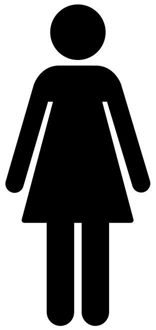
Differentially Private Medians and Interior Points for Non-Pathological Data

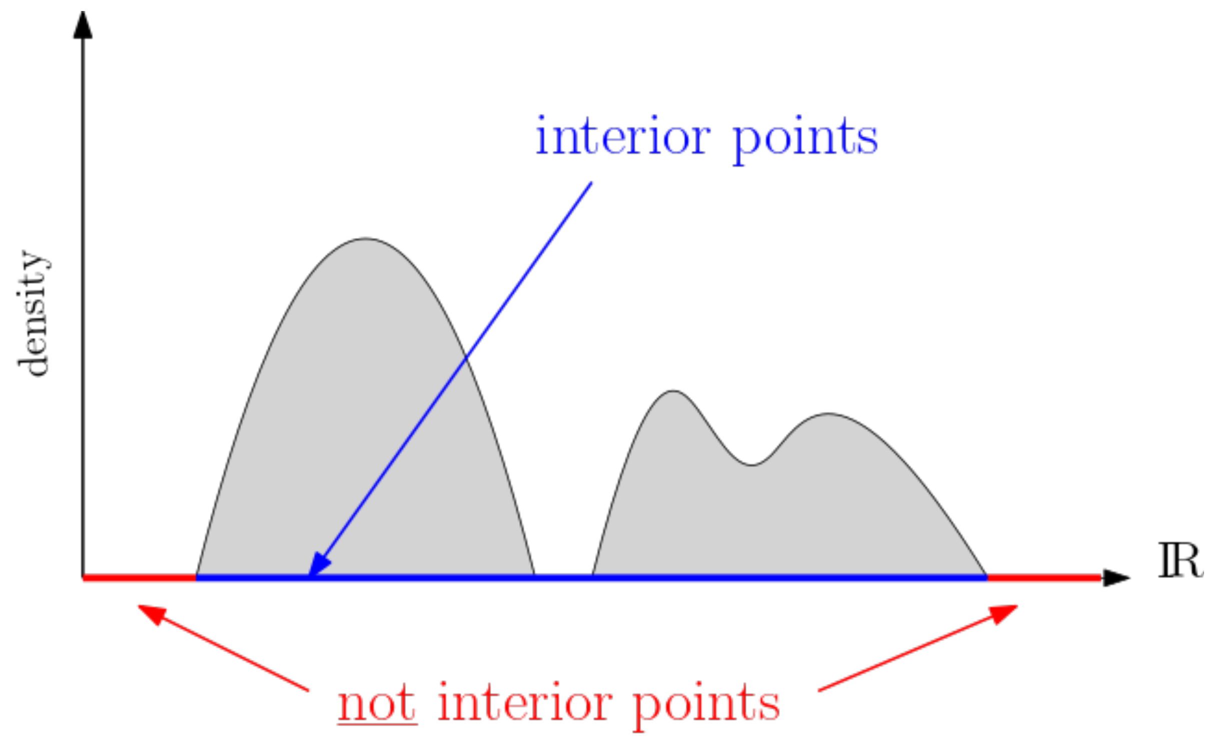
Maryam Aliakbarpour, Rose Silver, Thomas Steinke, Jonathan Ullman

The Interior Point Problem

Isn't this trivial?

Just return any sample X_i.



