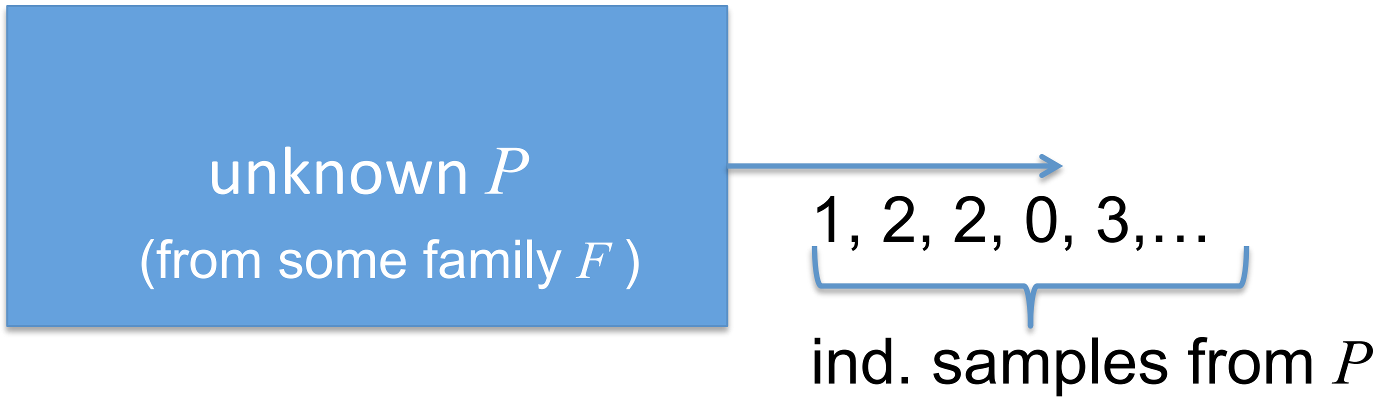
Beyond Berry-Esseen: *Structure* and *Learning* Sums of Random Variables

Constantinos Daskalakis
EECS, MIT

Distribution Learning Problem

- **Input: -** Sample access to distribution over $\{0,1,\ldots,n\}$



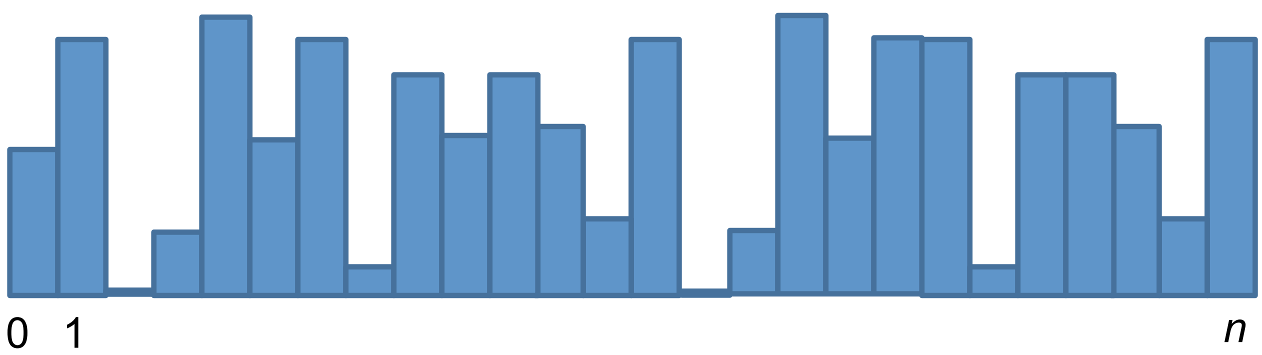
  – $\varepsilon > 0$

$$\frac{1}{2}\sum_{i=0}^{n}|P(i)-Q(i)|$$

- **Goal:**
  - Find some Q s.t. $d_{TV}(P, Q) \le \varepsilon$
  - **(proper learn)** Find $Q \in F$ s.t. $d_{TV}(P, Q) \le \varepsilon$
  - **Minimize number of samples, computation time**

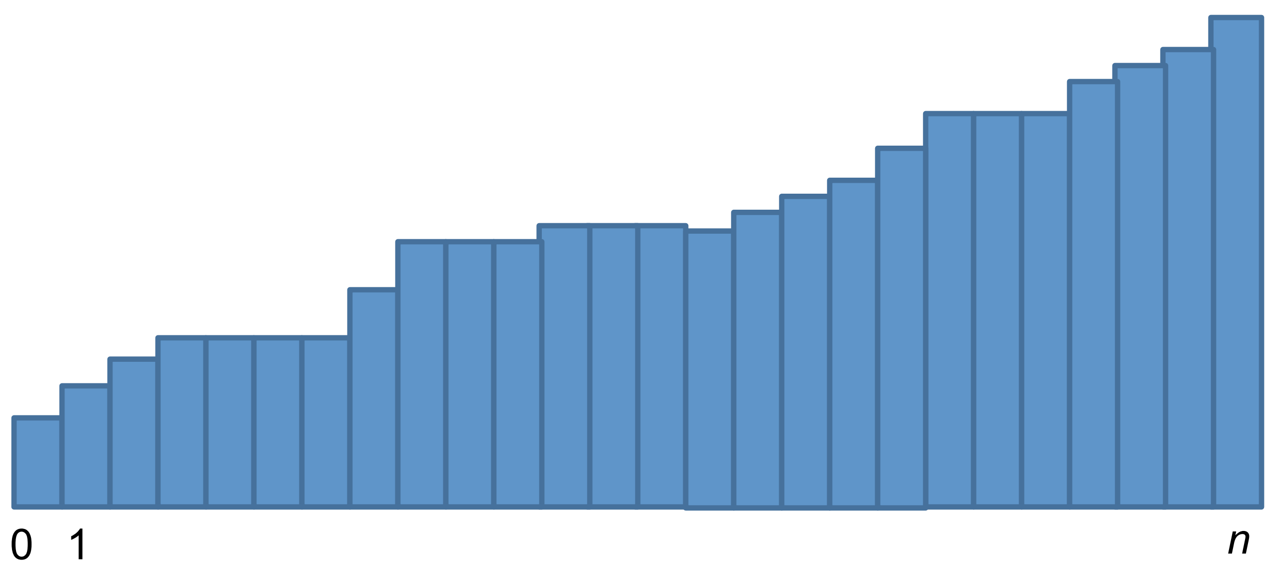
Distribution Learning Problem (cont'd)

- **Refresher:**
  - Arbitrary distribution over {0,…,*n*} requires time and sample complexity of $\Theta(n/\epsilon^2)$ (folklore)



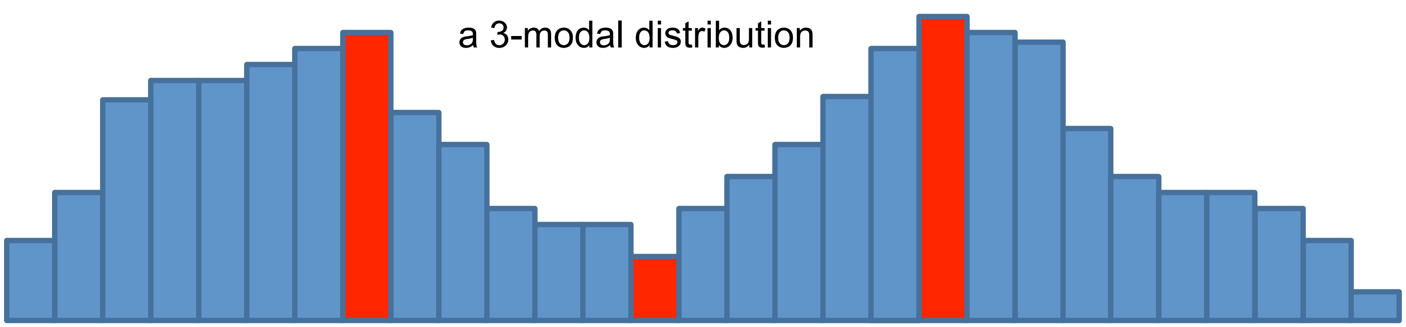
Distribution Learning Problem (cont'd)

