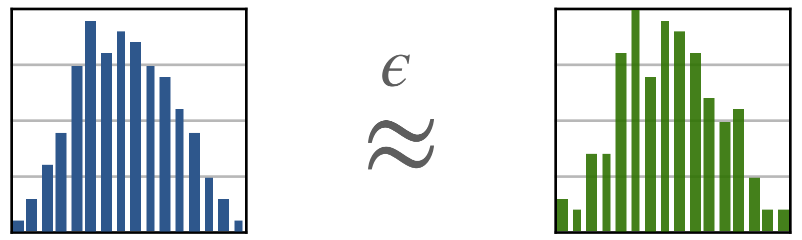
Previous results

Sums of Independent Integer Random Variables (SIIRVs)

We focus on a fundamental specific type of integer random variables:

$$\sum_{i=1}^{n} X_i \text{ with independent, integer valued terms}$$

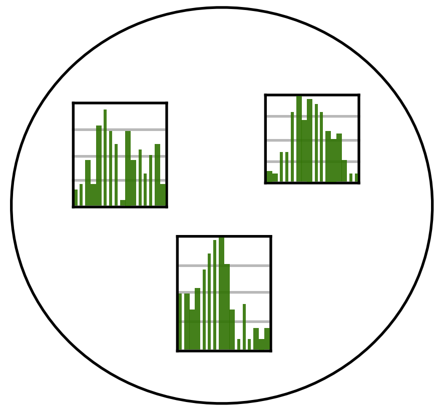
Challenge 1.
Sample Complexity independent from n



Distribution of Samples (from the sum)

Output Distribution

Challenge 2.
Representatives are SIIRVs themselves (proper covering)



However, *in the worst case*, if the terms X_i have:

Collective support of size $\geq$ 4:
The sample complexity scales with the maximum value of the support
[DLS18]

Unbounded support:
The sample complexity scales (polynomially) with n
[DDO+13]

Our results

Warm-up: Addressing Challenge 1 for "nice" unbounded distributions

We focus on a specific type of integer random variables:

$\sum_{i=1}^{n} X_i$ with independent, integer valued terms

&

Assumption 1.

Each term X_i is "nice", i.e.,

1. Unimodal & far from deterministic
2. Modes within a bounded region
3. Bounded fourth central moment

Theorem. Under Assumption 1, the distribution of an unknown SIIRV can be estimated up to error ϵ in statistical distance, using poly$(1/\epsilon)$ independent samples from the sum. (challenge 1)

Moreover, any family of SIIRVs that satisfy Assumption 1, can be ϵ-covered in statistical distance by the union of a collection of $2^{\text{poly}(1/\epsilon)}$ SIIRVs with the set of Discretized Gaussian random variables.

Our results

Main Result: SIIERVs

We focus on a specific type of integer random variables:

$\sum_{i=1}^{n} X_i$ with independent, integer valued terms

&

Assumption 2.

Each term X_i is belongs in a given “nice” exponential family $\mathscr{E}$ with k parameters.

Remark: we then call the sum an $\mathscr{E}$-SIIERV

Theorem. Under Assumption 2, the distribution of an unknown $\mathscr{E}$-SIIERV can be estimated up to error ϵ in statistical distance, using $\tilde{O}(k/\epsilon^2)$ independent samples from the sum. The output of the learning algorithm is itself an $\mathscr{E}$-SIIERV. (challenge 1)

Moreover, the family of $\mathscr{E}$-SIIERVs admits a proper sparse cover of size $2^{\tilde{O}(k/\epsilon^2)} + n^2 \cdot O(1/\epsilon)^k$. (challenge 2)

Proof Ingredients

- Sparsely covering a "nice" exponential family $\mathscr{E}$. *(case $n = 1$).*

  Geometric properties of polyhedral cones $\Rightarrow$ Bound the range of parameters

  Covering the bounded version of the parameter space $\Rightarrow$ Covering $\mathscr{E}$

- (Proper) structural results for $\mathscr{E}$-SIIERVs.

  Sparse Case (small n): Use sparse covers for $\mathscr{E}$.

  Dense Case (large n): Use appropriate Berry-Esseen type bound, unimodality and continuity of moments.

- Learning $\mathscr{E}$-SIIERVs.

  Given structural results, carefully apply standard methods to design a learning algorithm.

Proof Ingredients

- Sparsely covering a “nice” exponential family $\mathscr{E}$. *(case $n = 1$).*

  Geometric properties of polyhedral cones $\Rightarrow$ Bound the range of parameters

  Covering the bounded version of the parameter space $\Rightarrow$ Covering $\mathscr{E}$

- (Proper) structural results for $\mathscr{E}$-SIIERVs.

  Sparse Case (small n): Use sparse covers for $\mathscr{E}$.

  Dense Case (large n): Use appropriate Berry-Esseen type bound, unimodality and continuity of moments.

- Learning $\mathscr{E}$-SIIERVs.

  Given structural results, carefully apply standard methods to design a learning algorithm.

Thank you!