Learning SICSIRVs

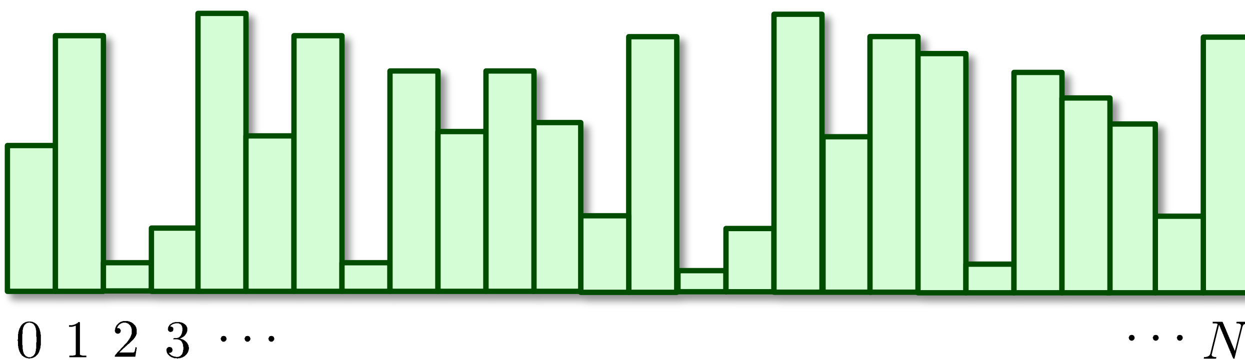
Anindya De
Northwestern University

Philip Long
Google

Rocco Servedio
Columbia University

Learnability of discrete distributions

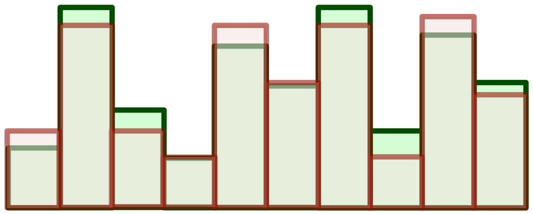
- Discrete distributions: distributions over $\mathbb{Z}$.
- Learning problem defined by class $\mathcal{C}$ of distributions. Unknown target distribution $\mathcal{D} \in \mathcal{C}$.



Learnability of discrete distributions

- Learner gets i. i. d. samples from distribution $\mathcal{D}$.

- Aim: with probability 9/10, the learner produces a hypothesis $\mathcal{D}'$ such that $\|\mathcal{D} - \mathcal{D}'\|_1 \leq 1/10$.



What is a SICSIRV?

- We'll talk about it later.
- Let's begin with an example.
- Consider the family of *Poisson Binomial distributions:* Sums of Independent Bernoulli random variables.
- In other words, each sample $Z \sim Z_1 + \ldots + Z_n$ where $Z_1, \ldots, Z_n$ are independent $\{0, 1\}$ r.v.s

Learnability of Poisson Binomial Distributions

- [Daskalakis, Diakonikolas, Servedio – 2012] The complexity of learning Poisson Binomial distributions is $\text{poly}(1/\epsilon)$. This complexity is independent of n!

- Strategy: Either

(i) The target distribution has large variance i.e. variance $\geq \text{poly}(1/\epsilon)$.

(ii) Target distribution has small variance $\leq \text{poly}(1/\epsilon)$.

Case Analysis

- Large variance (non-degenerate case): If the variance is at least poly$(1/\epsilon)$, then the distribution is $O(\epsilon)$ close to a discretized Gaussian (with the population mean and variance).

- Small variance (degenerate case): If the variance is at most poly$(1/\epsilon)$, then the effective support is poly$(1/\epsilon)$.

Case Analysis

- Large variance (non-degenerate case): Reduces to learning an approximate Gaussian distribution. Learning both the mean and variance to error ϵ takes poly$(1/\epsilon)$ samples.

- Small variance (degenerate case): The size of the effective support is poly$(1/\epsilon)$. Can be learnt by brute force in time poly$(1/\epsilon)$.

PBD $\Longrightarrow$ k-SIIRV

- [DDS]:

PBDs – Sums of independent $\{0,1\}$ r. v. s.