- **Refresher:**
  - Arbitrary distribution over {0,…,*n*} requires time and sample complexity of $\Theta(n/\epsilon^2)$ (folklore)
  - Monotone distributions over {0,…,*n*} require time and sample complexity of $\Theta(\log n/\epsilon^3)$ [Birgé 1987]



Distribution Learning Problem (cont'd)

- **Refresher:**
  - Arbitrary distribution over {0,…,*n*} requires time and sample complexity of $\Theta(n/\epsilon^2)$ (folklore)
  - Monotone distributions over {0,…,*n*} require time and sample complexity of $\Theta(\log n/\epsilon^3)$ [Birgé 1987]
  - *k*-modal distributions over {0,…,*n*} can be learned from $O\left(\frac{k \log n}{\epsilon^3} + \frac{k^3 \log k/\epsilon}{\epsilon^3}\right)$ samples in time $\mathrm{poly}(k \log n/\epsilon)$ [D-Diakonikolas-Servedio 2012]



Distribution Learning Problem (cont'd)

- **Refresher:**
  - Arbitrary distribution over {0,…,n} requires time and sample complexity of $\Theta(n/\epsilon^2)$ (folklore)
  - Monotone distributions over {0,…,n} require time and sample complexity of $\Theta(\log n/\epsilon^3)$ [Birgé 1987]
  - k-modal distributions over {0,…,n} can be learned from $O\left(\frac{k \log n}{\epsilon^3} + \frac{k^3 \log k/\epsilon}{\epsilon^3}\right)$ samples in time $\mathrm{poly}(k \log n/\epsilon)$ [D-Diakonikolas-Servedio 2012]
  - Log-concave distributions $P(i)^2 > P(i-1)\, P(i+1)$ can be learned from $\tilde{O}(1/\epsilon^{2.5})$ samples [Chan-Diakonikolas-Servedio-Sun 2014]

Focus of This Talk: PBDs and SIIRVs

- **Def 1:** A Poisson Binomial Distribution (PBD) is
  - the distribution of the sum $X = \Sigma X_i$ of *n* independent r.v.'s $X_i \in \{0,1\}$
  - support: $\{0,1,\ldots,n\}$

Sharp structural results

- **Def 2:** A *k*-SIIRV is
  - the distribution of the sum $X = \Sigma X_i$ of *n* independent r.v.'s $X_i \in \{0,\ldots,k\text{-}1\}$
  - support: $\{0,1,\ldots,n\cdot(k\text{-}1)\}$

Learning from $\Theta(1/\epsilon^2)$ / respectively $\mathrm{poly}(k/\epsilon)$ samples

- **Objectives:** *Structure* and *Learning*

Menu

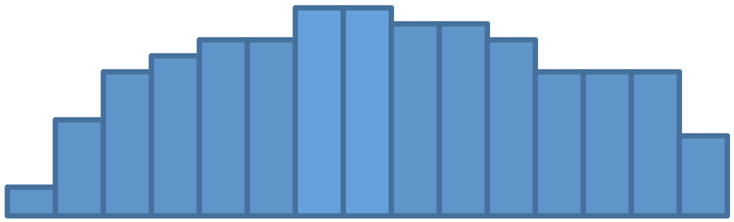
- *Refresher*
- *Objectives for this talk*
- PBD Structure and Learning
- *k*-SIIRV Structure and Learning

Menu

- *Refresher*
- *Objectives for this talk*
- **PBD Structure and Learning**
- *k*-SIIRV Structure and Learning

Structure of PBDs

- Unimodal; in fact log-concave distributions
  - so can be (non-properly) learned from $\tilde{O}(1/\epsilon^{2.5})$ samples



- [Berry 1941, Esseen 1942]: If $X_1,\ldots,X_n$ are independent and bounded then

$$d_K\left(\sum_i X_i, \mathcal{N}(\mu,\sigma^2)\right) \le C\frac{\sum_i \mathbb{E}[|X_i|^3]}{\sigma^3}$$

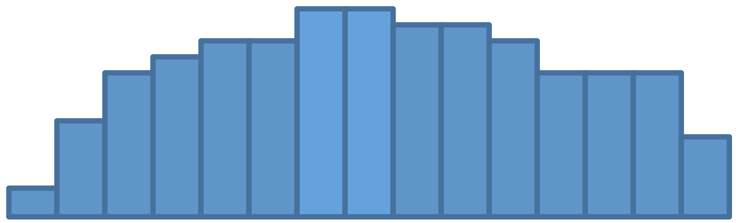
$$d_K(P,Q) = \max_i |P(\le i) - Q(\le i)|$$

$$\mu = \mathbb{E}\left[\sum_i X_i\right], \sigma^2 = \mathrm{Var}\left[\sum_i X_i\right]$$

$$\text{[Esseen 1956]} \quad 0.4097 \le C \le 0.5600 \quad \text{[Shevtsova 2010]}$$

Structure of PBDs

- Unimodal; in fact log-concave distributions
  - so can be (non-properly) learned from $\tilde{O}(1/\epsilon^{2.5})$ samples



- [Berry 1941, Esseen 1942]: If $X_1,\ldots,X_n$ are independent and bounded then

$$d_K\left(\sum_i X_i, \mathcal{N}(\mu,\sigma^2)\right) \le C\frac{\sum_i \mathbb{E}[|X_i|^3]}{\sigma^3}$$

- specializing to PBDs: $d_K\left(\sum_i X_i, N(\mu,\sigma^2)\right) \le C\frac{\mu}{\sigma^3}$
- e.g.

$$d_K(B(n,p), \mathcal{N}(np, np(1-p))) \le C\frac{1}{\sqrt{np}(1-p)^{1.5}}$$

quality of bound decays with $n \cdot p$ – poor if, eg, $p = 1/n$

Structure of PBDs

- [Berry 1941, Esseen 1942]: $d_K\left(\sum_i X_i, N(\mu,\sigma^2)\right) \le C\frac{\mu}{\sigma^3}$
- e.g. $d_K(B(n,p), \mathcal{N}(np, np(1-p)) \le C\frac{1}{\sqrt{np}(1-p)^{1.5}}$
- [Le Cam 1960]: $d_{\mathrm{TV}}\left(\sum_i X_i, \mathrm{Poisson}(\mu)\right) \le \sum_i p_i^2$
- e.g. $d_{\mathrm{TV}}(B(n,p), \mathrm{Poisson}(np)) \le np^2$
- good when, e.g., $p = 1/n$

$p_i = \mathbb{E}[X_i]$

Structure of PBDs

- [Berry 1941, Esseen 1942]: $d_K\left(\sum_i X_i, N(\mu,\sigma^2)\right) \le C\frac{\mu}{\sigma^3}$
- [Le Cam 1960]: $d_{\mathrm{TV}}\left(\sum_i X_i, \mathrm{Poisson}(\mu)\right) \le \sum_i p_i^2$
- [Chen-Goldstein-Shao 2011]: $d_{\mathrm{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) \le \frac{O(1)}{\sigma}$

rounded $\mathcal{N}(\mu,\sigma^2)$

Structure of PBDs

- [Berry 1941, Esseen 1942]: $d_K\left(\sum_i X_i, N(\mu,\sigma^2)\right) \le C\frac{\mu}{\sigma^3}$
- [Le Cam 1960]: $d_{\mathrm{TV}}\left(\sum_i X_i, \mathrm{Poisson}(\mu)\right) \le \sum_i p_i^2$
- [Chen-Goldstein-Shao 2011]: $d_{\mathrm{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) \le \frac{O(1)}{\sigma}$
- [Röllin 2007]: $d_{\mathrm{TV}}\left(\sum_i X_i, TP(\mu,\sigma^2)\right) \le \frac{1}{\sigma} + \frac{2}{\sigma^2}$

