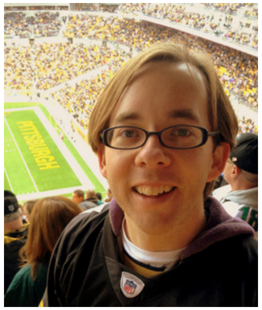
Learning Sums of Independent Integer Random Variables

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

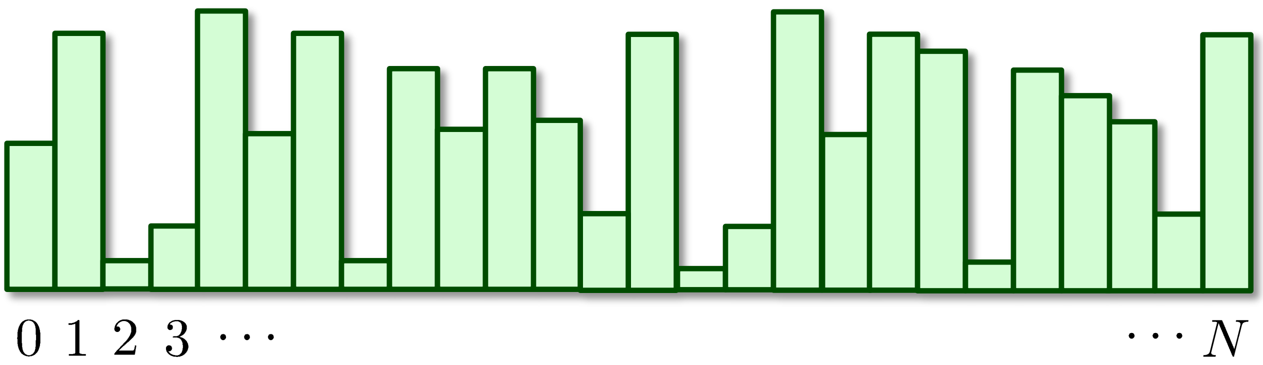





FOCS 2013, Berkeley CA

learning discrete distributions

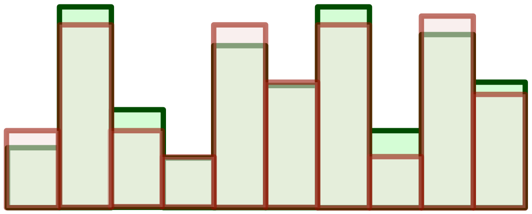
Probability distributions on $[N] = \{0, 1, \ldots, N\}$



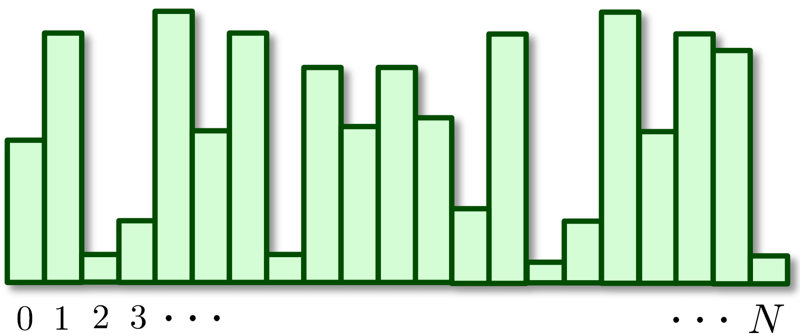
- Learning problem defined by class $\mathcal{C}$ of distributions
- Target distribution $\mathcal{D} \in \mathcal{C}$ unknown to learner
- Learner given sample of i.i.d. draws from $\mathcal{D}$

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

$$d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') := \frac{1}{2}\|\mathcal{D} - \mathcal{D}'\|_1 \leq \varepsilon$$



analogies with PAC learning Boolean functions



x	$f(x)$
10101010010	1
10111111110	1
10101010000	0
⋮	⋮

- 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 $\mathcal{D}$

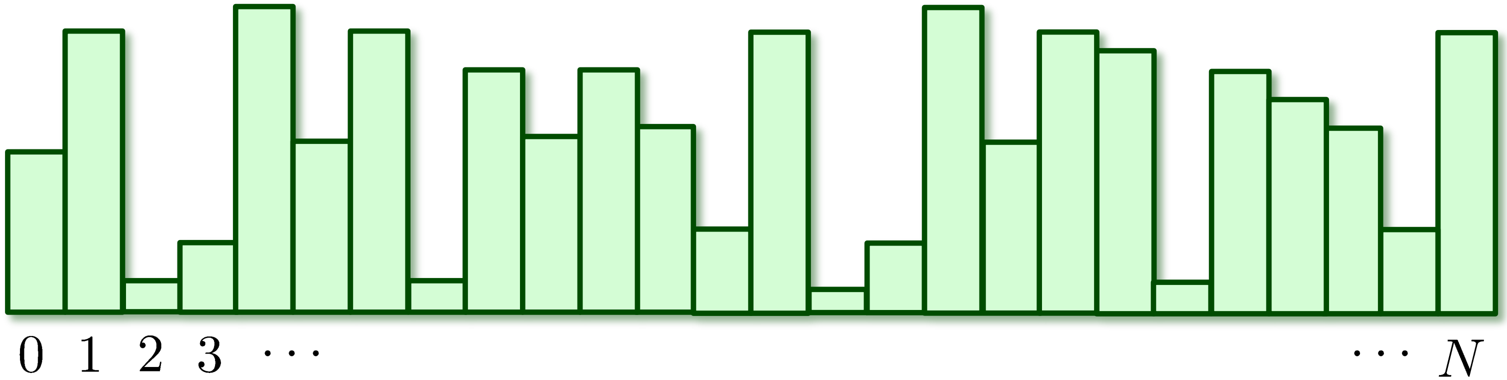
- Class $\mathcal{C}$ of **Boolean functions**
- Unknown target $f \in \mathcal{C}$
- Learner gets **labeled samples** $\langle x, f(x) \rangle$
- Output approximation f' of f