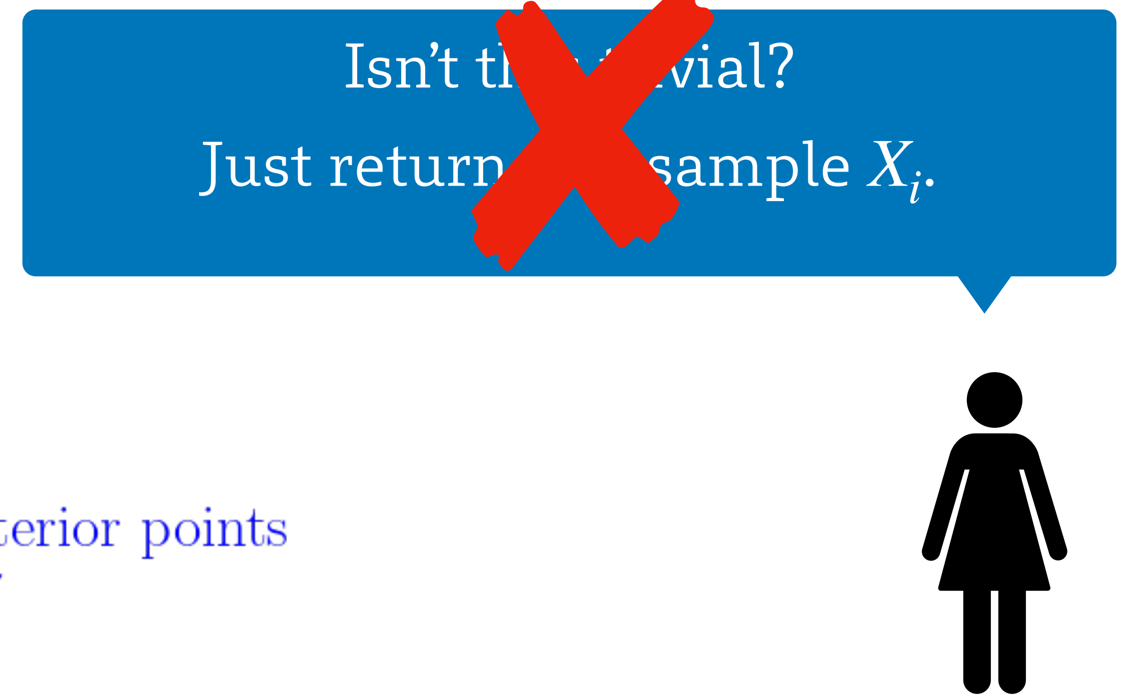


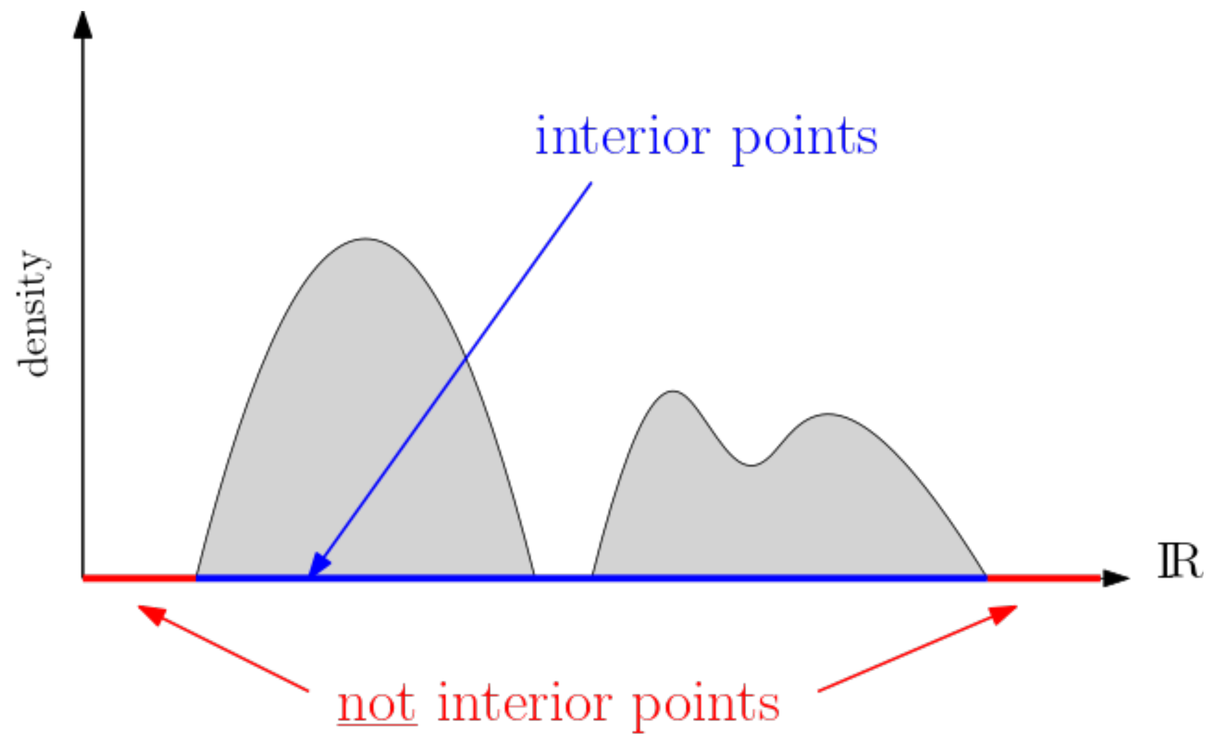
Interior Point Problem:

- Given n i.i.d. samples $X_1, \ldots, X_n \sim P$, return a point y s.t.