$\mathrm{Poisson}(\sigma^2 + \{\mu - \sigma^2\}) + \lfloor \mu - \sigma^2 \rfloor$

TP stands for "translated Poisson"

Structure of PBDs

- [Berry 1941, Esseen 1942]: $d_K\left(\sum_i X_i, N(\mu,\sigma^2)\right) \leq C\frac{\mu}{\sigma^3}$
- [Le Cam 1960]: $d_{\text{TV}}\left(\sum_i X_i, \text{Poisson}(\mu)\right) \leq \sum_i p_i^2$
- [Chen-Goldstein-Shao 2011]: $d_{\text{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) \leq \frac{O(1)}{\sigma}$
- [Röllin 2007]: $d_{\text{TV}}\left(\sum_i X_i, TP(\mu,\sigma^2)\right) \leq \frac{1}{\sigma} + \frac{2}{\sigma^2}$
- …
- Bounds *only use first two moments*
  - Question 1: Bounds for arbitrary approximation accuracy *ε* ?
  - Question 2: Distance of two PBDs with same first two moments?
- Approximating distributions *are from a different family*
  - Question 3: Are there meaningful *proper* approximations?

The first log(1/*ε*)-moments suffice

[D-Papadimitriou '09]: Let $X = \sum_i X_i$ and $Y = \sum_i Y_i$ be two PBDs s.t. $\mathbb{E}[X_i] \leq 1/2$ and $\mathbb{E}[Y_i] \leq 1/2$ for all *i.*

If $\mathbb{E}[X^\ell] = \mathbb{E}[Y^\ell], \forall \ell = 1, \ldots, d$

then: $d_{\mathrm{TV}}(X, Y) \leq 2^{-\Omega(d)}$.

Corollary: For all *ε>0*, agreement in the first log(1/*ε*) moments suffices for variation distance *ε*.

The Structure of PBDs

- S_n: set of all PBDs on n variables
- **[D-Papadimitriou '09]:** For all $\varepsilon > 0$, there exists a ***proper*** ε-cover $S_{n,\varepsilon} \subseteq S_n$ of size:

$$|S_{n,\epsilon}| \leq n^2 + n \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$$

$\propto$: integer multiple

$\forall P \in S_n, \exists Q \in S_{n,\epsilon}$ s.t. $d_{\mathrm{TV}}(P,Q) \leq \epsilon$

Naïve upper bound for cover size: $|S_{n,\epsilon}| \leq (\frac{n}{\epsilon})^n$

- obtained by discretizing every X_i so that its expectation is $\propto \frac{\epsilon}{n}$ which suffices given that:

$$d_{\mathrm{TV}}(\sum_i X_i, \sum_i Y_i) \leq \sum_i d_{\mathrm{TV}}(X_i, Y_i) = \sum_i |\mathbb{E}[X_i] - \mathbb{E}[Y_i]|$$

The Structure of PBDs

- S_n: set of all PBDs on n variables
- **[D-Papadimitriou '09]:** For all $\varepsilon > 0$, there exists a ***proper*** ε-cover $S_{n,\varepsilon} \subseteq S_n$ of size:

$$|S_{n,\epsilon}| \leq n^2 + n \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$$

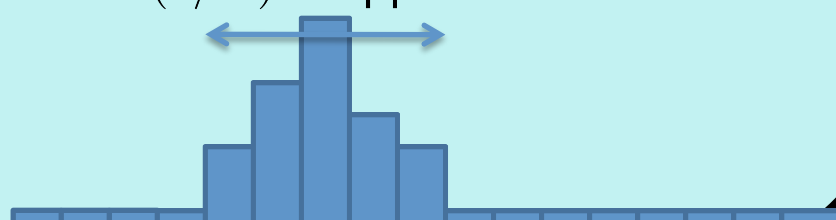
only keep subset of these with different log(1/ε) first moments

- In particular:

$$S_{n,\epsilon} = \left[\begin{array}{c}\textbf{Binomials}\\ \mathrm{Bin}(n', p)\\ n' \leq n \ \text{ and } \ p \propto \frac{1}{n}\end{array}\right] \cup \left[\begin{array}{c}\textbf{shifted-sparse PBDs}\\ n' + \sum_{i=1}^{1/\epsilon^3} Y_i\\ n' \leq n \ \text{ and } \ \mathbb{E}[Y_i] \propto \epsilon^2, \forall i\end{array}\right]$$

2-parameter distributions

$O(1/\epsilon^3)-$support



The Structure of PBDs

- S_n: set of all PBDs on n variables
- **[D-Papadimitriou '09]:** For all $\varepsilon > 0$, there exists a ***proper*** ε-cover $S_{n,\varepsilon} \subseteq S_n$ of size:

$$|S_{n,\epsilon}| \leq n^2 + n \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$$

only keep subset of these with different log(1/ε) first moments

- In particular:

$$S_{n,\epsilon} = \left(\begin{array}{c} \textbf{Binomials} \\ \mathrm{Bin}(n', p) \\ n' \leq n \quad \text{and} \quad p \propto \frac{1}{n} \end{array} \right) \cup \left(\begin{array}{c} \textbf{shifted-sparse PBDs} \\ n' + \sum_{i=1}^{1/\epsilon^3} Y_i \\ n' \leq n \quad \text{and} \quad \mathbb{E}[Y_i] \propto \epsilon^2, \forall i \end{array} \right)$$

- **Corollary:** For all $\varepsilon>0$, every PBD on n variables is either ε-close to a Binomial or ε-close to a shifted PBD on $1/\varepsilon^3$ variables.

Implications to Learning

[D-Diakonikolas-Servedio'12]: Let ***P*** be an unknown PBD in S_n.

- [Properly Learning PBDs] Given $\tilde{O}(1/\epsilon^2)$ independent draws from ***P*** and computation time

$$\left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)} \cdot \log n$$

can compute a PBD ***Q*** such that d_{TV} (***P***, ***Q***) < ε.

- Any algorithm requires $\Omega(1/\epsilon^2)$ samples (even for *n*=1).

Proof of Learning Result (Attempt 1)

- Use a cover based approach
- **[D-Kamath'14, Acharya et al'14]:** Suppose F_ε is an ε-cover (in TV distance) of a family of distributions F.

Then can learn any $P \in F$ to within O(ε)-distance using $O\left(\frac{\log |F_\epsilon|}{\epsilon^2}\right)$ samples from P, in time $O(\frac{|F_\epsilon| \log |F_\epsilon|}{\epsilon^2})$.

- **How?** Setup tournament among distributions in F_ε whose winner cannot be too far from P.
- *Improves long line of similar algorithms* ***[Devroyé-Lugosi'01, etc]*** *quadratically in the runtime by designing a better tournament*

- **In our PBD context:** Exists cover of size $n^2 + n \cdot (1/\epsilon)^{O(\log^2(1/\epsilon)}$

➔ sample complexity of problem is $\tilde{O}\left(\frac{\log n}{\epsilon^2}\right)$

- Fell short from our goal of $\tilde{O}(1/\epsilon^2)$.