PBD $\Longrightarrow$ k-SIIRV

- [DDS + O’Donnell and Tan]:

PBDs – Sums of independent $\{0, 1\}$ r. v. s.

k-SIIRV

$\{0,1\ ,\ ..,\ \mathrm{k}\}$

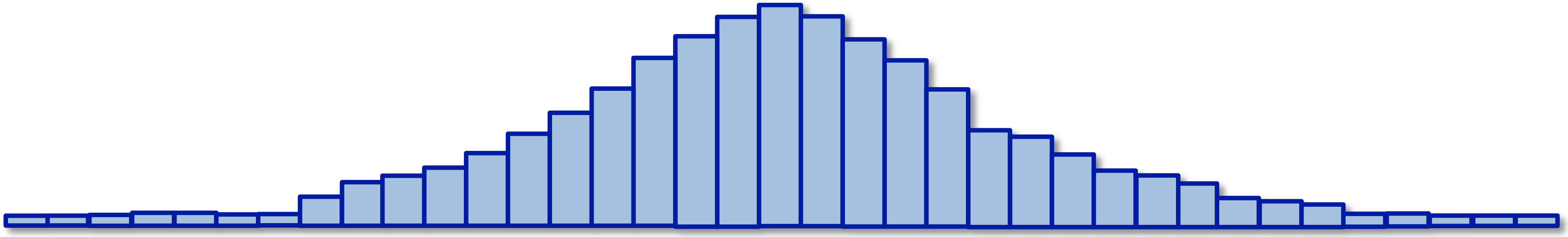
[DDOST] – k-SIIRVs can be learnt using poly(k/ϵ) samples in the same time.

Learning algorithm for k-SIIRVs

- [DDOST] : Similar proof structure as PBDs.

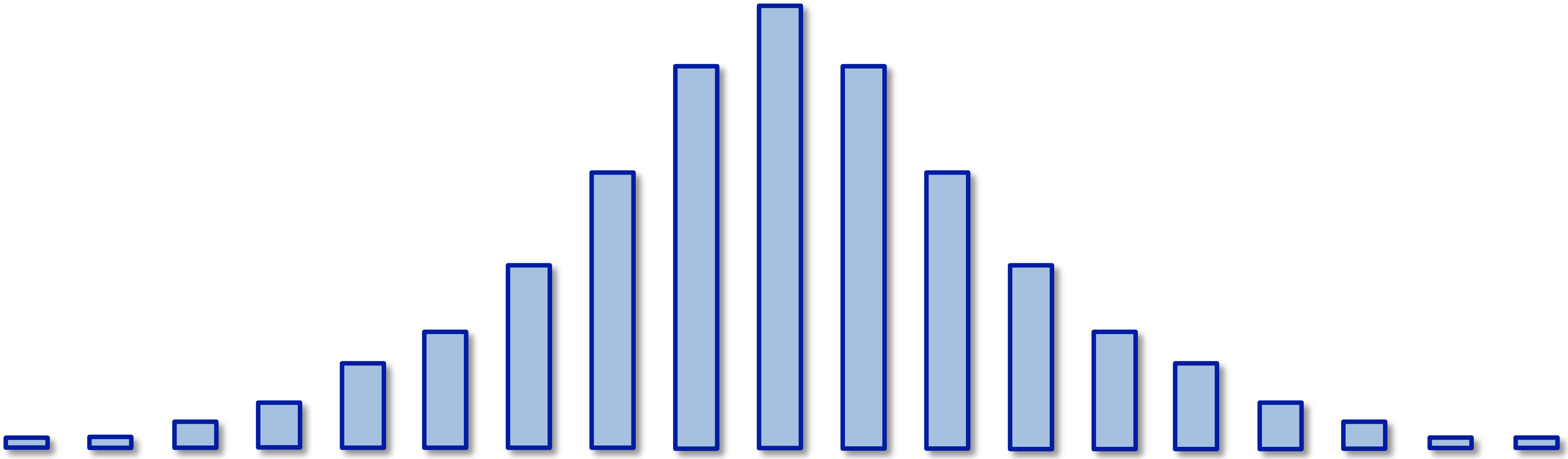
[Main structure theorem:] Every k-SIIRV can be approximately written as a convolution of a sparse (poly(k/ϵ)) distribution with a scaled discrete Gaussian for some scaling factor in $\{1, 2, \ldots, k\}$.

What is a scaled discrete Gaussian?



Discrete Gaussian

What is a scaled discrete Gaussian?



Scaled discrete Gaussian

Summary so far

- Distributions which are sums of independent commonly supported integer random variables (SICSIRVs) supported on $S = \{0, 1, \ldots, k\}$ can be learnt in time and samples poly(k/ϵ).

- What about the complexity of learning SICSIRVs supported on other sets S of small size?

Learning SICSIRVs

- For any set $|S| = 2$, SICSIRV over S is a linear translation of a PBD.

- What about sets S of size strictly more than 2?

Main result(s)

- Given any set S of size 3, SICSIRV over S can be learnt in time $\text{poly}(1/\epsilon)$.

- There exists infinitely many sets S such that $S = \{0 \leq r \leq q \leq p\}$ learning SICSIRV over S requires $\Omega(\log \log p)$ samples.

Sharp transition between sets of sizes 3 and 4!

