Efficient Sketches for Earth-Mover Distance, with Applications

Khanh Do Ba

MIT

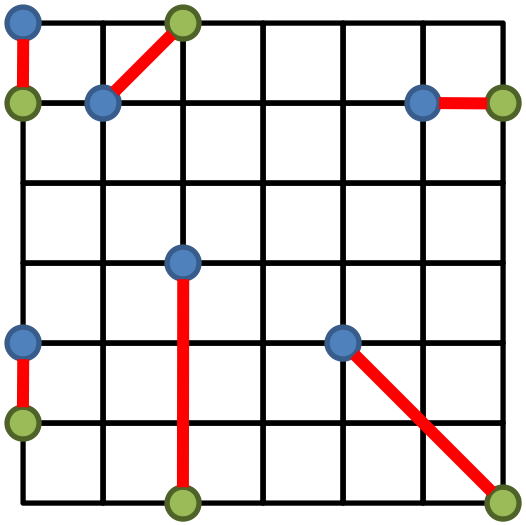
Joint work with Alexandr Andoni, Piotr Indyk and David Woodruff

(Planar) Earth-Mover Distance

- For multisets A, B of points in $[\Delta]^2$, $|A|=|B|=N$,

$$\mathrm{EMD}(A,B) = \min_{\pi:A\to B} \sum_{a\in A} \|a-\pi(a)\|$$

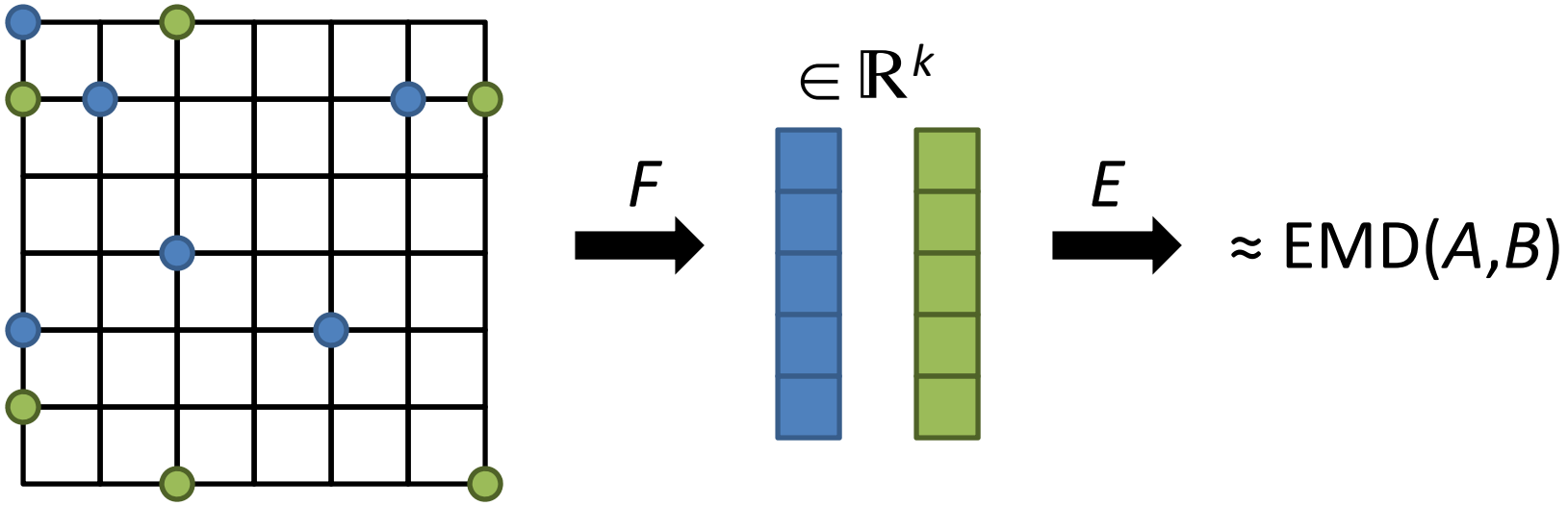
i.e., min cost of perfect matching between A and B



EMD(●,●) = 6 + 3√2

Geometric Representation of EMD

- Map A, B to k-dimensional vectors $F(A)$, $F(B)$
  - Image space of F "simple," e.g., k small
  - Can estimate EMD(A,B) from $F(A)$, $F(B)$ via some efficient recovery algorithm E



