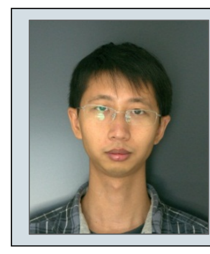
MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS



Wei Yu
Aarhus University

Danmarks Grundforskningsfond
Danish National Research Foundation

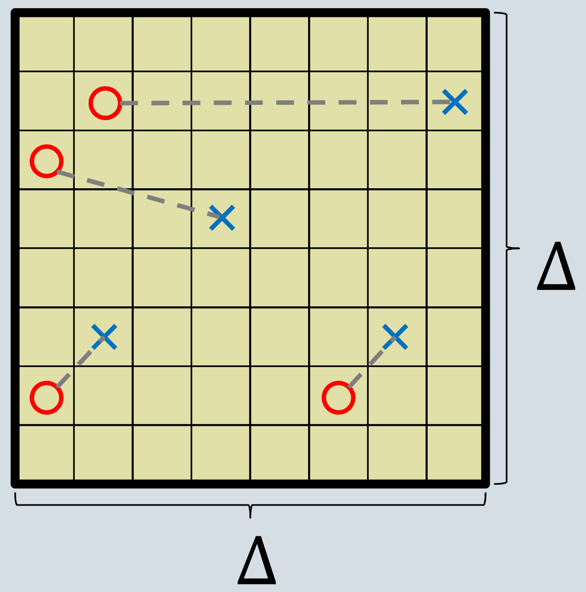
Error Correcting under Earth-Mover Distance

The Problem

Earth-Mover Distance

- Two point sets in d dimension
  - $x = \{x_1, x_2, \dots, x_n\} \in [\Delta]^d$
  - $y = \{y_1, y_2, \dots, y_n\} \in [\Delta]^d$
- Earth Mover Distance (EMD):

$$\mathrm{EMD}(x,y) = \min_{\pi:[n]\to[n]} \sum_{1\le i\le n} \|x_i - y_{\pi(i)}\|_2$$