  $$\inf \text{support}(P) \leq y \leq \sup \text{support}(P)\,.$$

The Interior Point Problem





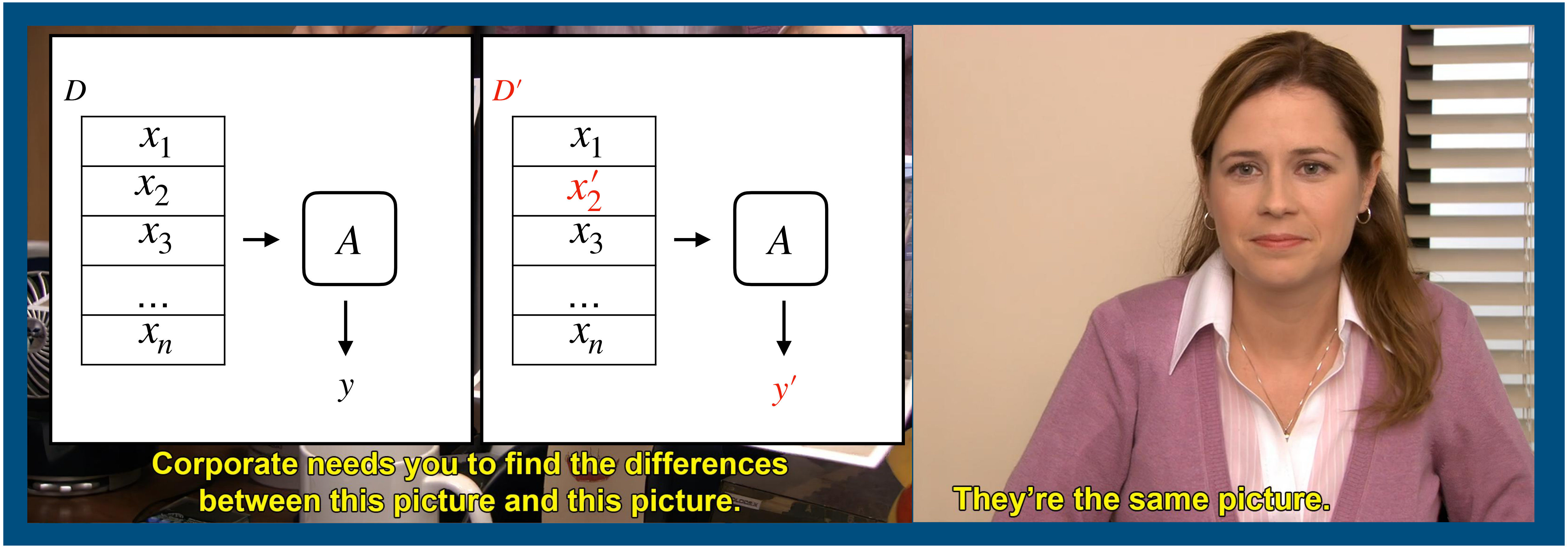
Private Interior Point Problem:

- Given n i.i.d. samples $X_1, \ldots, X_n \sim P$, privately return a point y s.t.

  $\inf \operatorname{support}(P) \le y \le \sup \operatorname{support}(P)\,.$

Differential Privacy (DMNS '06)

Differential Privacy (DMNS '06)