Explicit emphasis on *computational efficiency*

learning distributions: an easy upper bound

Learning *arbitrary* distributions:

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

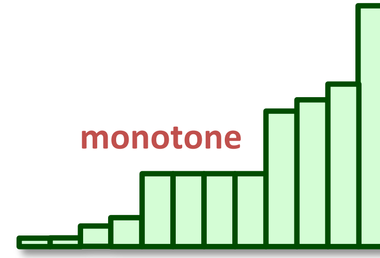


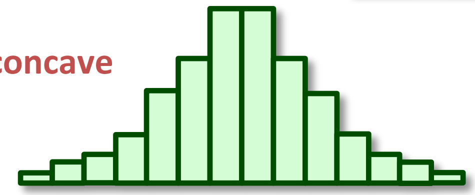
When can we do better?

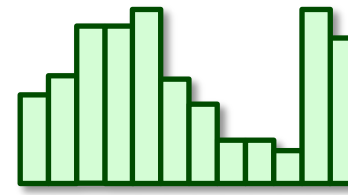
Which distributions are easy to learn, which are hard?

two types of structured distributions

- Distributions with "shape restrictions"

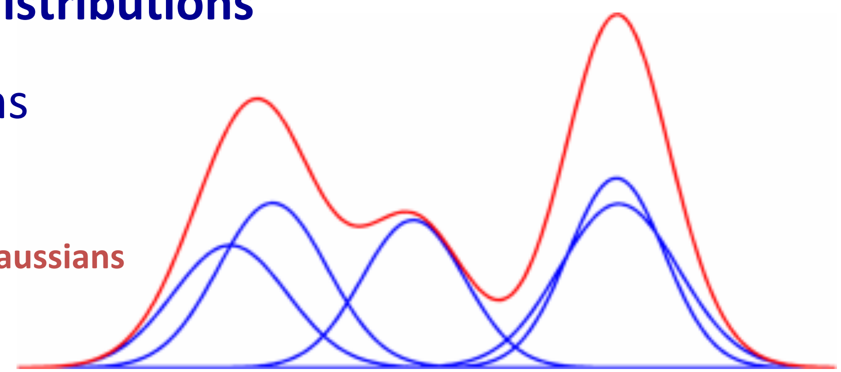




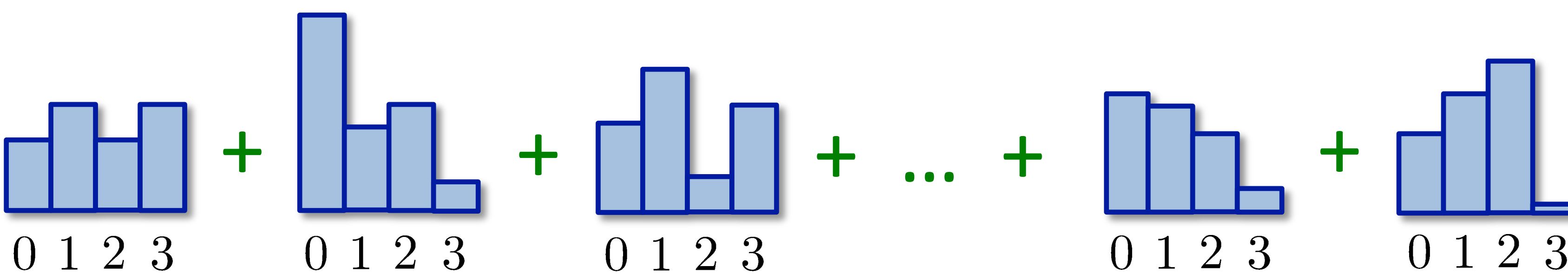


- ***Simple combinations* of *simple* distributions**

  Mixtures of simple distributions

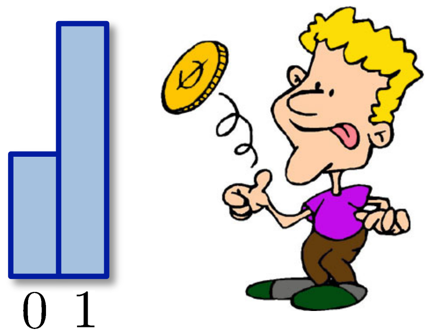


This work: *Sums* of independent, simple random variables

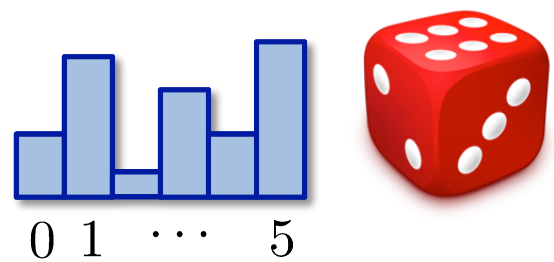


one piece of terminology

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

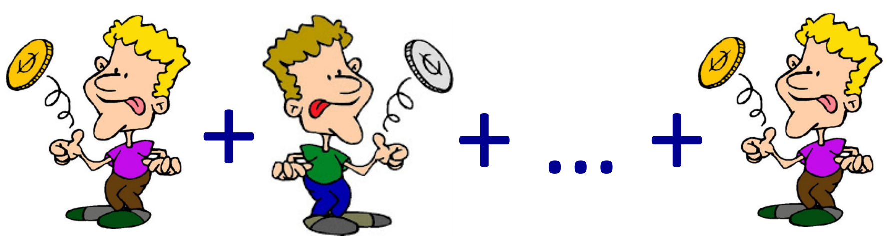


2-IRV



6-IRV

k-SIIRV: Sum of *n* Independent (*not necessarily identical*) *k*-IRVs



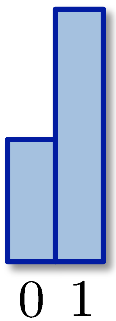
2-SIIRV

k-SIIRV

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?