Positive result

- Given any set S of size 3, SICSIRV over S can be learnt in time /samples $\text{poly}(1/\epsilon)$.

Without loss of generality, assume that $S = \{0, p, q\}$,

i.e. summands are supported on the set $\{0, p, q\}$.

Positive result

- Given any set S of size 3, SICSIRV over S can be learnt in time poly$(1/\epsilon)$.

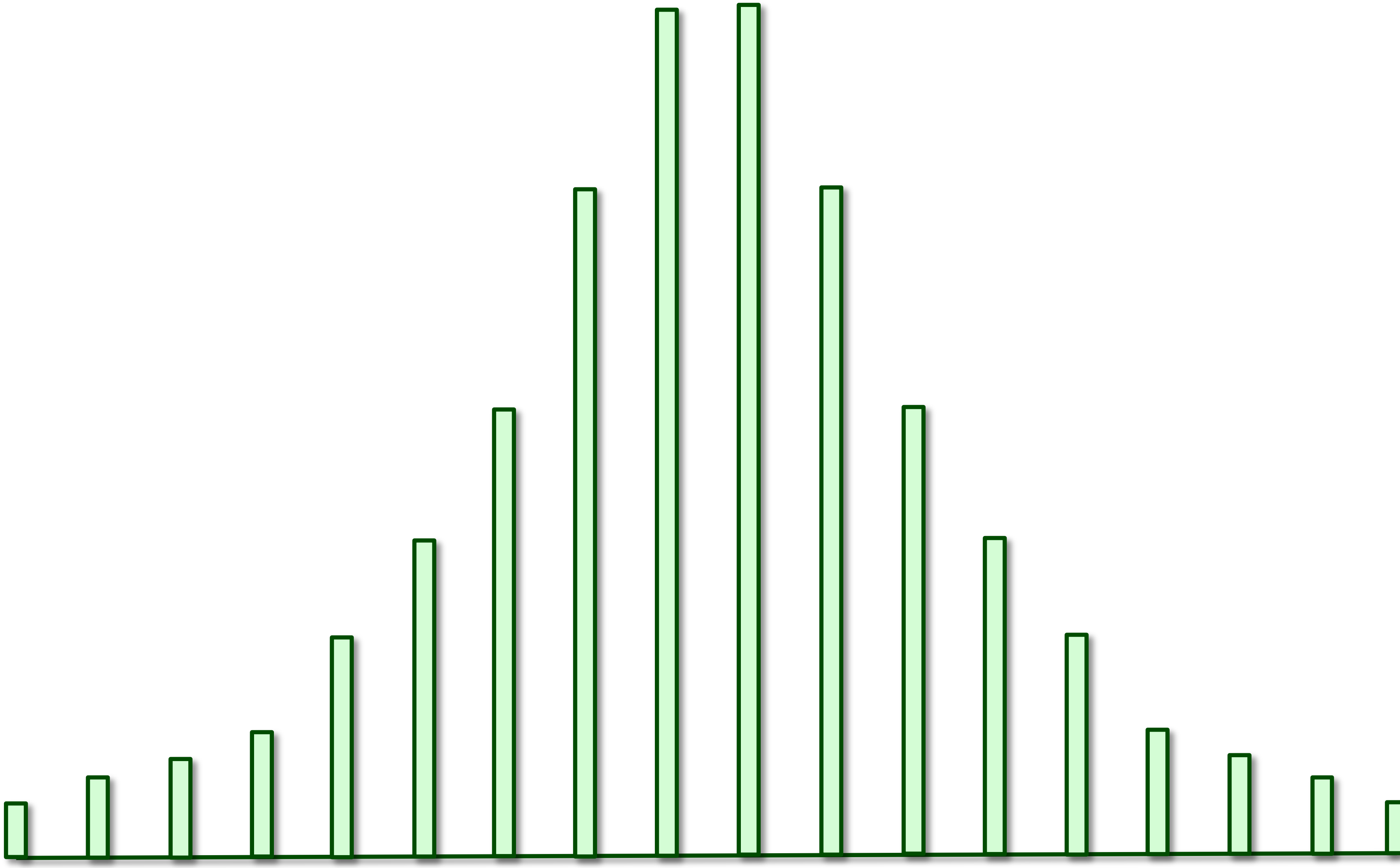
With~~out~~ loss of generality, assume that the summands are supported on the set $\{0, p\}$ and $\{0, q\}$. In other words, the target distribution is $p \cdot X^{(p)} + q \cdot X^{(q)}$ where $X^{(p)}, \ X^{(q)}$ are independent Poisson Binomial distributions.