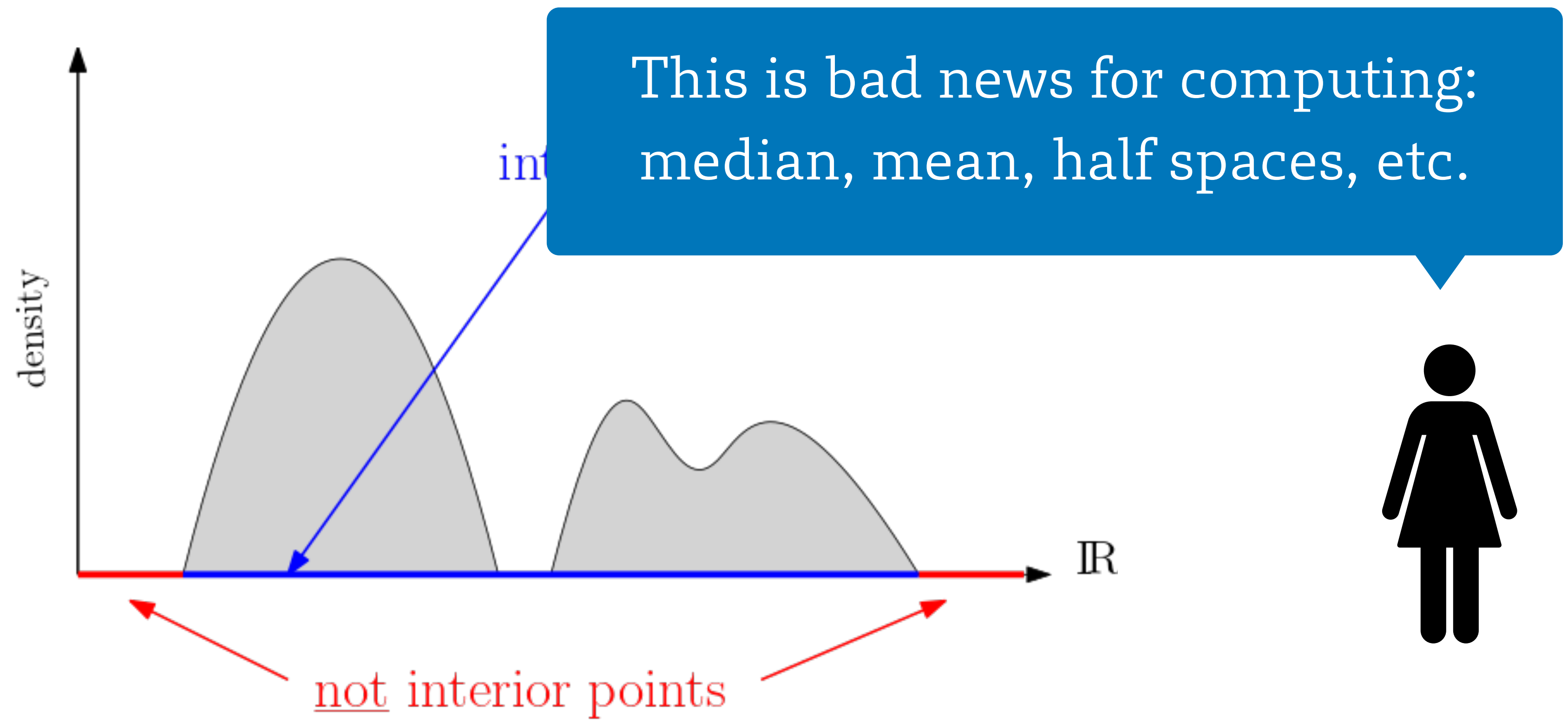


An algorithm A is (ϵ, δ)-**Differentially Private** if

- for all pairs of "neighboring" datasets D, D'
- for all events $E \subseteq \text{Range}(A)$

$$\Pr[A(D) \in E] \leq e^{\epsilon} \Pr[A(D') \in E] + \delta$$

A Surprising Impossibility Result (BNSV '15)



Theorem (BNSV'15). Any (ϵ, δ)-differentially private algorithm that solves the interior point problem must use at least n samples, where

$$n = \Omega(\log^* |\,\text{domain of } P\,|)$$

Immediate Corollary: When P is continuous, the problem is intractable!

Related Work: Bypassing the Impossibility Result

	Distributional Assumption
[KV18]	Assume data is Gaussian
[DL09, TVGZ20, BAM20, AD20]	Assume probability density is lower bounded at every point in some fixed-sized interval around the median
[HRS20]	Assume a smoothness property everywhere

Takeaway: Bun et al's lower bound seems to apply only to very "unusual" distributions.

This Work

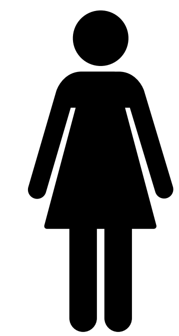
Theorem 1 (Informal). Assume P satisfies C-bounded normalized variance. Then there is an (ϵ, δ)-DP algorithm that:

1. returns an interior point of P and
2. uses $n = \mathrm{poly}(C\epsilon^{-1} \log \delta^{-1})$ samples.

We define this distributional assumption next!

What is a C-bounded distribution?

Examples include:
Uniform, Gaussian, Exponential,
Laplace, Binomial, Poisson, etc.



Definition:

- A distribution P with mean μ satisfies C-bounded normalized variance if

$$\underbrace{\sqrt{\mathbb{E}_{X \leftarrow P}[\,|X-\mu|^2\,]}}_{\text{standard deviation}} \leq C \cdot \underbrace{\mathbb{E}_{X \leftarrow P}[\,|X-\mu|\,]}_{\text{expected absolute deviation}}$$

Intuition:

- Distributions with $O(1)$-bounded normalized variance are those for which the standard deviation serves as a constant-factor proxy for the expected absolute deviation $\mathbb{E}[\,|X-\mu|\,]$.

