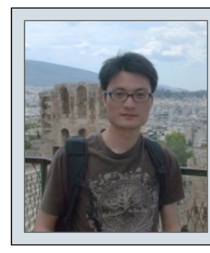
MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS


Zengfeng Huang
Aarhus University

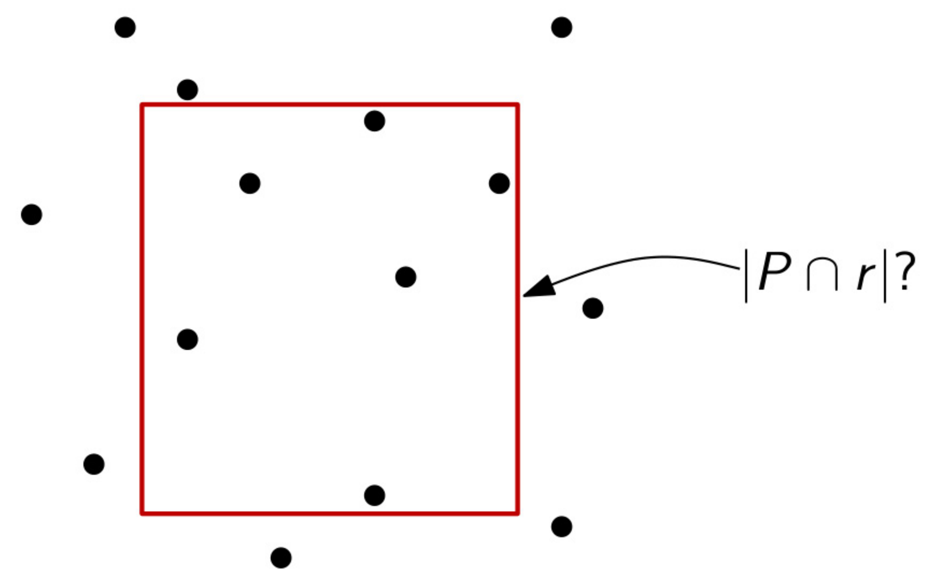
Danmarks Grundforskningsfond
Danish National Research Foundation

The Communication Complexity of Distributed ε-approximations

Background

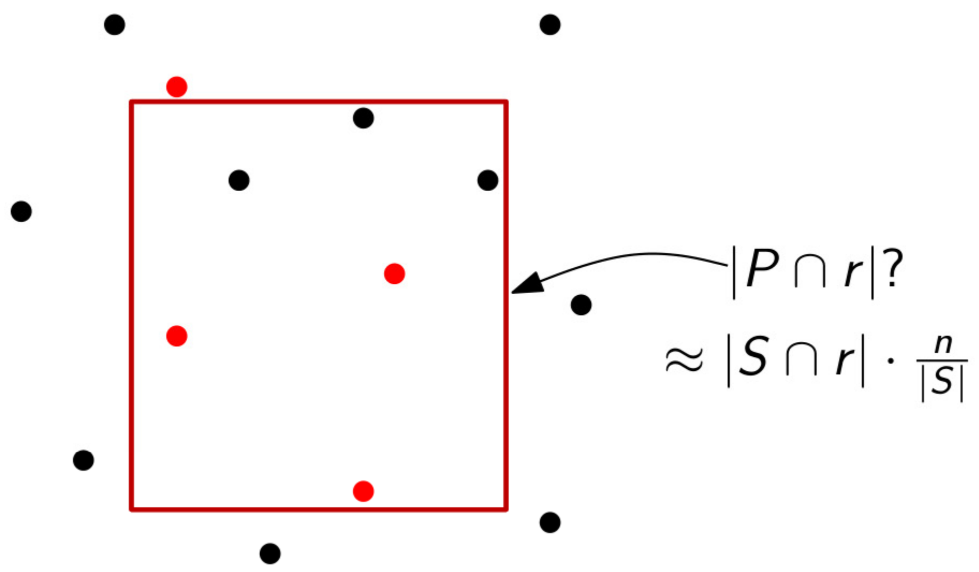
Motivations for *ε-approximations*

- Approximate range counting



Let P be a set of n points in the plane. Compute a summary structure so that, for any range r (from a certain range space), $|P \cap r|$ can be extracted with error εn