Proof of Learning Result (Attempt 2)

- Exploit not just the *size* of the cover, but also its *structure*.
- We know that every PBD $X = \sum_i X_i$ is ε-close to
  - A binomial: $\mathrm{Bin}(n', p)$, $n' \leq n$
  - OR a shifted PBD on 1/ε² variables: $n' + \sum_{i=1}^{1/\epsilon^3} Y_i$, $n' \leq n$

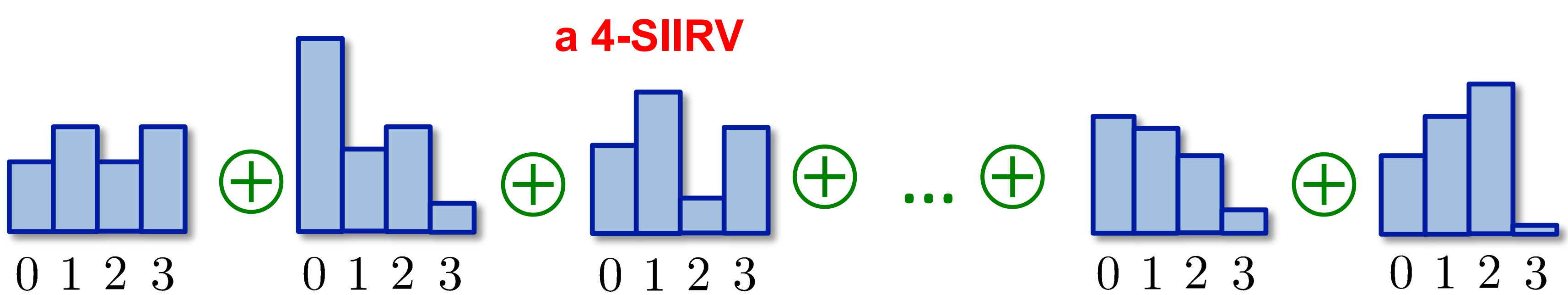
1. Using $O(1/\epsilon^2)$ samples estimate mean and variance of *X*.
2. Find Binomial distribution D_1 matching learned mean and variance.

3. In this case, all but ε probability mass of *X* is on support of length 1/ε³.
4. With $\tilde{O}(1/\epsilon^2)$ samples:
   i. Find the support of 1-ε mass of *X*. This gives estimate of shift *n'*.
   ii. Run tournament on ε-subcover of shifted by ≈*n'* PBDs on 1/ε³-*variables*
   iii. Let D_2 be the winner of the tournament.

5. Run tournament between D_1 and D_2.

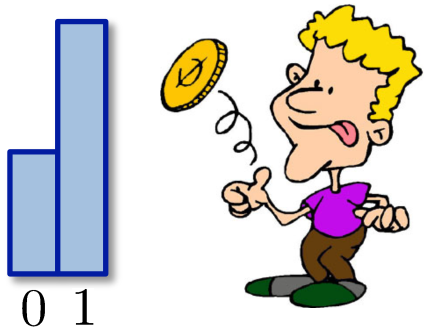
Menu

- *Refresher*
- *Objectives for this talk*
- *PBD Structure and Learning*
- ***k*-SIIRV Structure and Learning**

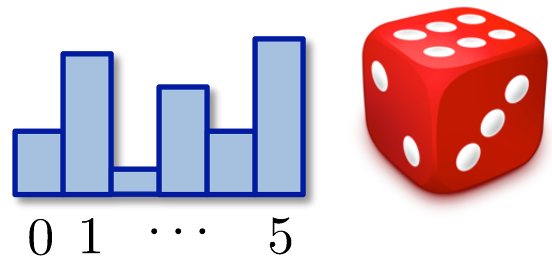


Terminology

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

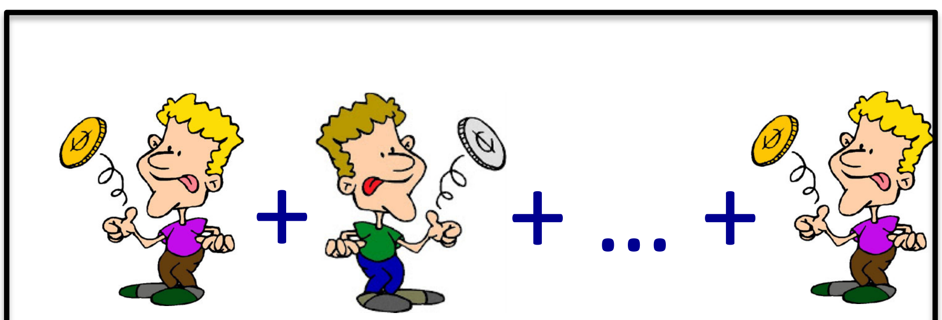


2-IRV



6-IRV

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



2-SIIRV

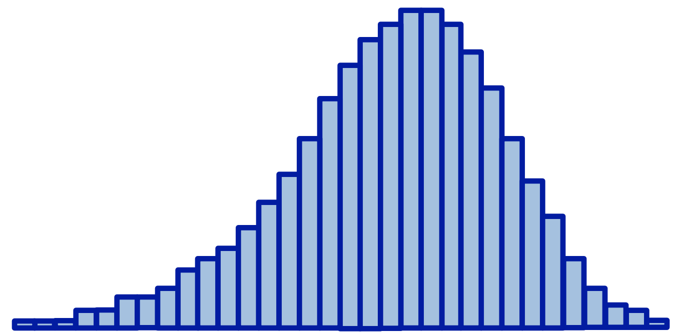
k-SIIRV

Structure ✓ Learning from $\tilde{O}(1/\epsilon^2)$ samples ✓

from 2 to *k*-SIIRVs: a whole new ball game

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

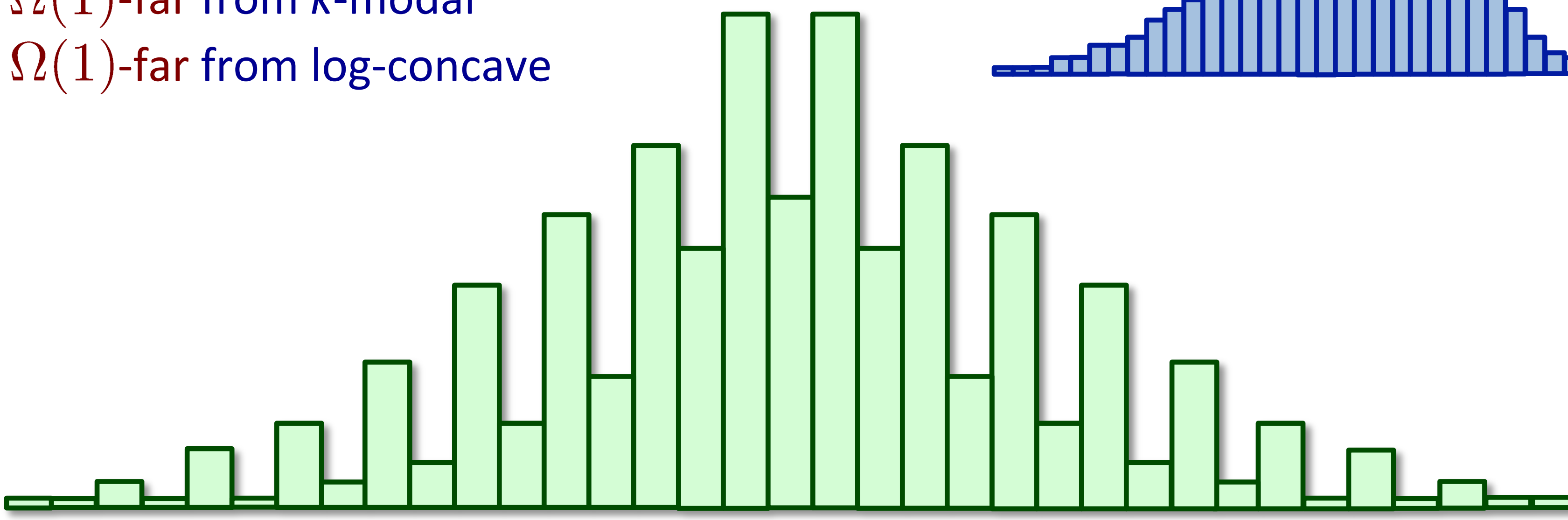
2-SIIRVs : unimodal, log-concave



3-SIIRVs :