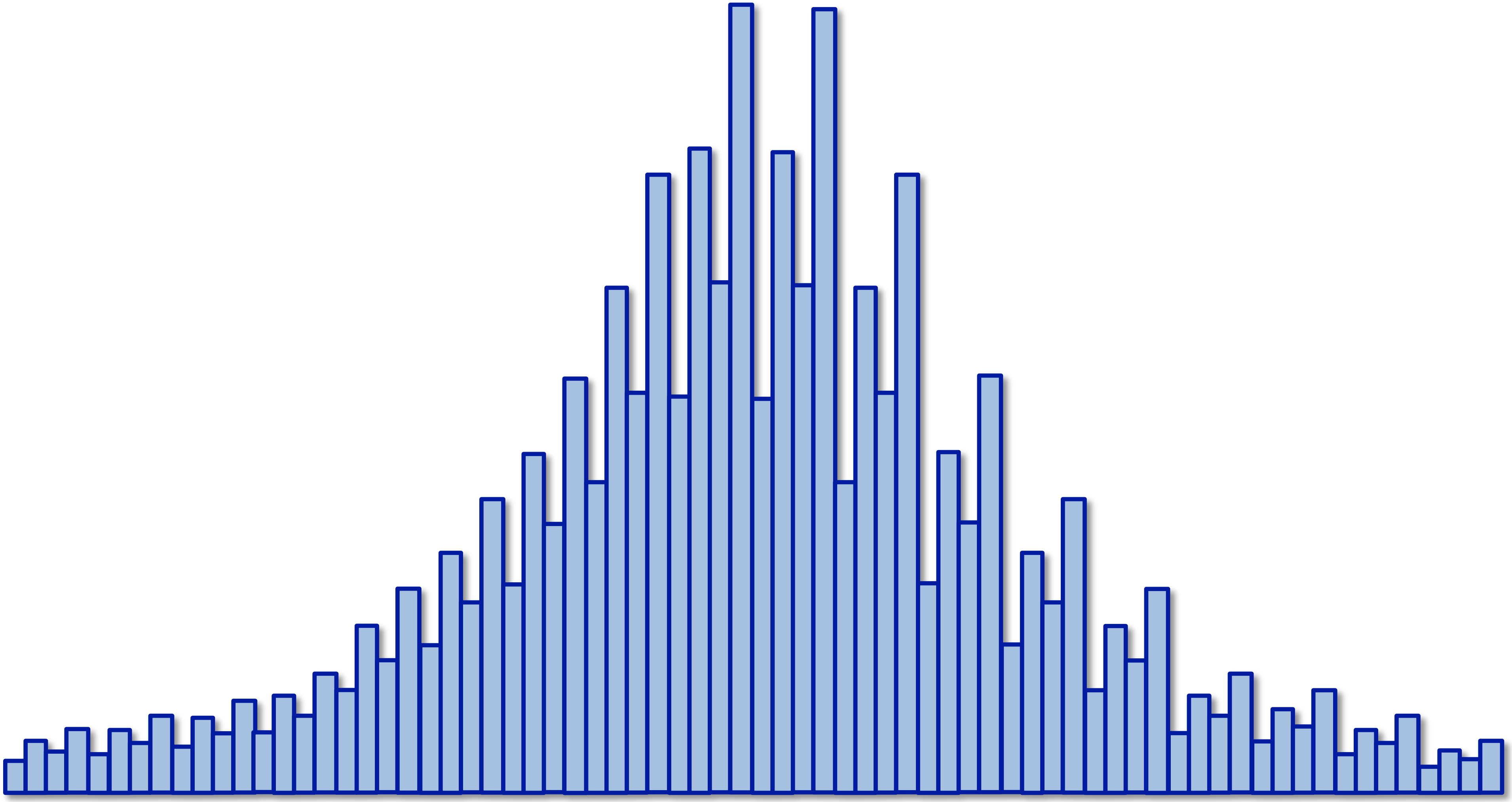
What does $p \cdot X^{(p)} + q \cdot X^{(q)}$ look like?