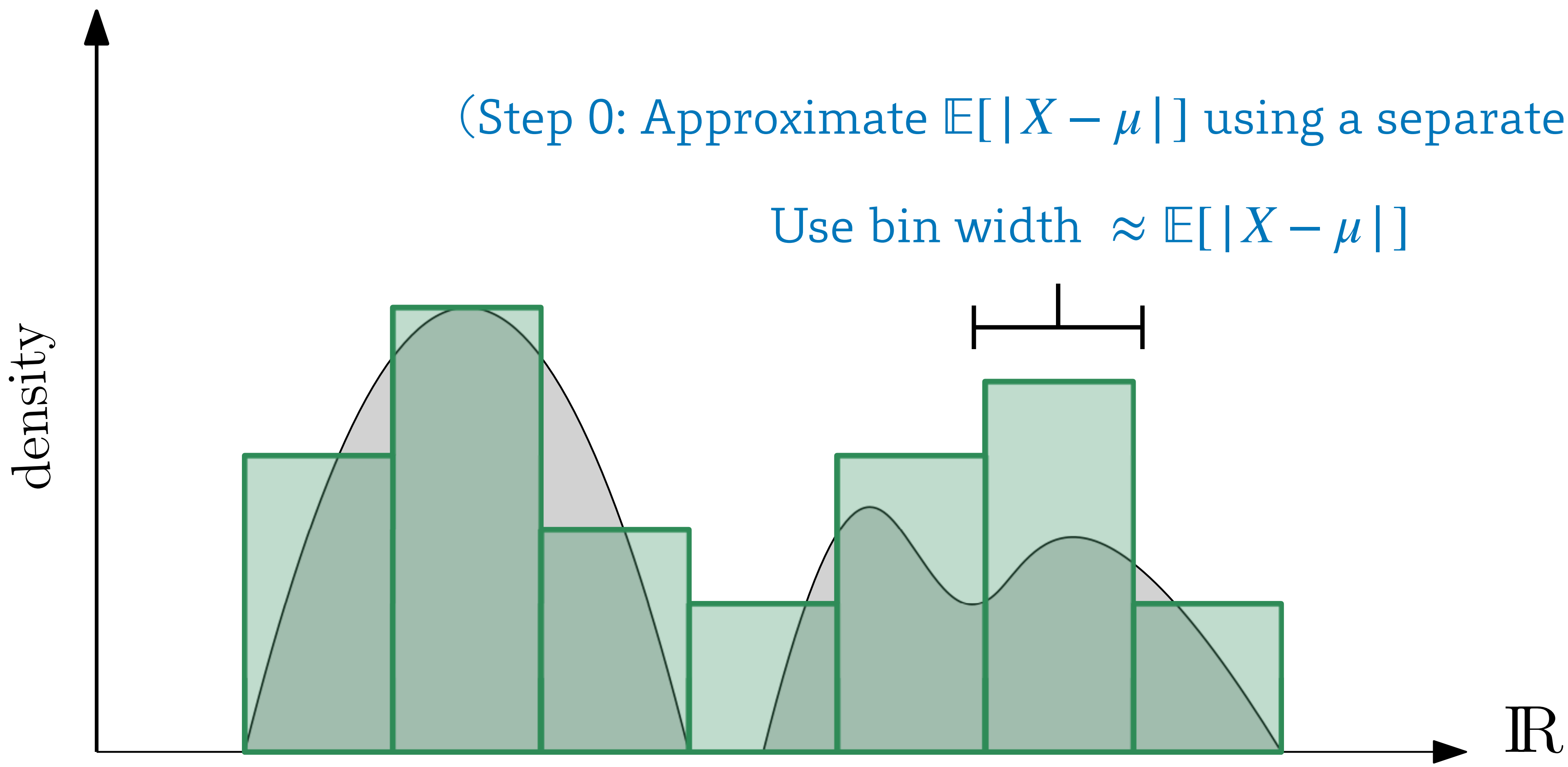
This Work

Theorem 1 (Informal). Assume P satisfies C-bounded normalized variance. Then there is an (ϵ, δ)-DP algorithm that:

1. returns an interior point of P and
2. uses $n = \mathrm{poly}(C\epsilon^{-1} \log \delta^{-1})$ samples.