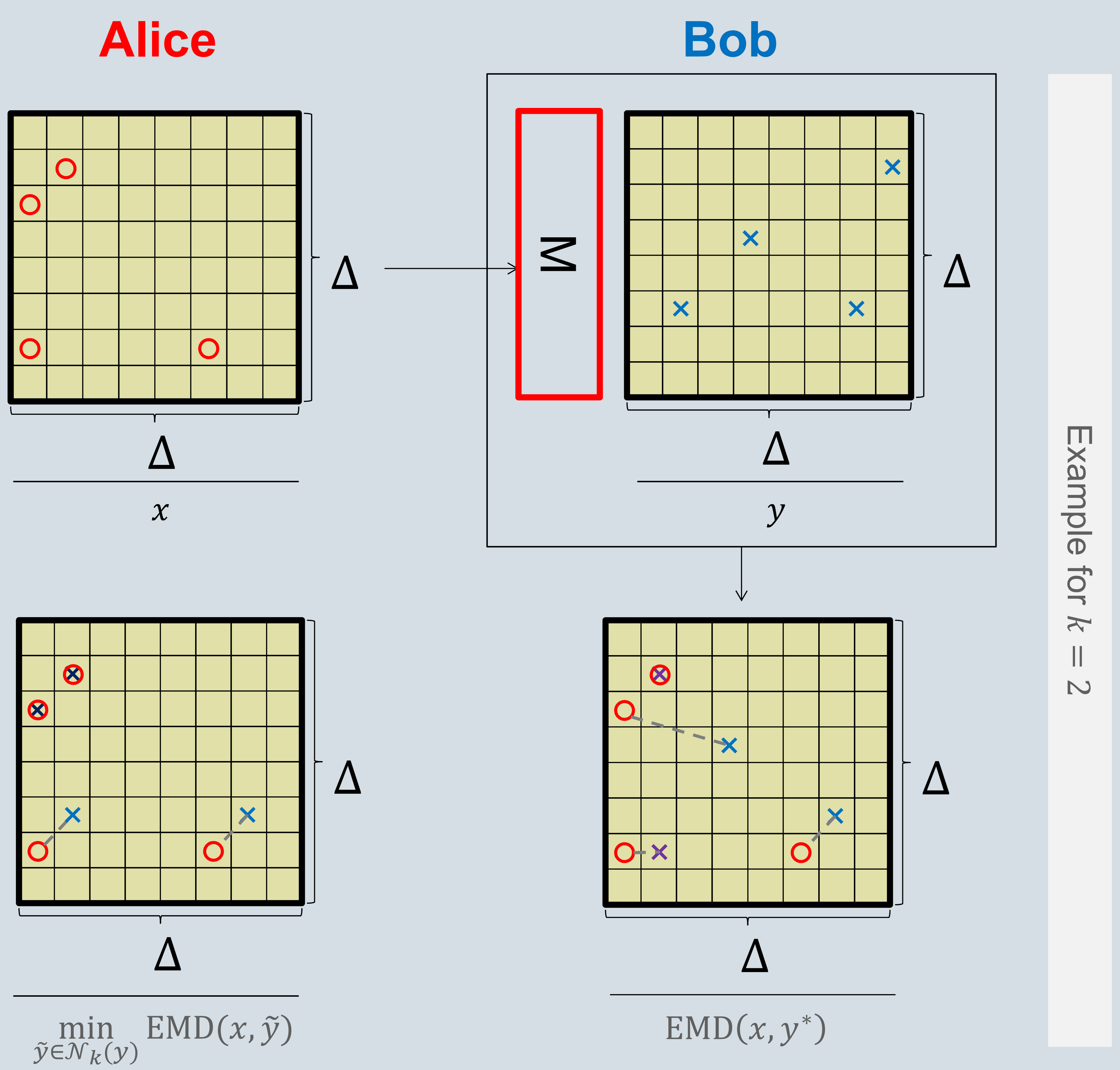


Earth-Mover Distance with Error Correcting

- Alice holds x, Bob holds y, both of them know parameter k
- Alice sends a message M to Bob
- Bob computes y^*
- For some approximation factor C
- We want

$$\mathrm{EMD}(x,y^*) \le C \cdot \min_{\tilde{y}\in\mathcal{N}_k(y)} \mathrm{EMD}(x,\tilde{y})$$