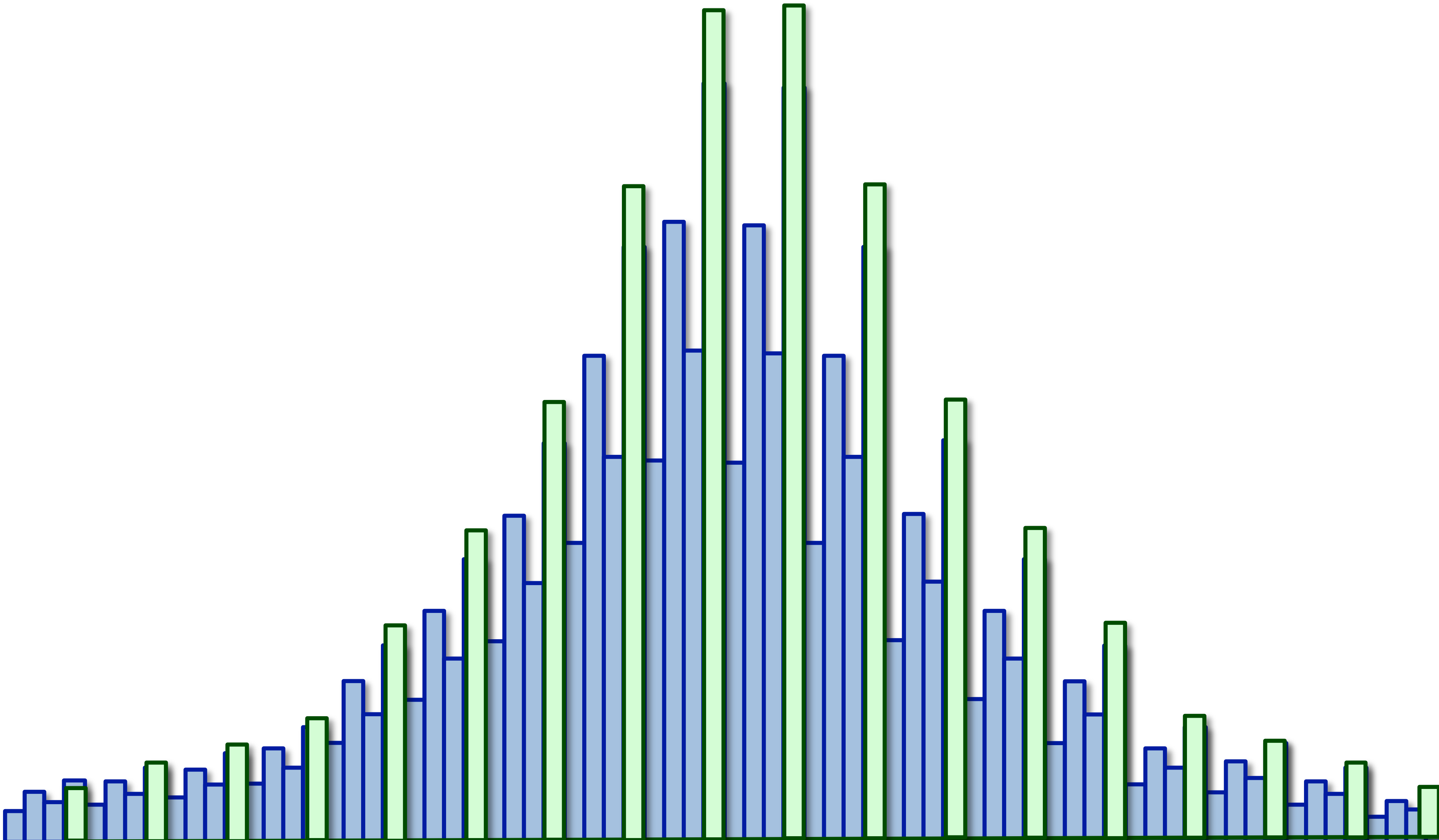
- ✓ Assume that Var$(X^{(p)})$, Var$(X^{(q)}) \geq$ poly$(1/\epsilon)$.
- ✓ Assume that Var$(p \cdot X^{(p)}) \geq$ Var$(q \cdot X^{(q)})$.