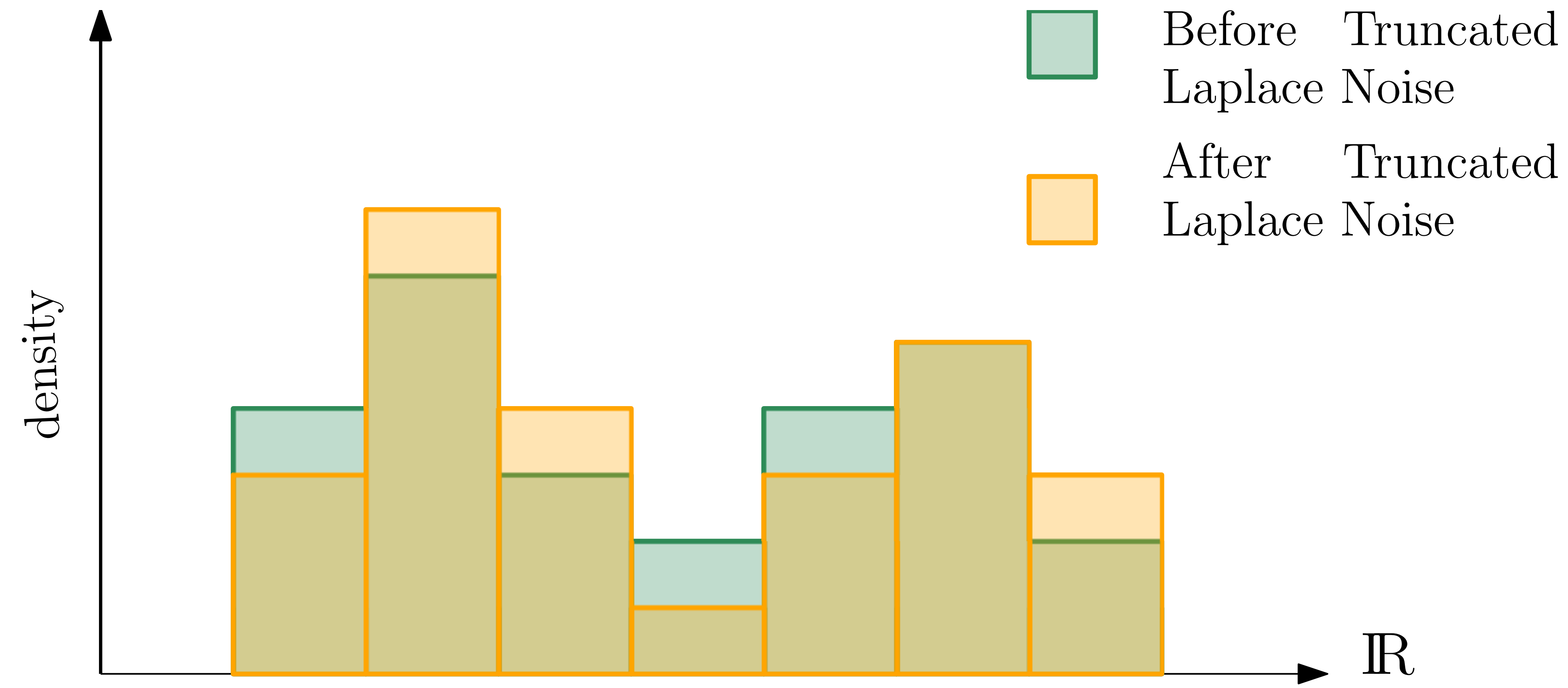
High-Level Algorithm Overview

Step 1: Place points into histogram bins



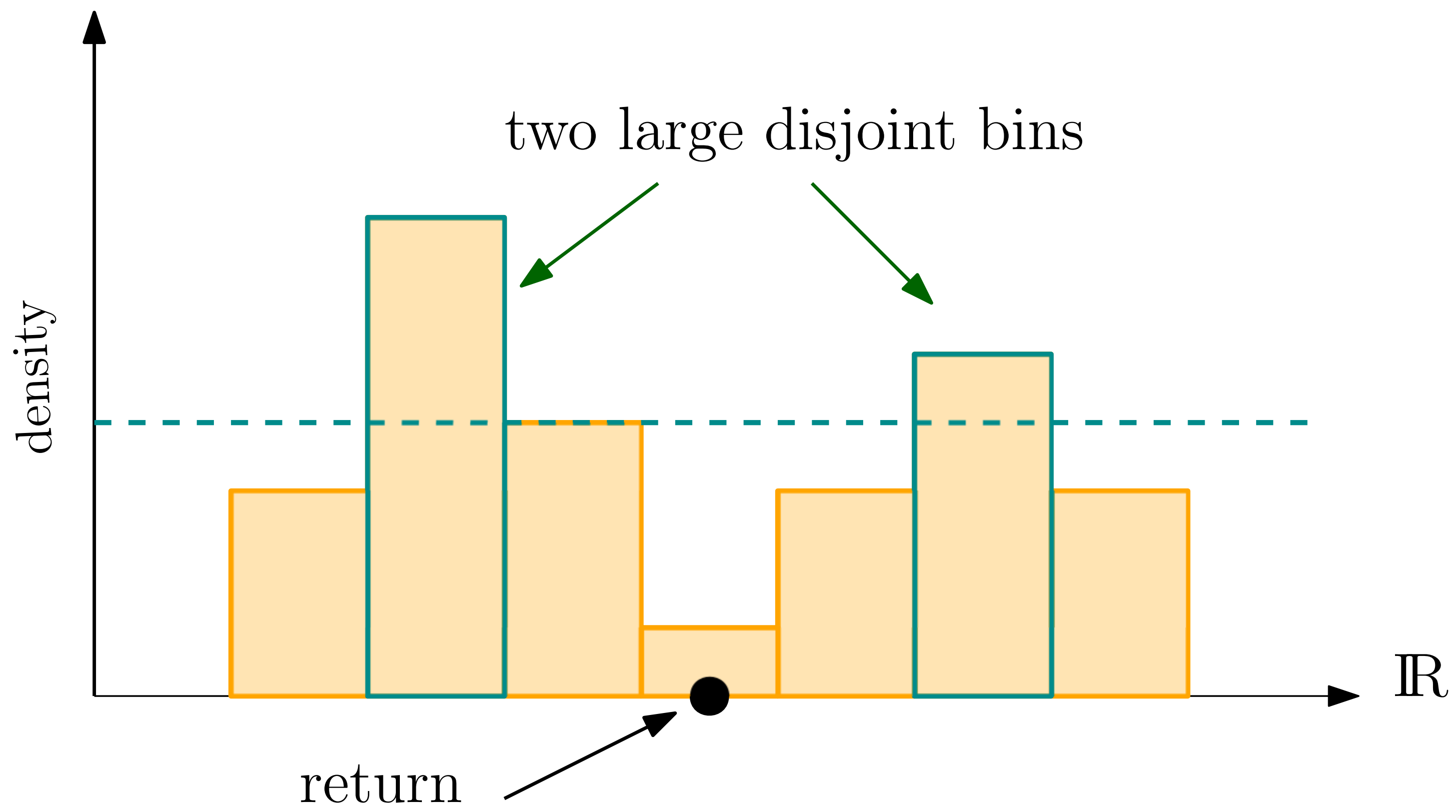
- The domain of P is partitioned into contiguous bins B of a fixed width
- Each bin counts the number of samples $X_1, \ldots, X_n$ that land in the bin