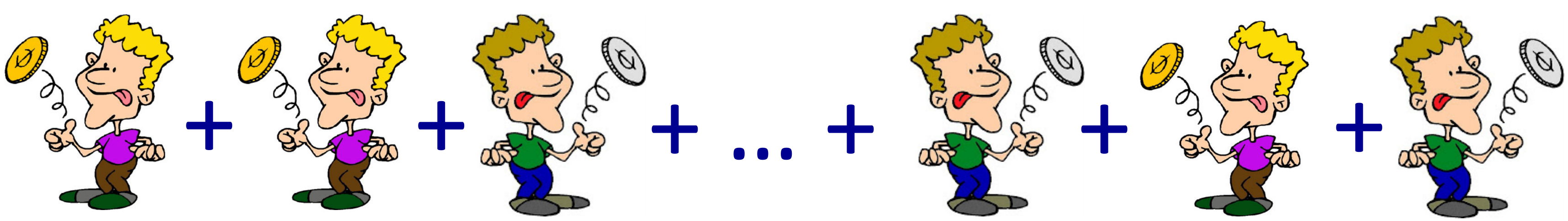


Daskalakis, Diakonikolas, Servedio [STOC 2012]

$\tilde{O}(1/\varepsilon^3)$

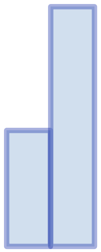
samples,

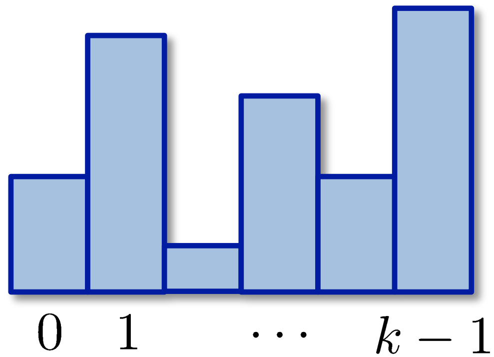
independent of *n*!



[Defined by Poisson in 1837]

more ambitious





Learning *k*-IRVs: $\Theta(k/\varepsilon^2)$ samples necessary and sufficient

Learning *k*-SIIRVs:

Sum of n independent die rolls, each with distinct biases, in $o(n)$ time?

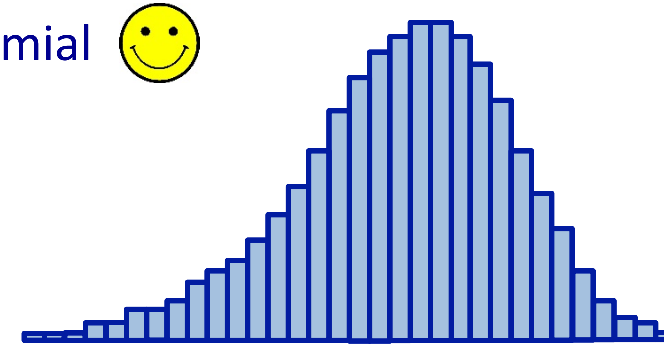
Our main result: **Yes!**

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

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

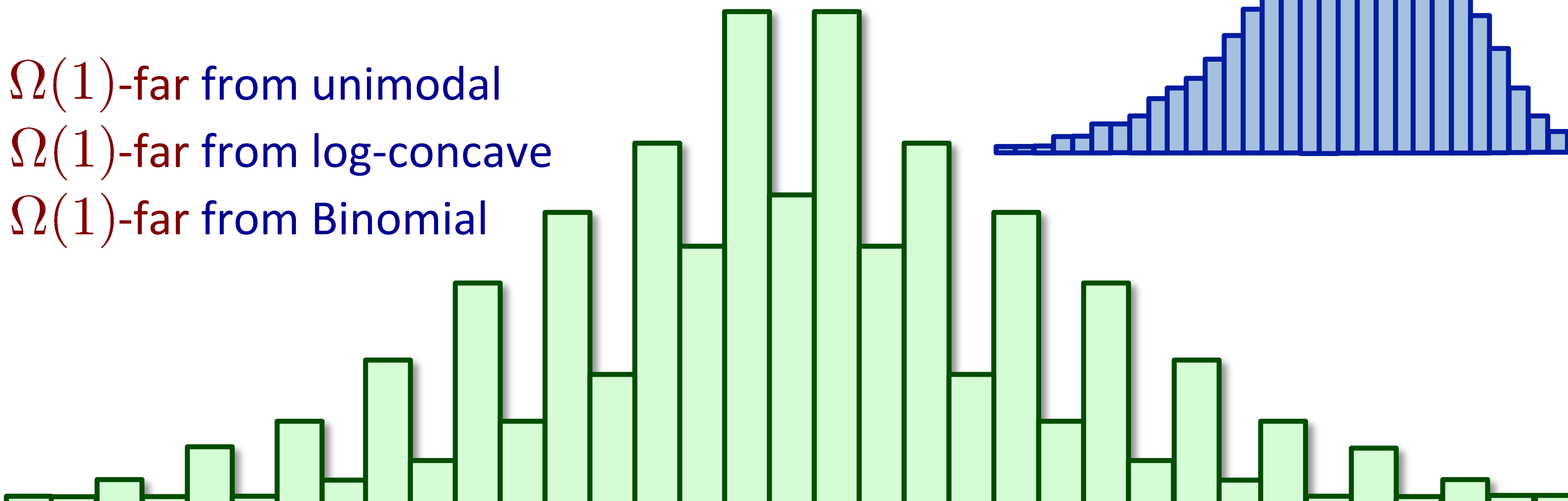


3-SIIRVs :