$\Omega(1)$-far from *k*-modal

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

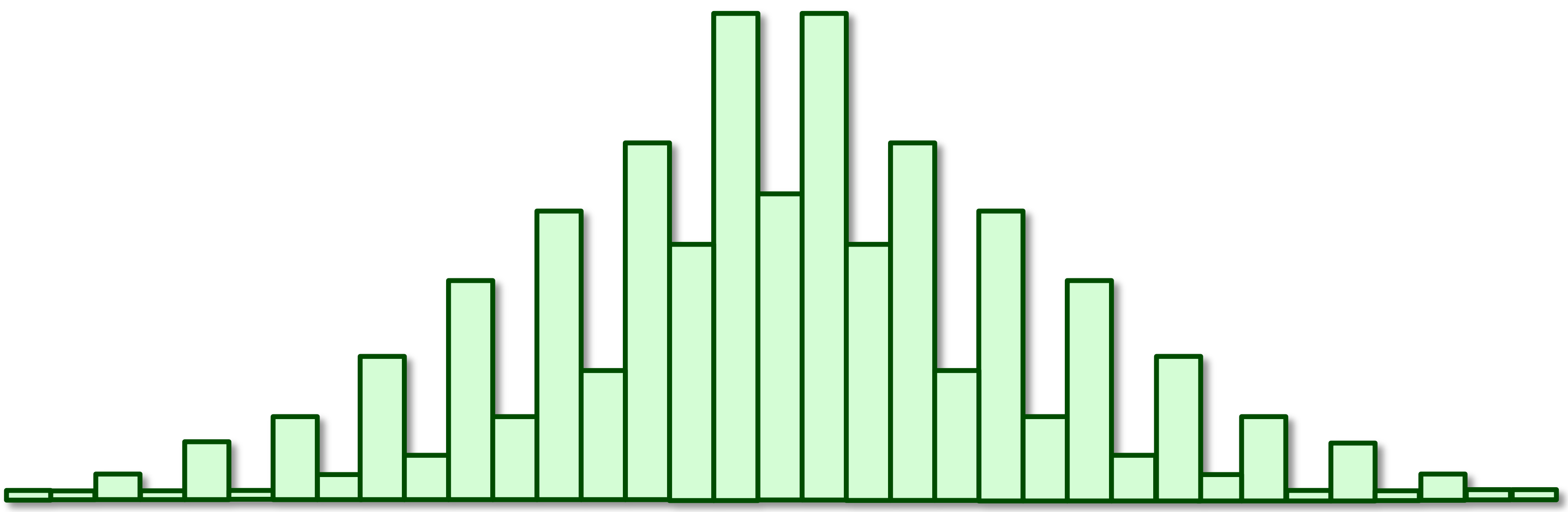


$$Y + 2 \cdot Z, \quad \text{where: } Y \sim \text{Bernoulli}(1/3) \;\; \text{and } Z \sim \text{Bin}(64, 1/2)$$

Structure of *k*-SIIRVs

- [Berry 1941, Esseen 1942]: If $X_1, \ldots, X_n$ are independent and bounded then

$$d_K\left(\sum_i X_i, \mathcal{N}(\mu, \sigma^2)\right) \le C \frac{\sum_i \mathbb{E}[|X_i|^3]}{\sigma^3}$$



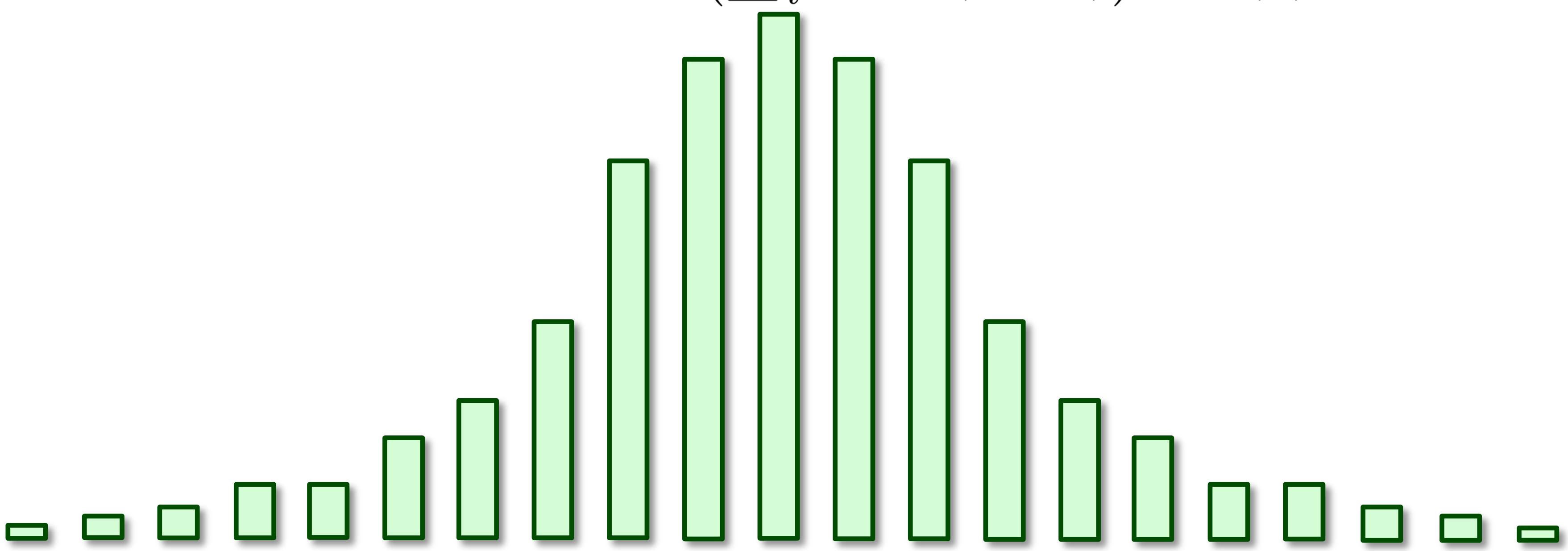
$$Y + 2 \cdot Z, \quad \text{where: } Y \sim \text{Bernoulli}(1/3) \;\; \text{and } Z \sim \text{Bin}(64, 1/2)$$

Structure of *k*-SIIRVs

- [Berry 1941, Esseen 1942]: If $X_1,\ldots,X_n$ are independent and bounded then

$$d_K\left(\sum_i X_i, \mathcal{N}(\mu,\sigma^2)\right) \le C\frac{\sum_i \mathbb{E}[|X_i|^3]}{\sigma^3}$$

- Clearly, in general: $d_{\mathrm{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) = \Omega(1)$
- Conditions under which $d_{\mathrm{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) = o(1)$?