- $\mathcal{N}_k(y)$: set of all point sets obtained by relocating k points in y



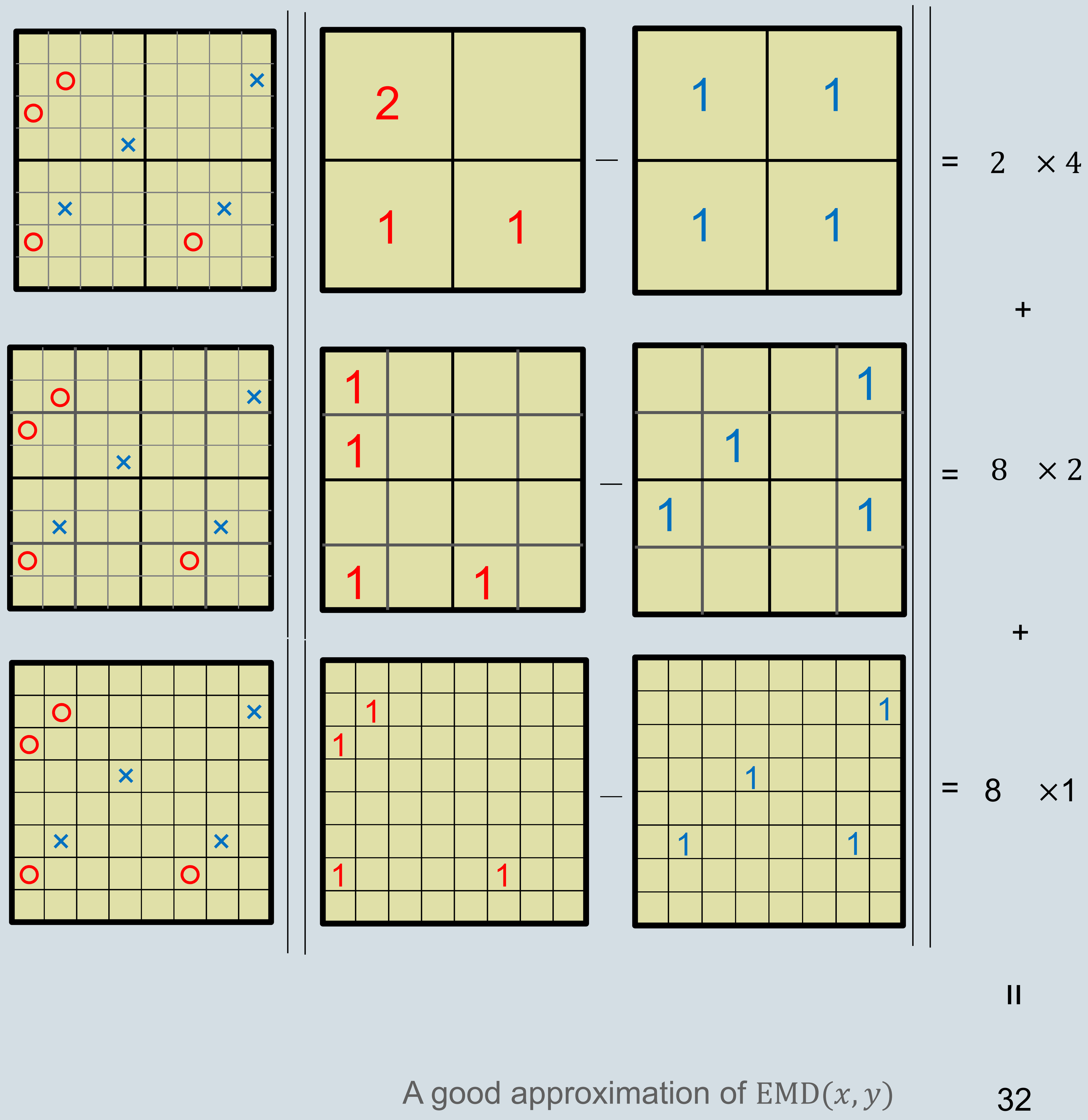
*Note: All examples are for $d = 2$

Upper Bound

Result

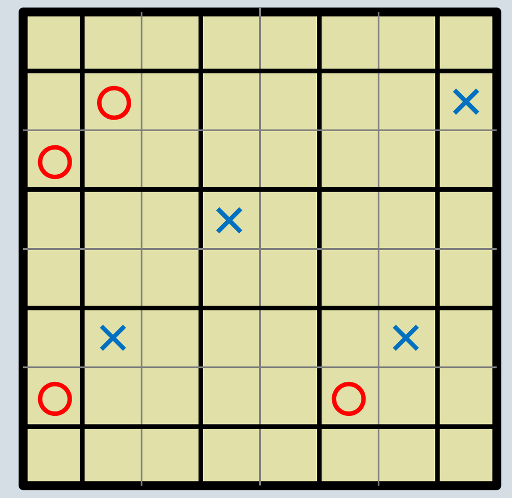
- Sketch size $|M| = O(k \log\Delta \log(n\Delta^d))$
- Approximation $C = O(d)$
- Encoding and decoding time: $O(k \log\Delta \log^3(n\Delta^d))$

Pyramid Array Approximation



Rounding Error and Random Shift

- Random shift the grid by a distance
- With high probability close pairs NOT being "cut"
- Distant pairs always get "cut" regardless of the shift
- Better approximation with high probability



Important Level (for $k = 2$)

Look at what happens on the same level of

$$A = \min_{\tilde{y}\in\mathcal{N}_k(y)} \mathrm{EMD}(x,\tilde{y})$$

Unimportant Levels

- ℓ_1 difference $> 2k$
- Random correcting is already a good approximation of A
- All levels below are unimportant

Important Levels

- ℓ_1 difference $\le 2k$
- Must rebuild the correct red pyramid array
- Error-correcting code for $2k$ errors ($O(k\log n)$) in Hamming distance
- Differences = Errors
- All level above are important

y^* =correct important levels + random unimportant levels

$\Rightarrow \mathrm{EMD}(x,y^*)$ is a good approximation of $\min_{\tilde{y}\in\mathcal{N}_k(y)} \mathrm{EMD}(x,\tilde{y})$

Motivation

- A good error-correcting code under EMD
- Sending image incrementally
- Detect (a few) moving objects
- Streaming computation of image sketches
- Also towards the (hard) problem of sketching EMD

Lower Bound

- Sketch size: $\Omega(k \log\Delta \log(\Delta^d/k)/\log C)$ for C approximation
- Communication Complexity
  - Reduction from Augmented Indexing
  - Randomized
- Hard case: $n = O(k\Delta^{\frac{d}{2}} \log\Delta)$
- Implying the upper bound is (nearly) optimal up to log factors

References

[1] Christian Konrad, Wei Yu and Qin Zhang. *Budget Error-Correcting under Earth-Mover Distance*. Manuscript, 2013.

MADALGO – Center for Massive Data Algorithmics, a Center of the Danish National Research Foundation