Step 2: Add truncated Laplacian noise to each bin



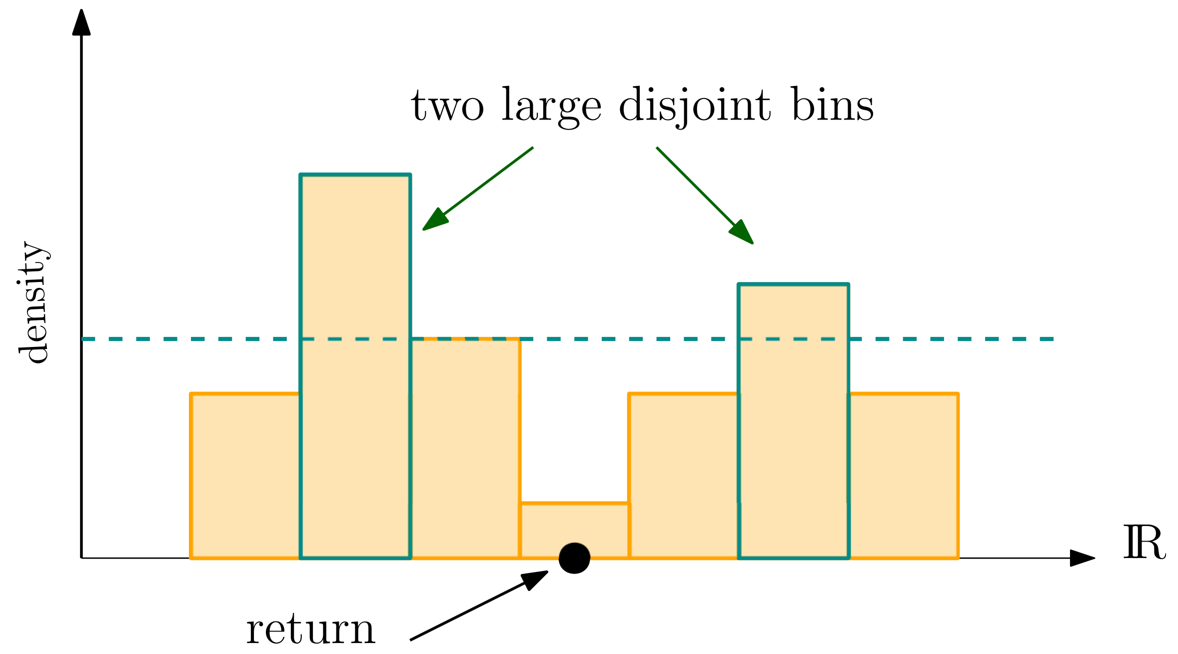
- Each bin receives a random amount of noise sampled from a truncated Laplace distribution
- The Laplace noise ensures differential privacy

Step 3: Find two bins with sufficiently large loads, and return any point between them.



- Intuitively, any point in-between two very full bins must be an interior point
- We are not required to know where the samples are in these two full bins, which is convenient for privacy

How C-Boundedness Helps



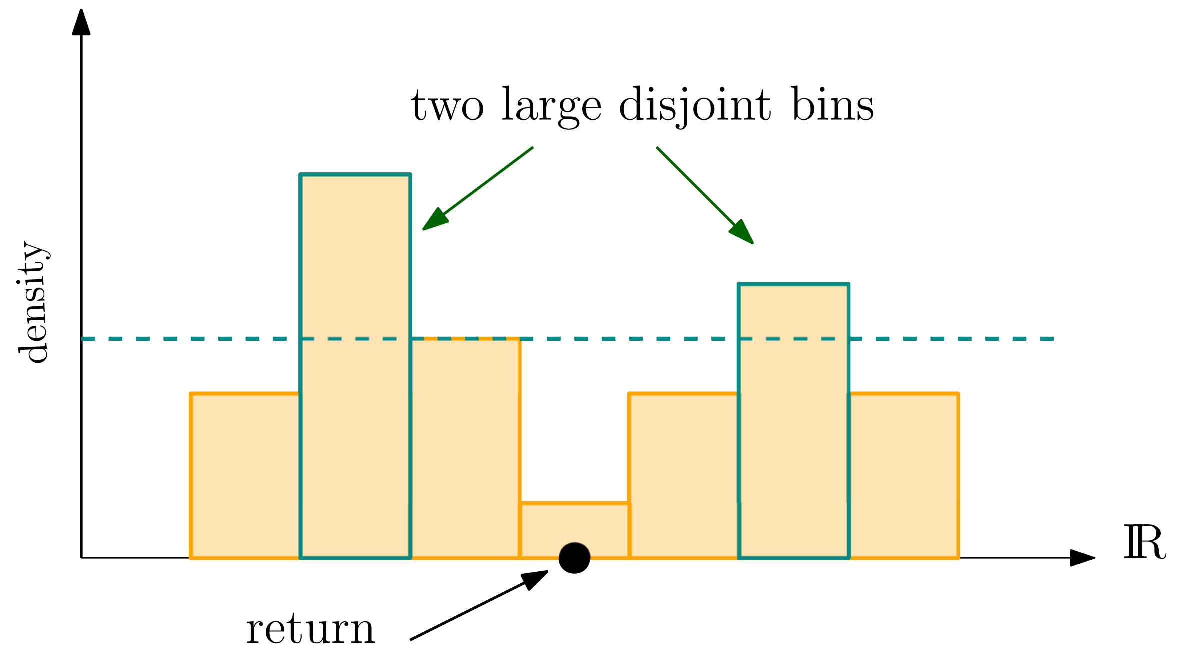
Problem 1:

- The samples could be too spread out, so that there are no large bins

Key Idea:

- By Chebyshev's Inequality, a large fraction of mass is concentrated within a constant number of standard deviations
- By C-boundedness, a large fraction of mass thus is concentrated within a constant number of bins

How C-Boundedness Helps



Problem 2:

- The samples could be too concentrated, so that there is only one large bin

Key Idea:

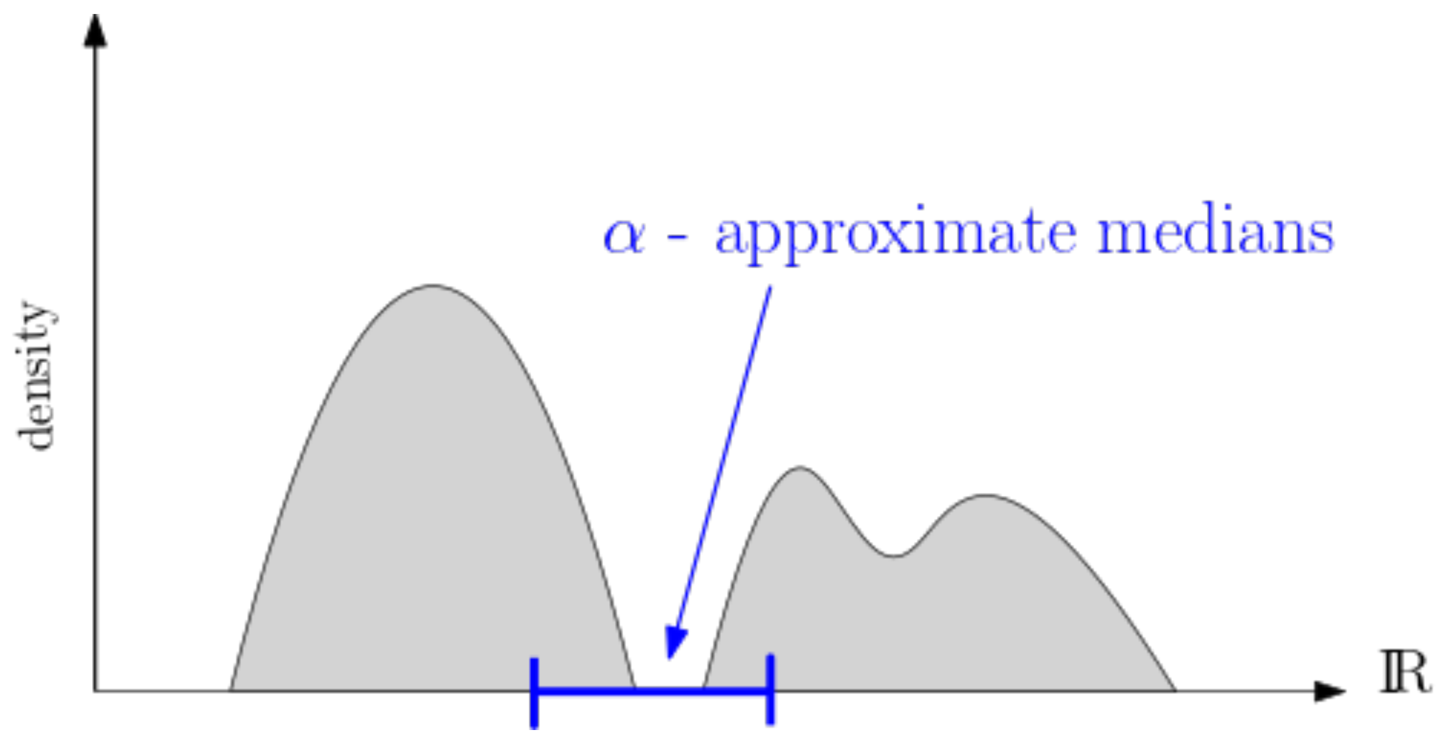
- C-boundedness tells us that outliers are not what determines std dev
- Suppose that the points are very concentrated in one bin
- Then the std dev is smaller than the size of a bin
- But this contradicts the definition of the bin size!

This Work

Theorem 1 (Informal). Assume P satisfies C-bounded normalized variance. Then there is an (ϵ, δ)-DP algorithm that:

1. returns an interior point of P and
2. uses $n = \mathrm{poly}(C\epsilon^{-1} \log \delta^{-1})$ samples.

The Private Approximate Median Problem



Private α-Approximate Median Problem:

- Given n i.i.d. samples $X_1, \ldots, X_n \sim P$, privately return a point y between the $0.5 - \alpha$ and $0.5 + \alpha$ quantiles of the distribution

This Work

Theorem 2 (**Informal**). Assume P satisfies C-bounded normalized variance around the median. Then there is an (ϵ, δ)-DP algorithm that:

1. returns an α-approximate median of P and

2. uses $n = \mathrm{poly}(C\alpha^{-1}\epsilon^{-1}\log\delta^{-1})$ samples.

Theorem 1 (**Informal**). Assume P satisfies C-bounded normalized variance. Then there is an (ϵ, δ)-DP algorithm that:

1. returns an interior point of P and

2. uses $n = \mathrm{poly}(C\epsilon^{-1}\log\delta^{-1})$ samples.

Conclusion

- Differential Privacy makes even the simplest problems challenging
- A single framework formalizing the intuition that the lower bound applies only to pathological distributions
- Algorithms for interior point problem and approximate medians

Thank you!