$2 \cdot Z$, where: $Z \sim \mathrm{Bin}(n, 1/2)$

Structure of *k*-SIIRVs

- [Berry 1941, Esseen 1942]: If $X_1,\ldots,X_n$ are independent and bounded then

$$d_K\left(\sum_i X_i, \mathcal{N}(\mu,\sigma^2)\right) \le C\frac{\sum_i \mathbb{E}[|X_i|^3]}{\sigma^3}$$

- Clearly, in general: $d_{\mathrm{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) = \Omega(1)$
- Conditions under which $d_{\mathrm{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) = o(1)$?
- [Chen-Goldstein-Shao 2011]: If $X_1,\ldots,X_n$ are independent *k*-IRVs and

$$d_{\mathrm{TV}}\left(\sum_{j\neq i} X_j, \sum_{j\neq i} X_j + 1\right) \le \delta, \forall i$$

then

$$d_{\mathrm{TV}}\left(\sum_i X_i, Z(\mu,\sigma^2)\right) = O(k)\left(\frac{1}{\sigma}+\delta\right)$$

[Daskalakis-Diakonikolas-O'Donnel-Servedio-Tan]

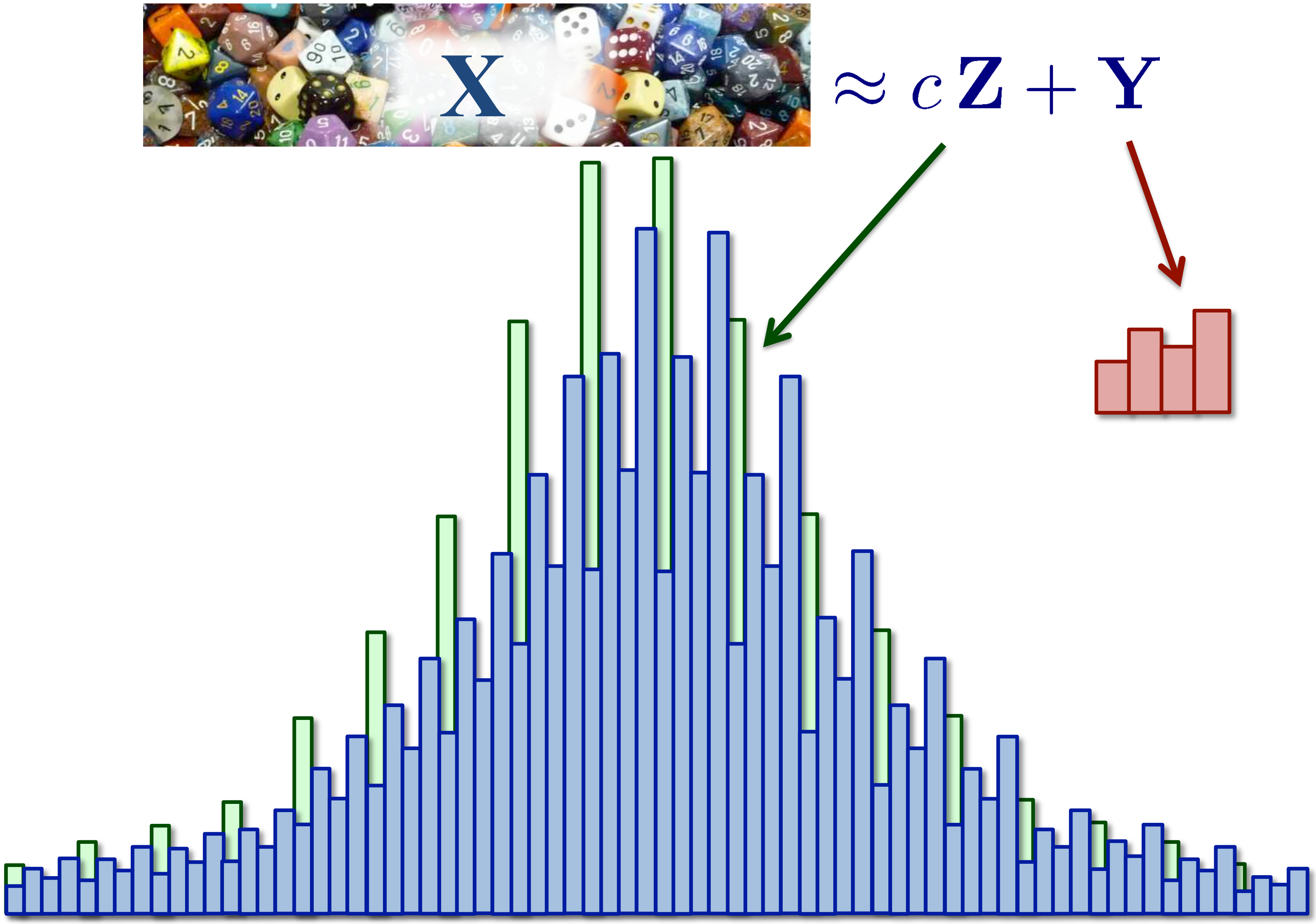
[DDOST'13]: Let X be a k-SIIRV with $\mathrm{Var}[X] \geq \mathrm{poly}(k/\epsilon)$.

Then X is ε-close to $cZ + Y$, where

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

Y, Z: independent

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



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

Structure of *k*-SIIRVs

[DDOST'13]: Let *X* be a *k*-SIIRV with $\mathrm{Var}[X] \geq \mathrm{poly}(k/\epsilon)$.

Then *X* is *ε*-close to $cZ + Y$, where

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

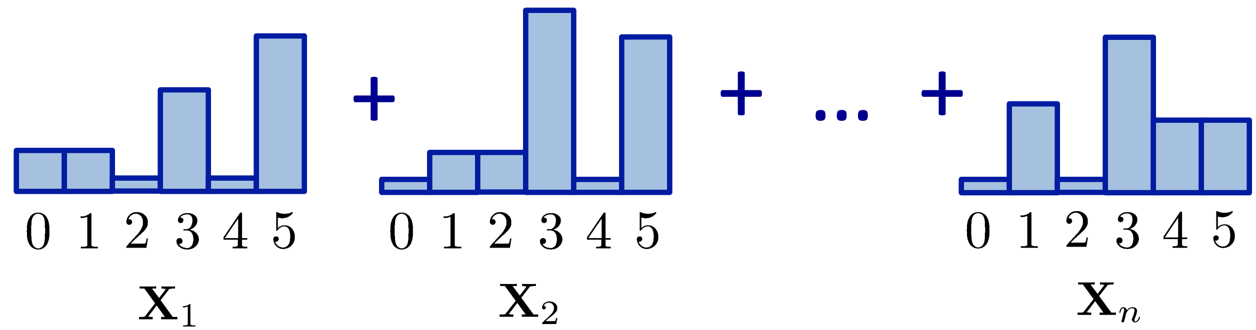
Y, *Z*: independent

Corollary: Let *X* be an arbitrary *k*-SIIRV. For all *ε*>0, *X* is *ε*-close to:

- a $\mathrm{poly}(k/\epsilon)$- IRV
- *OR c Z + Y,* where:
  - *c* ∈ {1,…, *k*-1}
  - *Z* = discretized normal
  - *Y* = *c*-IRV