$\Omega(1)$-far from unimodal

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

$\Omega(1)$-far from Binomial



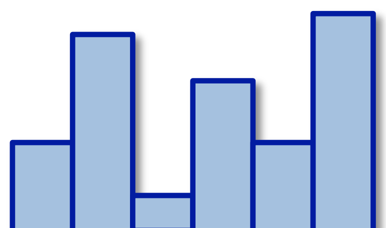
Prior to our work nothing known, even about sample complexity, even for 3-SIIRVs.

our main theorem

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

$$\mathbf{S} = \mathbf{X}_1 + \ldots + \mathbf{X}_n$$

where $\mathbf{X}_i' s$ are independent, distinct r.v.'s supported on $\{0, 1, \ldots, k-1\}$. There is an algorithm that learns $\mathcal{C}$ with time and sample complexity $\mathrm{poly}(k, 1/\varepsilon)$, 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:

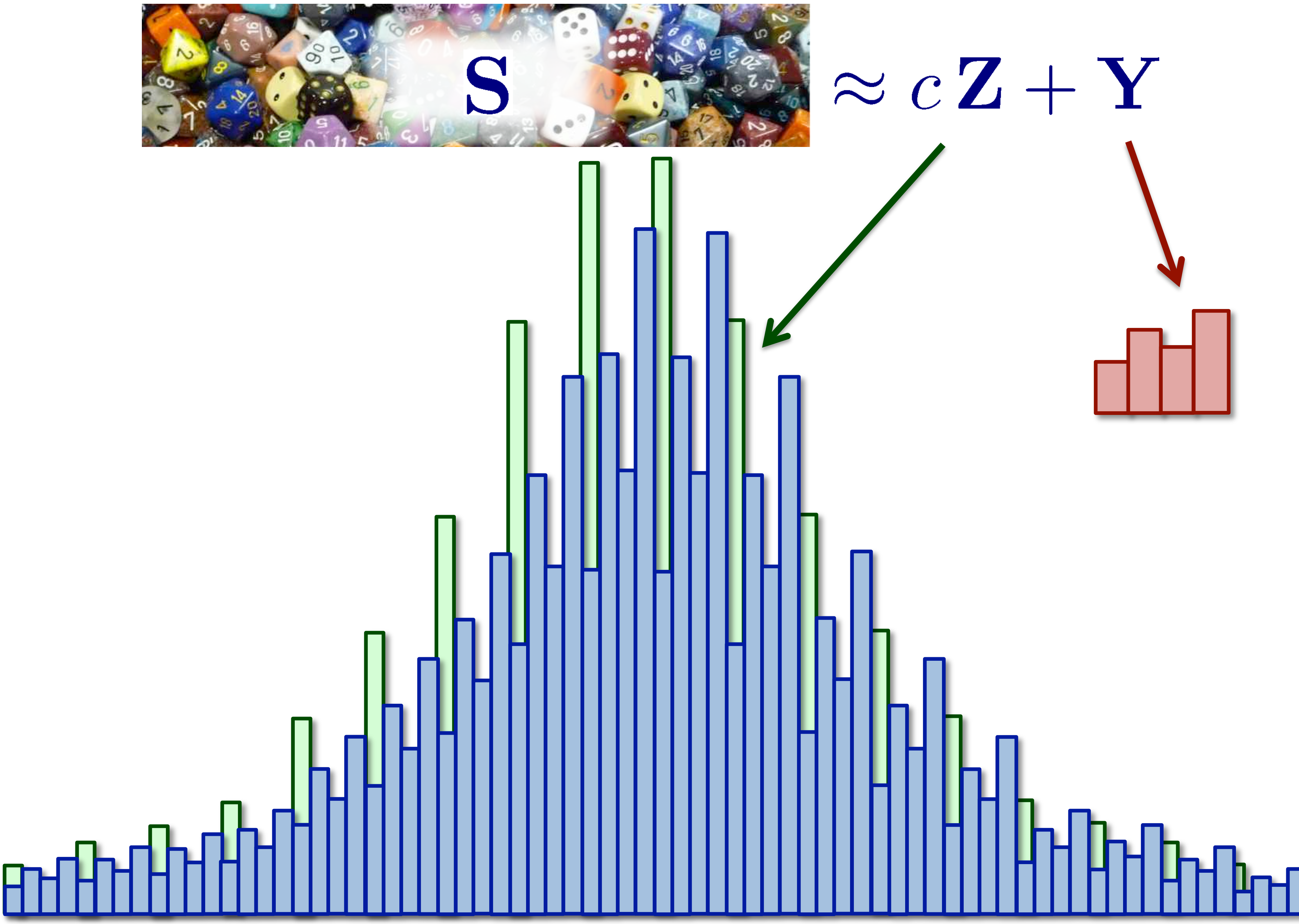
“Every *k*-SIIRV is close to sum of two simple random variables”