Lemma: The r.v. $Z = p \cdot X^{(p)} + q \cdot X^{(q)}$ looks like a

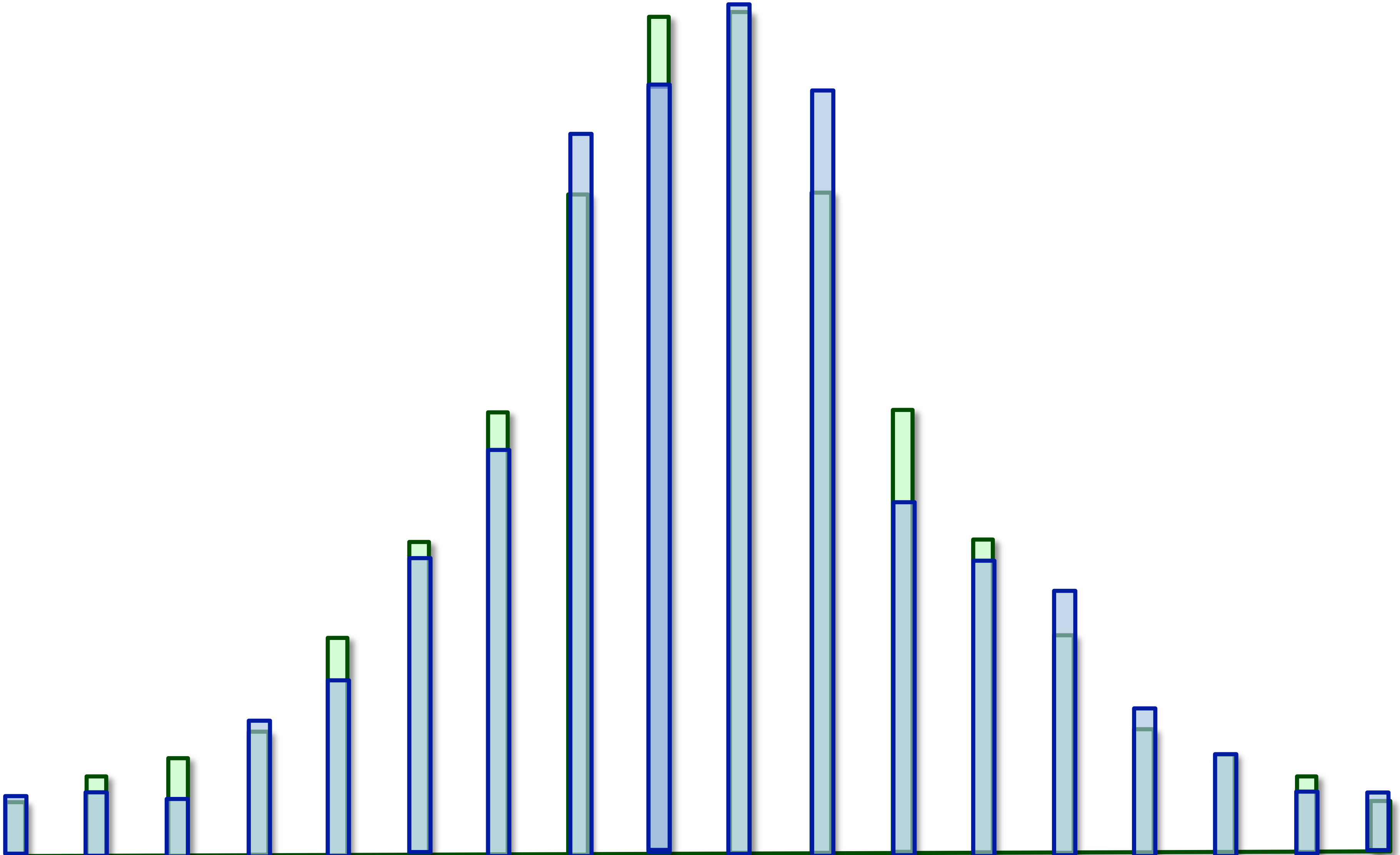
discretized Gaussian if you *blur your eyes at the scale of p.*







Total variation distance between the distributions may be large.



If you round the distributions to nearest multiples of p, they are close to each other.

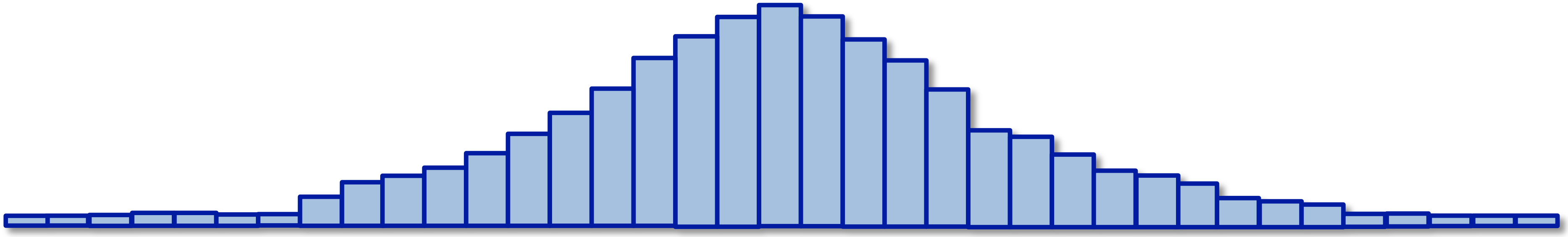
What does $p \cdot X^{(p)} + q \cdot X^{(q)}$ look like?