Geometric Representation of EMD: Motivation

- Visual search and recognition:
  - Approximate nearest neighbor under EMD
    - Reduces to approximate NN under simpler distances
    - Has been applied to fast image search and recognition in large collections of images [Indyk-Thaper'03, Grauman-Darrell'05, Lazebnik-Schmid-Ponce'06]

- Data streaming computation:
  - Estimating the EMD between two point sets given as a stream
    - Need mapping F to be linear: adding new point a to A translates to adding $F(a)$ to $F(A)$
    - Important open problem in streaming ["Kanpur List '06"]

Prior and New Results

Geometric representation of EMD:

Paper	Recovery	Dimension	Approx.
[Charikar'02, Indyk-Thaper'03]	ℓ_1	$O(\Delta^2)$	$O(\log \Delta)$
[Naor-Schechtman'06]	ℓ_1	Any	$\Omega(\log^{1/2} \Delta)$
Our result	**Non-norm**	$\boldsymbol{O(\Delta^{\varepsilon})}$	$\boldsymbol{O(1/\varepsilon)}$

Main Theorem

For any $\varepsilon \in (0,1)$, there exists a distribution over linear mappings $F: \mathbb{R}^{\Delta^2} \to \mathbb{R}^{\Delta^{\varepsilon}}$ s.t. for multisets $A,B \subseteq [\Delta]^2$ of equal size, we can produce an $O(1/\varepsilon)$-approximation to EMD(A,B) from $F(A)$, $F(B)$ with probability 2/3.

Implications

- Streaming:

Paper	Space	Approximation
[Indyk'04]	$\log^{O(1)}(\Delta N)$	$O(\log \Delta)$
Our result	$\boldsymbol{\Delta^{\varepsilon} \log^{O(1)}(\Delta N)}$	$\boldsymbol{O(1/\varepsilon)}$

* N = number of points

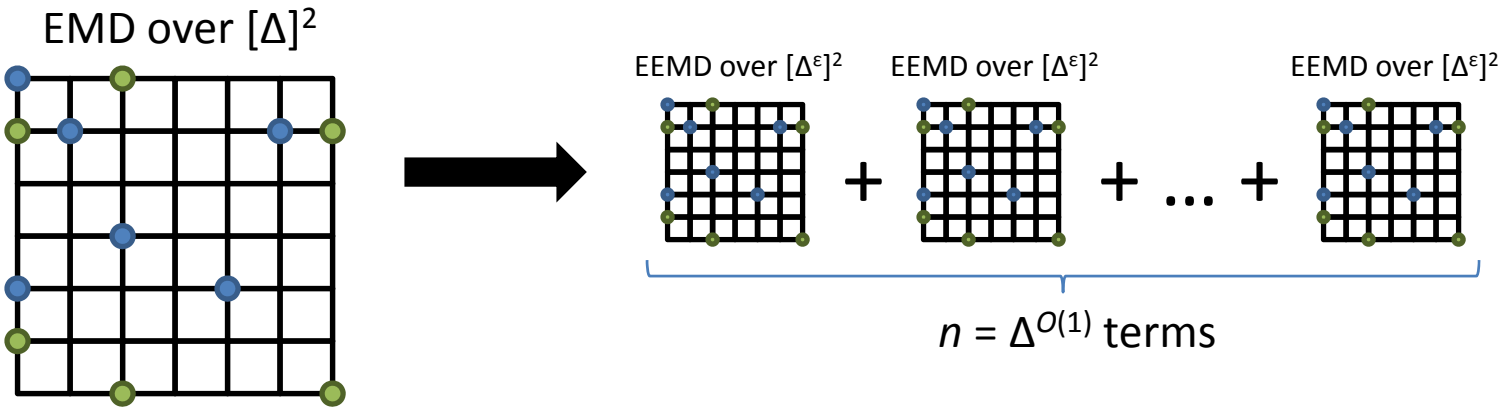
- Approximate nearest neighbor:

Paper	Space	Query time	Approximation
[Andoni-Indyk-Krauthgamer'09]	$s^{2+\varepsilon}\, 2^{\Delta^{1/\alpha}}$	$\Delta^{O(1)}\, s^{\varepsilon}$	$O((\alpha/\varepsilon) \operatorname{loglog} s)$
Our result	$\boldsymbol{2^{\Delta^{\varepsilon} \log(s\Delta N)^{O(1)}}}$	$\boldsymbol{(\Delta \log(s\Delta N))^{O(1)}}$	$\boldsymbol{O(1/\varepsilon)}$

* s = number of data points (multisets) to preprocess
$\alpha>1$ free parameter

Proof Outline

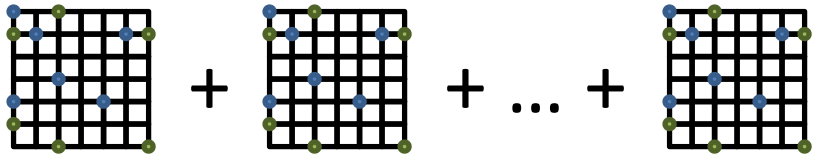
- Old [Agarwal-Varadarajan'04, Indyk'07]:
  - Extend EMD to EEMD which:
    - Handles sets of unequal size
    - Is induced by a norm $\|\cdot\|_{EEMD}$, i.e., $EEMD(A,B) = \|\chi(A) - \chi(B)\|_{EEMD}$, where $\chi(A) \in \mathbb{R}^{\Delta^2}$ is the characteristic vector of A
  - Decomposition of EEMD into weighted sum of small EEMD's
    - $O(1/\varepsilon)$ distortion



- New:
  - Linear sketching of "sum-norms"

Main Technical Theorem

$$\|x\|_{1,X} = \underbrace{\text{[grid]}}_{\|x_1\|_X} + \underbrace{\text{[grid]}}_{\|x_2\|_X} + \ldots + \underbrace{\text{[grid]}}_{\|x_n\|_X}$$



For normed space $X = (\mathbb{R}^m, \|\cdot\|_X)$ and $x \in X^n$, denote $\|x\|_{1,X} = \sum_i \|x_i\|_X$.

Fix $n \in \mathbb{N}$, $M > 0$ and $\gamma > 1$. There exists a distribution over linear mappings

$$\mu: X^n \to X^{(\gamma \log n)^{O(1)}}$$

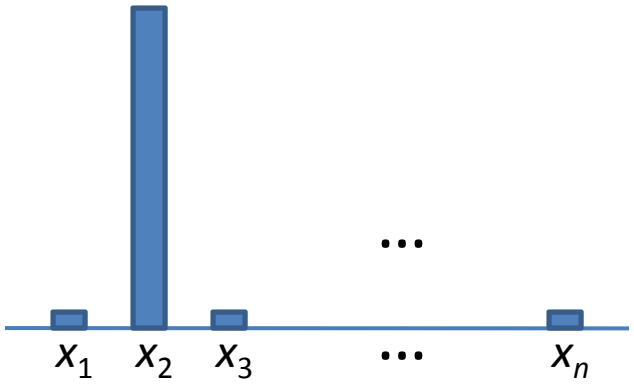
such that for any $x \in X^n$ satisfying

$$M/\gamma \le \|x\|_{1,X} \le M,$$

we can produce an $O(1)$-approximation to $\|x\|_{1,X}$ from $\mu(x)$ whp.

Proof Outline: Sum of Norms

- First attempt:
  - Sample (uniformly) a few x_i's to compute $\|x_i\|_X$
  - Problem: sum could be concentrated in 1 block