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

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

$\mathbf{Y}, \mathbf{Z}$ independent

$$\approx c\,\mathbf{Z} + \mathbf{Y}$$

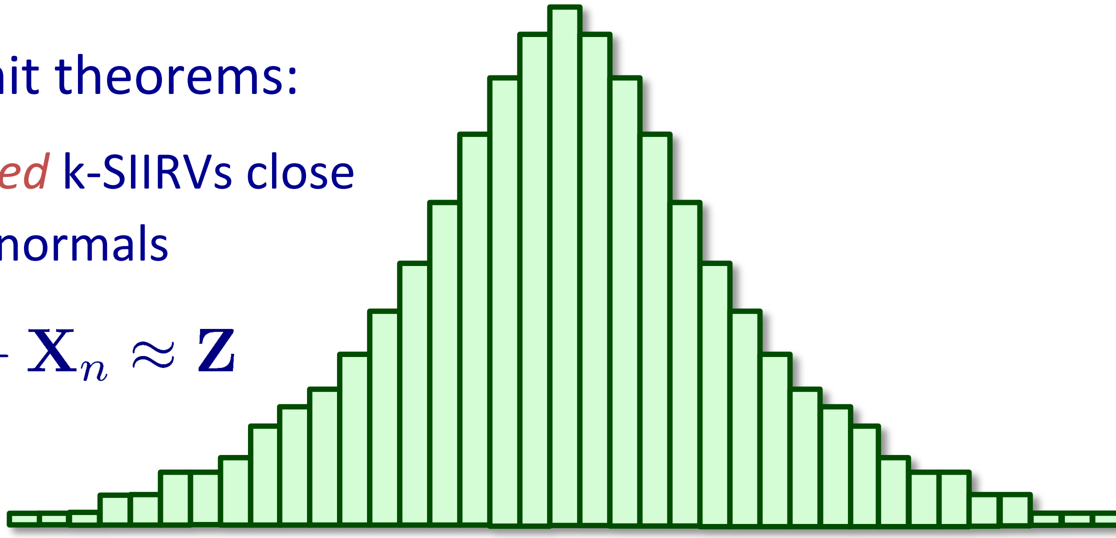

$\mathbf{S} \approx c\,\mathbf{Z} + \mathbf{Y}$

previous limit theorems

Existing *k*-SIIRV limit theorems:

Certain highly *structured* k-SIIRVs close to discretized normals

$$\mathbf{S} = \mathbf{X}_1 + \ldots + \mathbf{X}_n \approx \mathbf{Z}$$



structure = “shift-invariance” of $\mathbf{X}_i's$

But general *k*-SIIRVs can be far from any disc. norm. $\mathbf{Z}$

Goal: limit theorem for ***arbitrary*** k-SIIRVs

k-SIIRVs can be far from $\mathbf{Z}$

Trivial but illustrative example:

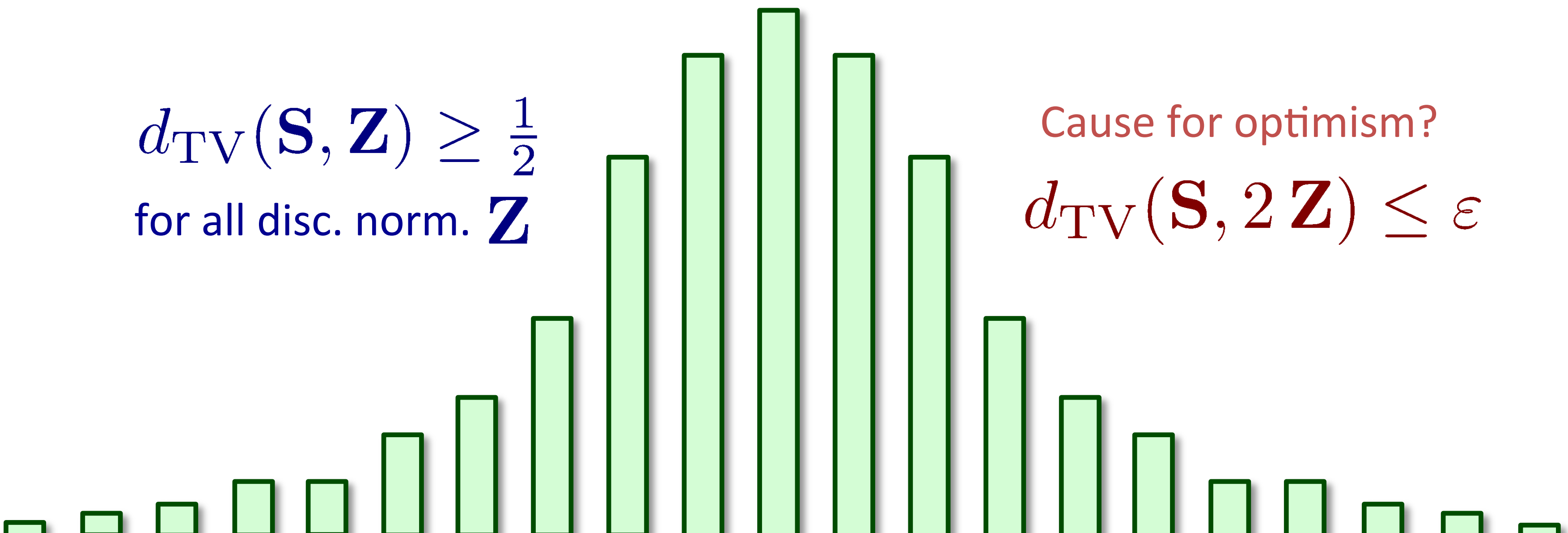
$$\mathbf{S} = \mathbf{X}_1 + \ldots + \mathbf{X}_n, \quad \text{all } \mathbf{X}_i \text{ uniform over } \{0, 2, 4, \ldots, k\}$$

Our main contribution:

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

$$d_{\mathrm{TV}}(\mathbf{S}, \mathbf{Z}) \geq \frac{1}{2}$$

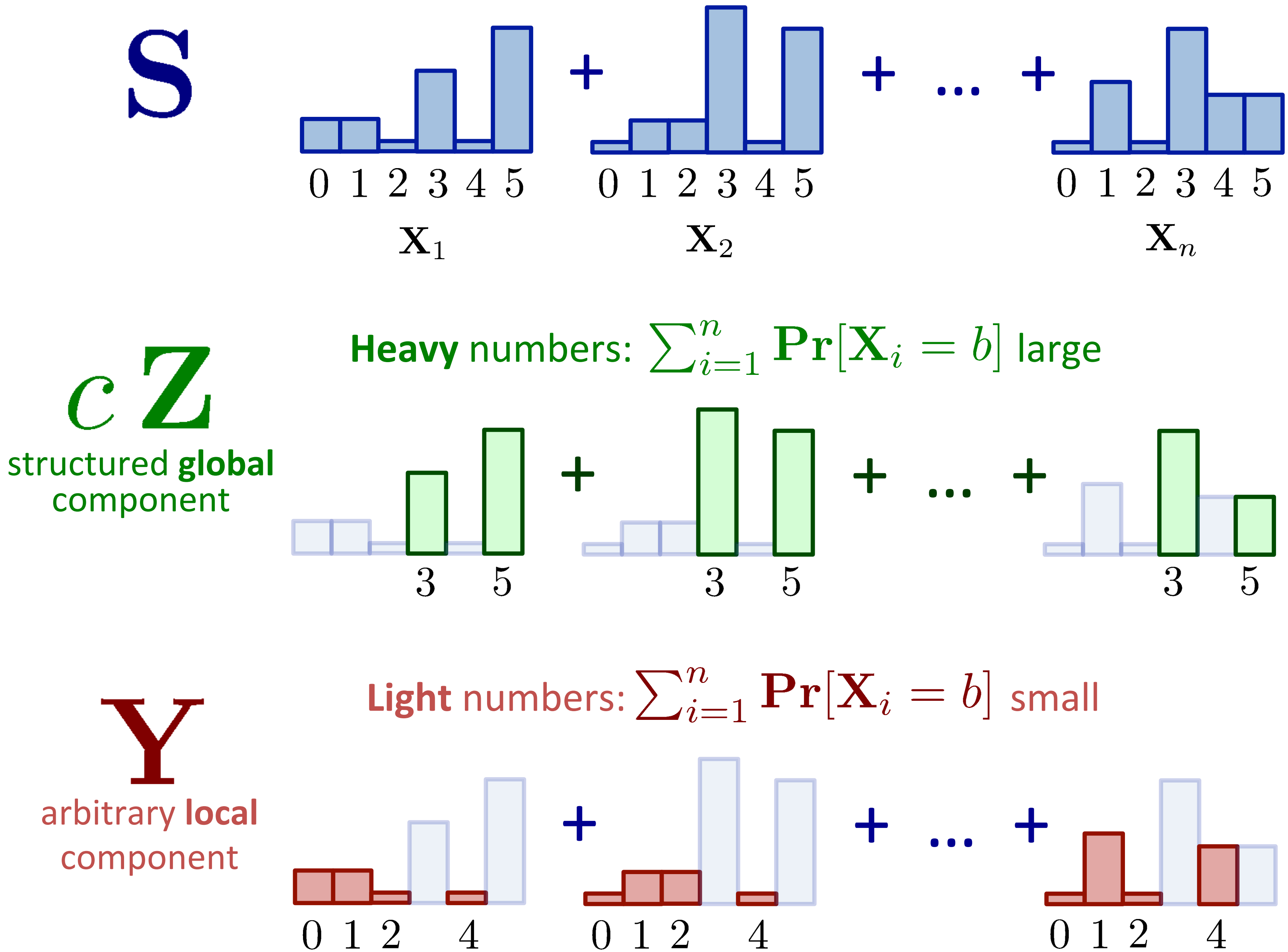
for all disc. norm. $\mathbf{Z}$



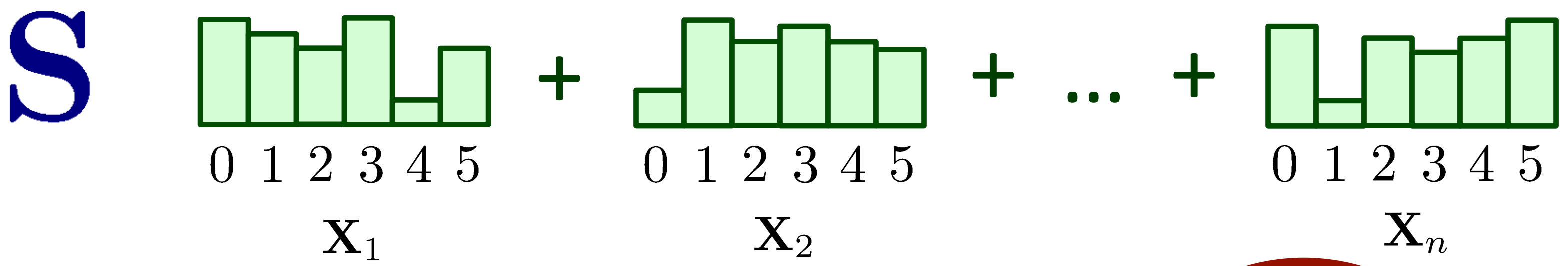
Cause for optimism?