Proof of Structural Theorem

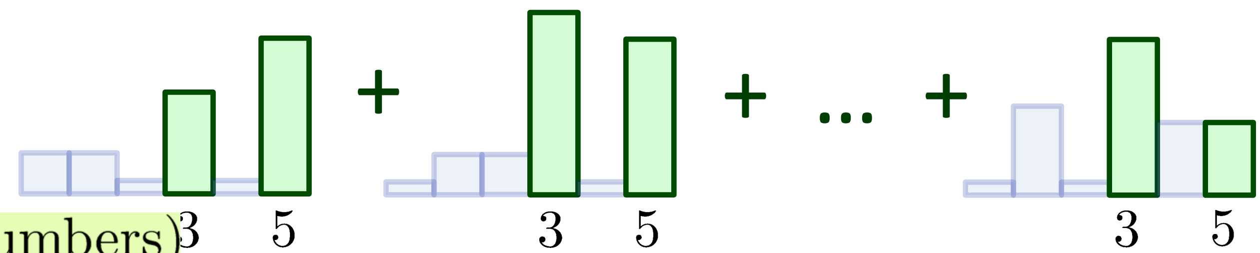
$\mathbf{X}$



$c\,\mathbf{Z}$

structured **global** component

Heavy numbers: $\sum_{i=1}^{n} \mathbf{Pr}[\mathbf{X}_i = b]$ large

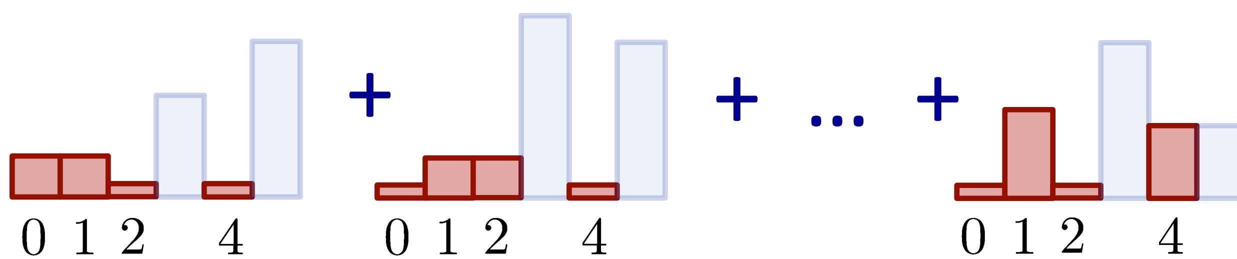


$c = \gcd(\text{heavy numbers})$

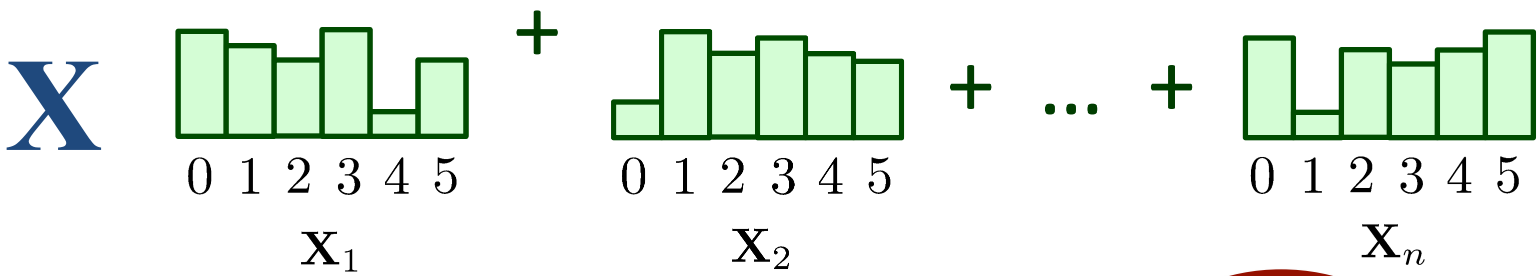
$\mathbf{Y}$

arbitrary **local** component

Light numbers: $\sum_{i=1}^{n} \mathbf{Pr}[\mathbf{X}_i = b]$ small



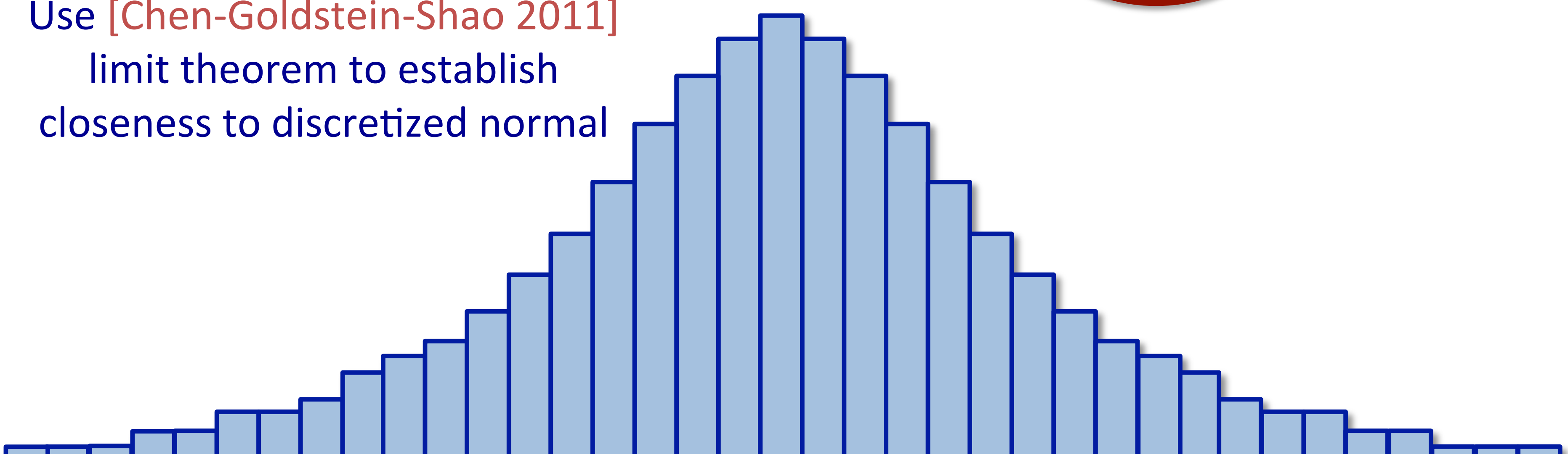
Special case: *all* numbers heavy



Intuition: No "mod structure" in **X**

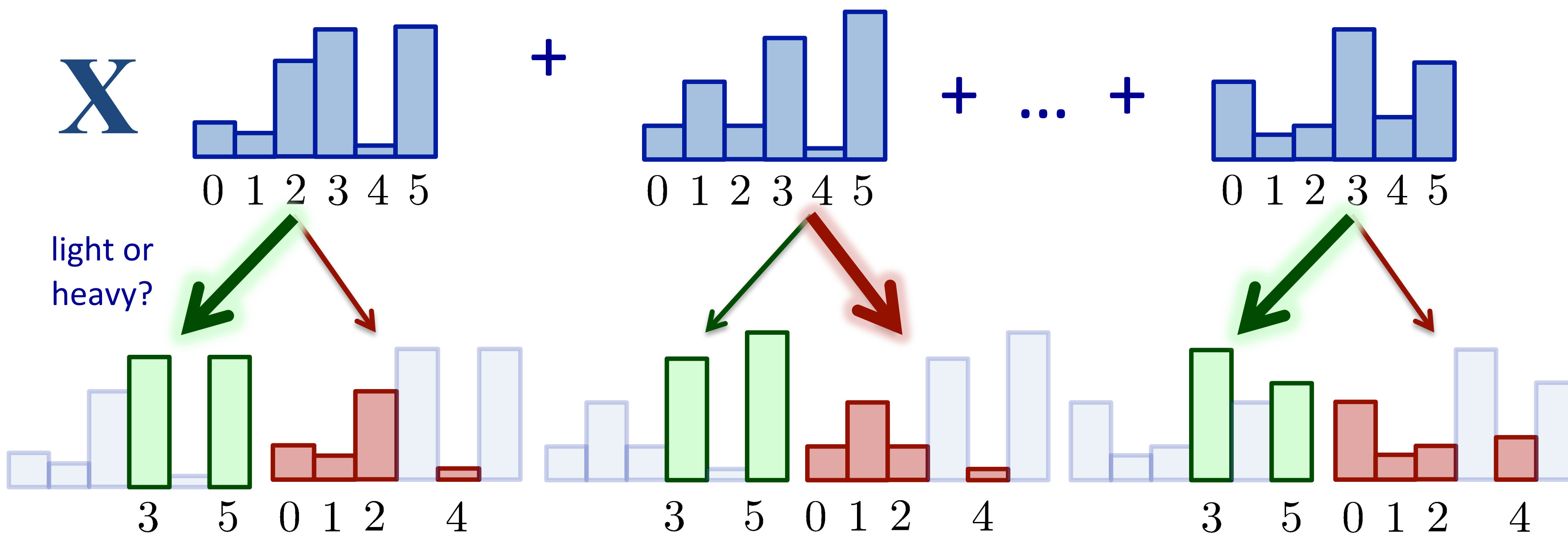
e.g. **X** equally likely to be 0 or 1 mod 2