Lemma: The r.v. $Z = p \cdot X^{(p)} + q \cdot X^{(q)}$ looks like a discretized Gaussian if you *blur your eyes at the scale of p.*

- ✓ What is the structure of Z mod p?
- ✓ The discretized Gaussian is uniformly distributed mod p
- ✓ Thus, we need to study the structure of $X^{(q)} \bmod p$

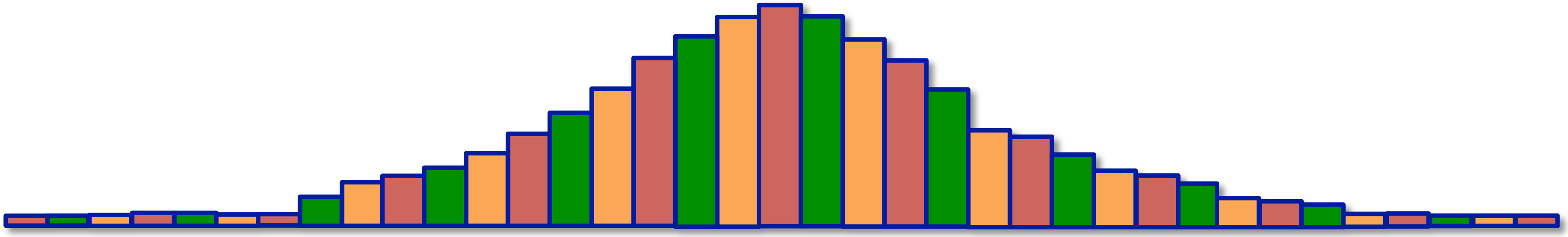
Structure of $X^{(q)}$ mod p

Lemma: If $\sigma(X^{(q)}) \gg p/\epsilon,$ then $q \cdot X^{(q)}$ is uniformly distributed in $\mathbb{Z}_p$. (Easy to prove)



Structure of $X^{(q)}$ mod p

Lemma: If $\sigma(X^{(q)}) \gg p/\epsilon,$ then $q \cdot X^{(q)}$ is uniformly distributed in $\mathbb{Z}_p$. (Easy to prove)



$:= 0 \mod 3$

$:= 1 \mod 3$

$:= 2 \mod 3$

All residue classes modulo 3 are roughly equidistributed.

Structure of $p \cdot X^{(p)} + q \cdot X^{(q)}$

Lemma: If $q \cdot X^{(q)}$ is uniformly distributed in $\mathbb{Z}_p$, then

$p \cdot X^{(p)} + q \cdot X^{(q)}$ is close to a discretized Gaussian.

(Not difficult to prove)

Proof: Requires some generalization of the notion of shift-Invariance from probability theory (measures smoothness of probability distributions). First used in CS by GMRZ.

What happens if $\sigma(X^{(q)}) \leq p/\epsilon$?

What happens if $\sigma(X^{(q)}) \ll p$?

Lemma: If $\sigma(X^{(q)}) \ll p$, then we can learn $X^{(q)}$.

Proof:

- ✓ Take samples of $p \cdot X^{(p)} + q \cdot X^{(q)}$ mod p.
- ✓ You learn $q \cdot X^{(q)}$ mod p and hence $X^{(q)}$ mod p. (Multiply samples by q^{-1} mod p).
- ✓ Since $\sigma(X^{(q)}) \ll p$, you *essentially* learn $X^{(q)}$.

To recap:

Two cases:

(a) If $\sigma(X^{(q)}) \gg p/\epsilon,$ then Z is essentially a Gaussian.

(b) If $\sigma(X^{(q)}) \ll p$, then taking the samples mod p, will reveal $X^{(q)}$. It is not difficult to infer $p \cdot X^{(p)}$ either.

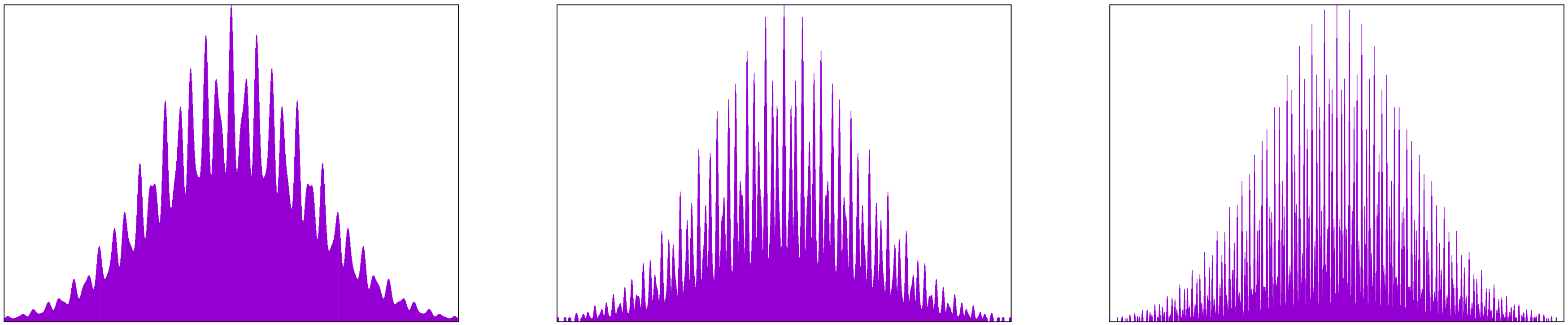
Lower bound

- There exists infinitely many sets S such that $S = \{0 \leq r \leq q \leq p\}$ learning SICSIRV over S requires $\Omega(\log \log p)$ samples.