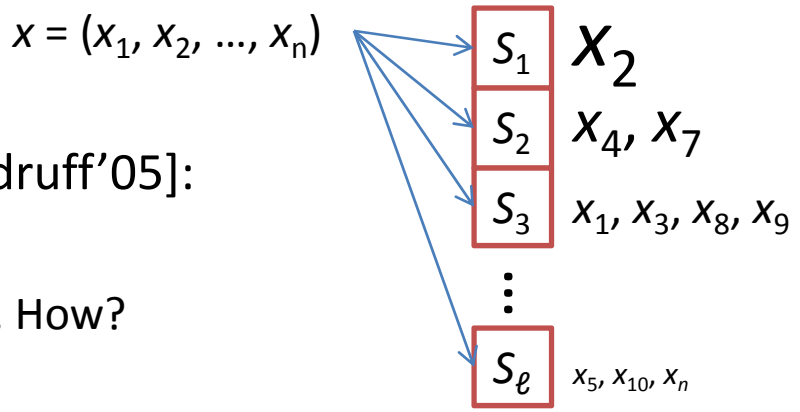


x_2 contains most of mass

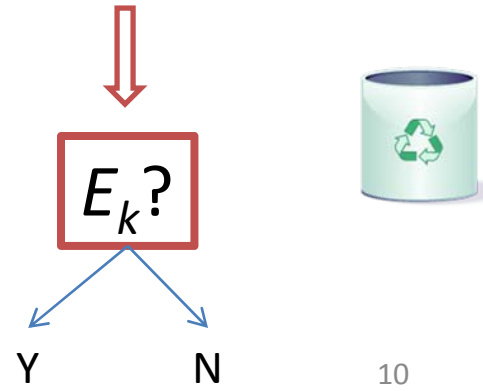
- Second attempt:
  - Sample x_i with probability roughly $\propto \|x_i\|_X$ [Indyk'07]
  - Problem: how to do online?
  - Techniques from [Jayram-Woodruff'09]?
    - Need to sample/retrieve blocks, not just individual coordinates

Proof Outline: Sum of Norms (cont.)

- Our approach:
  - Split into exponential levels [Indyk-Woodruff'05]:
    - $S_k = \{i \in [n] \mid \|x_i\|_X \in (T_k, 2T_k]\}$, $T_k = M/2^k$
    - Suffices to estimate $|S_k|$ for each level k. How?
  - For each level k, subsample from $[n]$ at a rate such that event E_k ("isolation" of level k) holds with probability $\propto |S_k|$
  - Repeat experiment several times, count number of successes



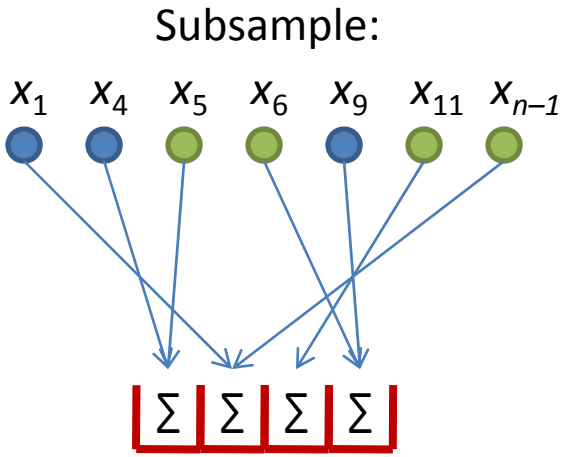
Subsample:



Proof Outline: Event E_k

- $E_k \Leftrightarrow$ "isolation" of level k:
  - Exactly one $i \in S_k$ gets subsampled
  - Nothing from $S_{k'}$ for $k'<k$

- Verification of trial success/failure
  - Hash subsampled elements
    - Each cell maintains sum of subsampled x_i's that hash there
  - E_k holds roughly (we "accept") when:
    - 1 cell has X-norm in $(0.9T_k, 2.1T_k]$
    - All other cells have X-norm $\leq 0.9T_k$
  - Check fails only if:
    - Elements from lighter levels contribute a lot to 1 cell
    - Elements from heavier levels subsampled and collide
  - Both unlikely if hash table big enough
  - Remark: triangle inequality of norm gives control over impact of collisions



Sketch and Recovery Algorithm

Sketch:

- For each level k, create t hash tables
- For each hash table:
  - Subsample from $[n]$, including each $i \in [n]$ w.p. p_k
  - Each cell maintains sum of x_i's that hash to it

Recovery algorithm:

- For each level k, count number c_k of "accepting" hash tables
- Return $\sum_k T_k \cdot (c_k/t) \cdot (1/p_k)$

Conclusion and Open Problems

- We achieve a linear embedding of EMD
  - with constant distortion, namely $O(1/\varepsilon)$,
  - into a space of strongly sublinear dimension, namely Δ^{ε}.

- Open problems:
  - Getting (1+ε)-approximation / proving impossibility
  - Reducing dimension to $\log^{O(1)}\Delta$ / proving lower bound