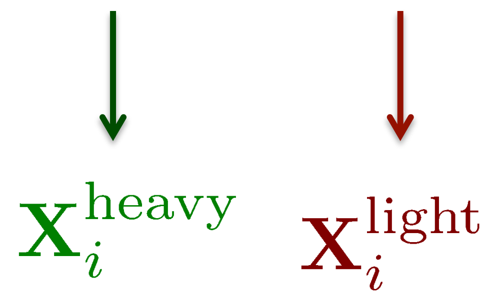
Use [Chen-Goldstein-Shao 2011] limit theorem to establish closeness to discretized normal



General Case: indirect sampling procedure

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

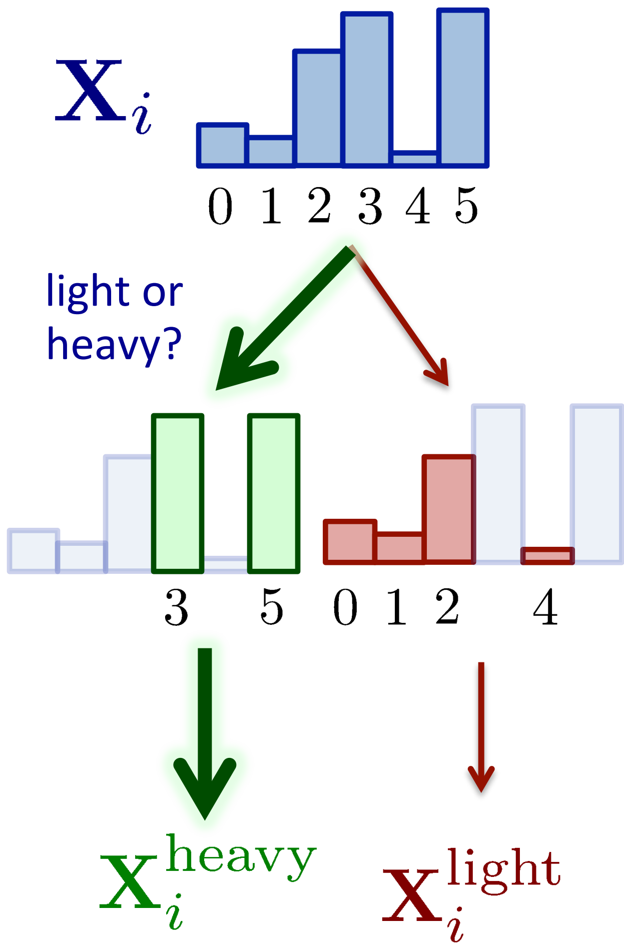




1. Decide independently for each $\mathbf{X}_i$ whether outcome will be heavy or light.

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

Analysis



Every outcome 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 O

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

where $\mathbf{Z}$ = disc. norm. *independent* of O.

- Proof uses "all numbers heavy" special case
- $c = \gcd(\text{heavy numbers})$

Structure of *k*-SIIRVs

[DDOST'13]: Let *X* be a *k*-SIIRV with $\mathrm{Var}[X] \geq \mathrm{poly}(k/\epsilon)$.

Then *X* is *ε*-close to $cZ + Y$, where

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

Y, *Z*: independent

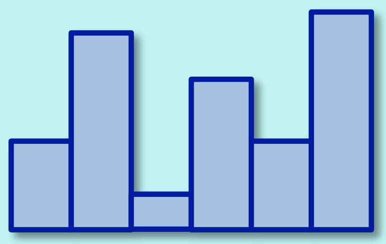
Corollary: Let *X* be an arbitrary *k*-SIIRV. For all *ε*>0, *X* is *ε*-close to:

- a $\mathrm{poly}(k/\epsilon)$- IRV
- *OR c Z + Y,* where:
  - *c* ∈ {1,…, *k*-1}
  - *Z* = discretized normal
  - *Y* = *c*-IRV

Learning *k*-SIIRVs

[DDOST'13]: Let $S_{n,k}$ be the class of *k*-SIIRVs, *i.e.* all distributions of a sum $X = \sum_{i=1}^{n} X_i$ of *n* independent *k*-IRVs.

There is an algorithm that learns an arbitrary $\boldsymbol{P} \in S_{n,k}$ with time and sample complexity $\mathrm{poly}(k/\epsilon)$, independent of *n*.



Recall: $\Omega(k/\epsilon^2)$ samples necessary even for a single *k*-IRV

Proof of Learning Result for *k*-SIIRVs

Corollary: Let X be an arbitrary k-SIIRV. For all $\varepsilon>0$, X is ε-close to:

- a $\mathrm{poly}(k/\epsilon)$- IRV
- *OR* $c\,Z + Y$, where:
  - $c \in \{1,\ldots, k\text{-}1\}$
  - Z = discretized normal
  - Y = c-IRV

- If $\boldsymbol{X}$ is ε-close to $\mathrm{poly}(k/\epsilon)$ - IRV: easy to learn from $\mathrm{poly}(k/\epsilon)$ samples
- Else: guess $c \in \{1,\ldots, k\text{-}1\}$.
- Learn Z from conditional distribution on integers: 0 mod k
- Learn Y as the appropriate mixing distribution
- Run tournament to choose among (k+1) distributions generated.

Summary

1. Ilias discussed how ***shape*** restrictions on a distribution (monotonicity, *k*-modality, log-concavity) permit faster learning algorithms

2. I discussed how ***syntactic*** restrictions on a distribution permit even faster learning:
- PBDs on *n* variables have support {0,…,*n*} but can be learned from $\tilde{O}(1/\epsilon^2)$ samples
- *k*-SIIRVS on *n* variables have support {0,..,*n* (*k*-1)} but can be learned from $\mathrm{poly}(k/\epsilon)$ samples

3. In turn, finding these improved algorithms requires stronger limit theorems, and tighter structural results for these distributions.

4. **Take-away 1:** Every PBD on *n* variables is ε-close to a Binomial or ε-close to a shifted PBD on $1/\varepsilon^3$ variables.

Summary

5. **Take-away 2:** Every *k*-SIIRV *X* on *n* variables is *ε*-close to a distribution of support poly(*k*/*ε*) OR

$$\approx_{\varepsilon} c\,\mathbf{Z} + \mathbf{Y}$$

Future Directions

6. Super Concrete Open Problems:

- Properly Learn PBDs from $1/\varepsilon^2$ samples in polynomial time
  (our algorithm was quasi-polynomial)
- Maybe further improve PBD cover size to poly(n/ε)
- Properly learn k-SIIRVs (our learner outputs a poly(k/ε)-SIIRV)
- Proper covers of k-SIIRVs

7. Multi-dimensional Case:

- Learn $\boldsymbol{d}$-dimensional $\boldsymbol{k}$-flat distributions, in poly($d\ k/\varepsilon$) time.
  Possible from poly($d\ k/\varepsilon$) samples (information theoretically)
- Generalize structural/learning results to Poisson Multinomial Distributions
  Preliminary results with Kamath and Tzamos

Thanks!