(a) Choose $r = 1$.

(b) $q \approx \sqrt{p}$ is chosen carefully. Construction exploits delicate properties of continued fractions.

Picture aided proof

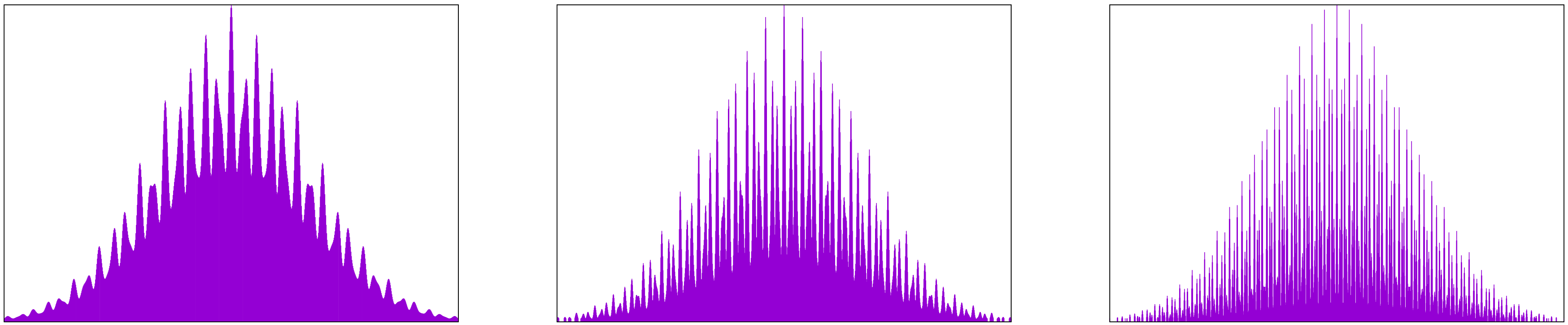


(a) (b) (c)

We construct a family of $\Omega(\log p)$ SICSIRVs over the set $\{0, 1, p, q\}$ such that

(1) All these families look like Gaussians at *the scale of p.*
(2) The "intra-p" structure is different among these distributions.
(3) The peak-valley structure becomes finer as we go from (a) to (c).
(4) Nearby peaks and valleys have mass ratio of at most 2.

Picture aided proof



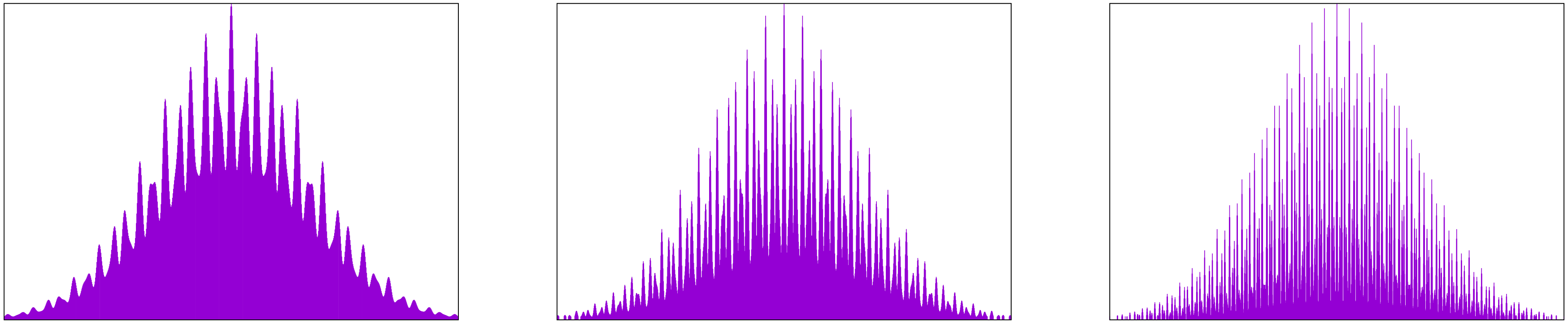
(a) (b) (c)

In other words, we obtain $\Omega(\log p)$ SICSIRVs over the set $\{0, 1, p, q\}$ such that

(1) ℓ_1 distance between any two of these distributions is > 0.1.

(2) KL-divergence between any two of these distributions is O(1).

Picture aided proof



(a) (b) (c)

This is sufficient for us to apply Fano's inequality and obtain a $\Omega(\log\log p)$ lower bound.

Thanks!