- An *ε-approximation* is a subset of the ground point set



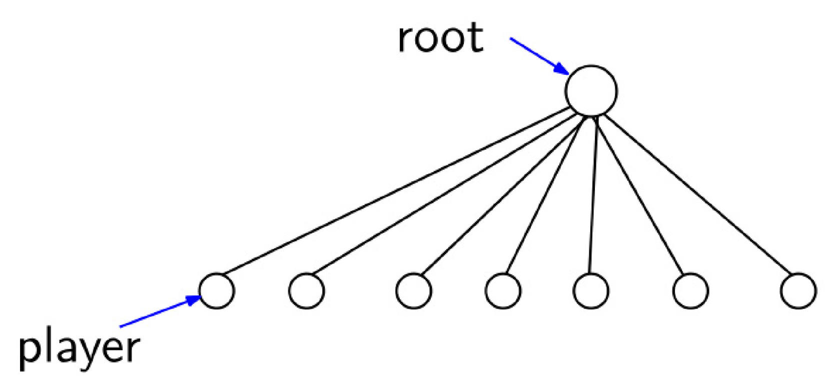
$S \subseteq P$ is an ε-approximation of P if for any r (from a certain range space),

$$\left| \frac{|P \cap r|}{n} - \frac{|S \cap r|}{|S|} \right| \leq \varepsilon$$

- In centralized setting, the goal is to minimize the size of S and the problem is well-studied
- In this work we study this problem in distributed setting

Distributed Model

Points are distributed across k players



- Each player i get a set X_i of n_i points
- All the player exchange messages with the root
- The root output an ε-approximation of the whole set

- This is a simplified and abstract model for real distributed systems
- In this setting, the goal is to minimize the communication cost
- Communication is a bottleneck in many applications

Simple Algorithms

Communication cost for straightforward algorithms

1. Random sampling

   $O(k + \frac{1}{\varepsilon^2})$

2. Combining local ε-approximations (deterministic)

   S_1 is an ε-approximation of P_1 and S_2 is an ε-approximation of P_2
   Then $S_1 \cup S_2$ is an ε-approximation of $P_1 \cup P_2$

   $O(\frac{k}{\varepsilon})$ for 1-d intervals

There is a trade-off between k and ε

- Each approach is better for certain values of k and ε
- Can we improve?

Our Results

Upper bounds

- We give an general algorithm for any *range space*, and the communication cost is better than the above two algorithms
- The algorithm is based on *combinatorial discrepancy*
- Results for some common range space are listed in the next table

	Deterministic O	Randomized O
Intervals in $\mathbb{R}^1$	k/ε	$\frac{\sqrt{k}}{\varepsilon}\sqrt{\log\frac{1}{\varepsilon}}$
Boxes in $\mathbb{R}^d$	$\frac{k}{\varepsilon}\log^{d+\frac{1}{2}}\frac{1}{\varepsilon}$	$\frac{\sqrt{k}}{\varepsilon}\log^{d+1}\frac{1}{\varepsilon}$
Halfspaces in $\mathbb{R}^d$	$k/\varepsilon^{\frac{2d}{d+1}}$	$k^{\frac{1}{d+1}}/\varepsilon^{\frac{2d}{d+1}} \cdot \log^{\frac{d}{d+1}}\frac{1}{\varepsilon}$

Lower Bounds

- We propose two methods to prove communication lower bounds for this problem
- Using these two methods, we prove near optimal lower bounds for interesting range space in any dimension
- We prove the first near optimal deterministic lower bounds, which shows that the algorithm of combining local ε-approximations is the best we can do deterministically
- Some lower bound results are listed below

	Deterministic Ω	Randomized Ω
Intervals in $\mathbb{R}^1$	k/ε	$\sqrt{k}/\varepsilon$
Boxes in $\mathbb{R}^d$	$\frac{k}{\varepsilon}\log\frac{1}{\varepsilon}$	$\frac{\sqrt{k}}{\varepsilon}\log\frac{1}{\varepsilon}$ for $d = 2, 3, 4$; $\frac{\sqrt{k}}{\varepsilon}\log^{\frac{d-3}{2}}\frac{1}{\varepsilon}$ for all d
Halfspaces in $\mathbb{R}^d$	$k/\varepsilon^{\frac{2d}{d+1}}$	$k^{\frac{1}{d+1}}/\varepsilon^{\frac{2d}{d+1}}$

References

[1] Zengfeng Huang and Ke Yi. *The Communication Complexity of Distributed ε-approximations*. FOCS 2014.