$$d_{\mathrm{TV}}(\mathbf{S}, 2\,\mathbf{Z}) \leq \varepsilon$$

two kinds of numbers



a useful special case: *all* numbers heavy

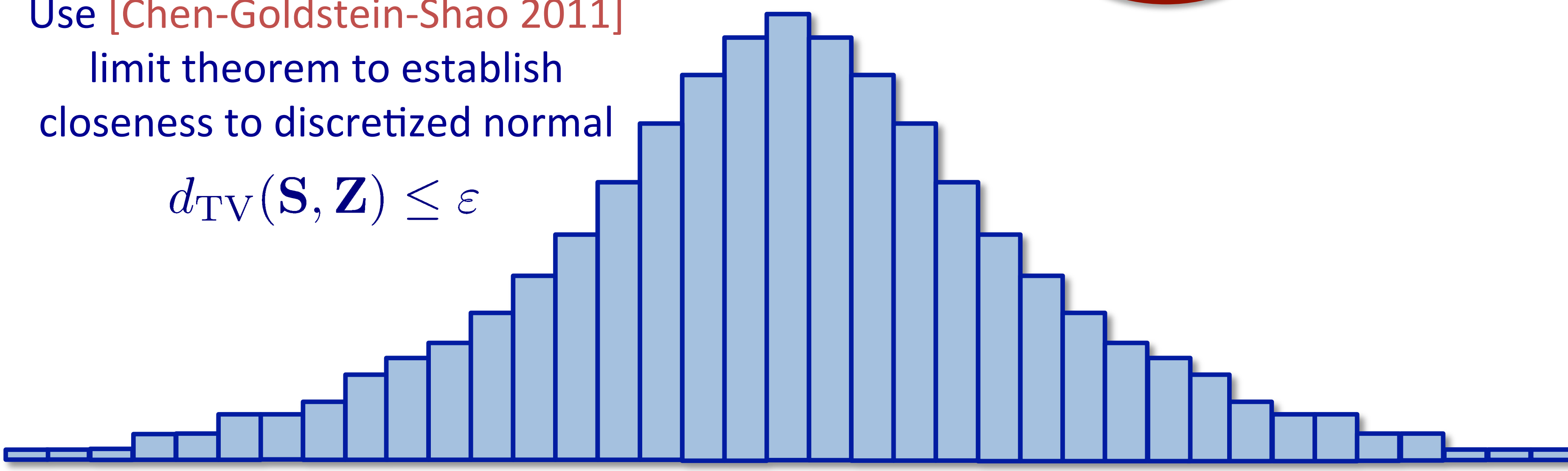


Intuition: No "mod structure" in $\mathbf{S}$

e.g. $\mathbf{S}$ equally likely to be 0 or 1 mod 2

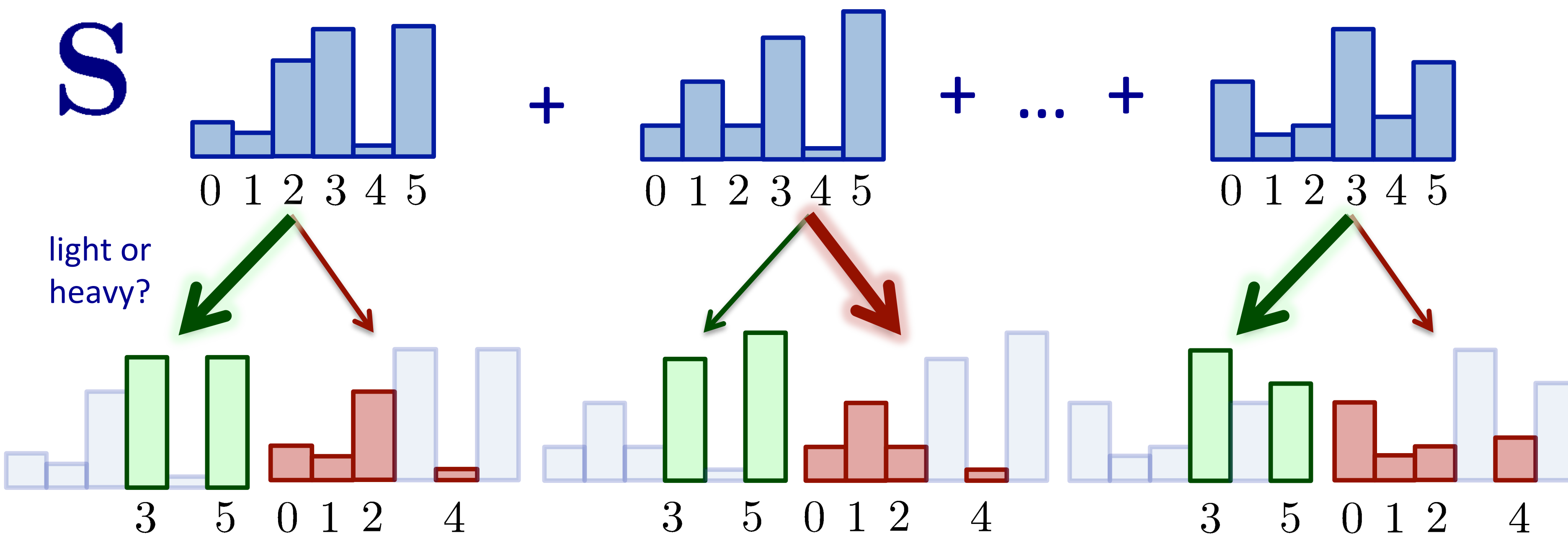
Use [Chen-Goldstein-Shao 2011] limit theorem to establish closeness to discretized normal

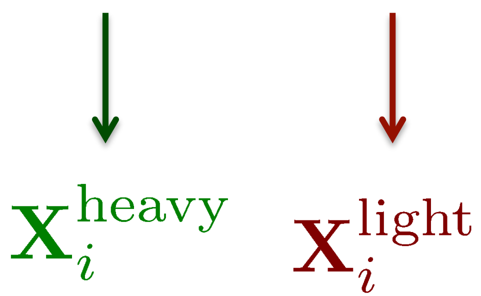
$$d_{\mathrm{TV}}(\mathbf{S}, \mathbf{Z}) \leq \varepsilon$$



a sampling procedure for *k*-SIIRVs

{3, 5} **heavy**, {0,1,2,4} **light**

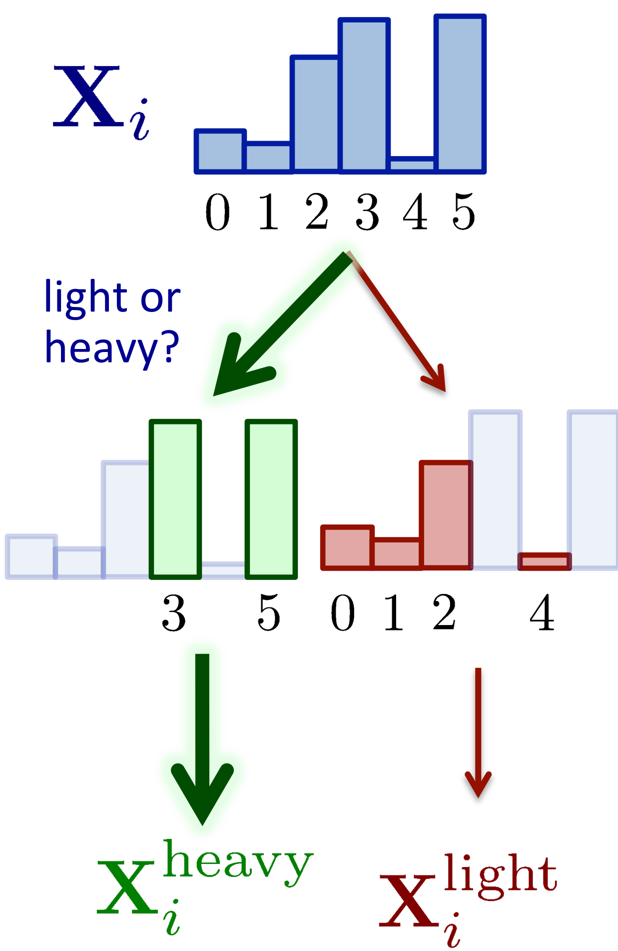




1. Decide independently for each $\mathbf{X}_i$ whether outcome will be heavy or light.

2. Draw either $\mathbf{X}_i^h$ or $\mathbf{X}_i^\ell$ according to respective conditional distributions.

analysis



Every outcome $\mathcal{O}$ of Stage 1 induces distribution

$$\mathbf{S}_{\mathcal{O}} = \sum_{i \in \text{heavy}(\mathcal{O})} \mathbf{X}_i^h + \sum_{j \in \text{light}(\mathcal{O})} \mathbf{X}_j^\ell$$

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

Key technical lemma:

With high probability over outcomes $\mathcal{O}$

$$\sum_{i \in \text{heavy}(\mathcal{O})} \mathbf{X}_i^h \approx c\,\mathbf{Z}$$

where $\mathbf{Z}$ = 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 $\mathbf{S}$ be a *k*-SIIRV with $\mathrm{Var}[\mathbf{S}] \geq \mathrm{poly}(k/\varepsilon)$.
Then $\mathbf{S}$ is ε-close to $c\,\mathbf{Z} + \mathbf{Y}$, where

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

$\mathbf{Y}$, $\mathbf{Z}$ independent

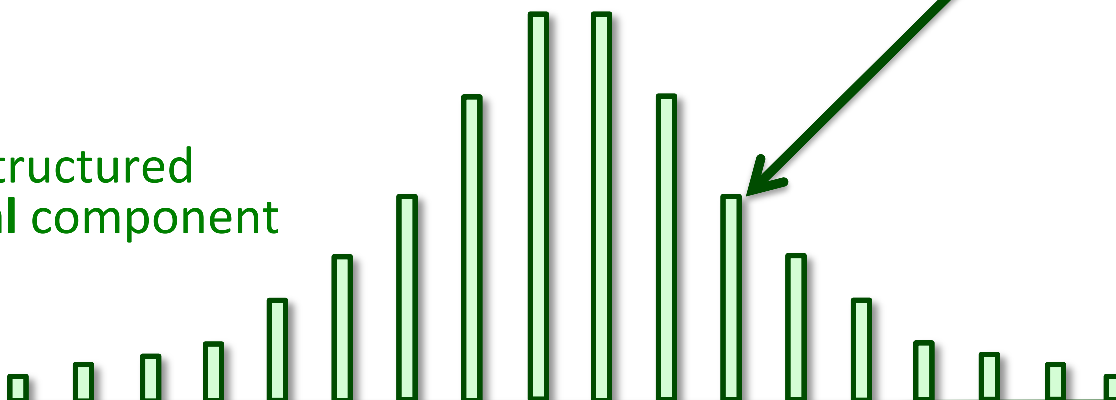
- If $\mathrm{Var}[\mathbf{S}] \leq \mathrm{poly}(k/\varepsilon)$, $\mathbf{S}$ is close to sparse.
  Easily learn by "brute force".
- Else guess $c \in \{1, 2, \ldots, k-1\}$
- For each c, learn $\mathbf{Y}$ and $\mathbf{Z}$ separately.
- Do hypothesis testing over all *k* possibilities.

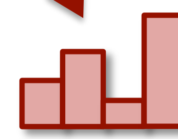
summary of contributions

1. A limit theorem for *k*-SIIRVs

$$\mathbf{S} \approx c\,\mathbf{Z} + \mathbf{Y}$$

structured
global component





arbitrary
local component

2. Efficient algorithm for learning *k*-SIIRVs

   $\mathrm{poly}(k, 1/\varepsilon)$ time and samples.


thank you!