Approximating edit distance

čtvrtek 30. října 2025 0:13

Thm: Edit distance can be approximated within $O(1)$-factor in time $\tilde{O}(n^{1.9})$.

Best known: $O(1)$-factor approx in time $n^{1+\varepsilon}$ $\forall$ const. $\varepsilon > 0$.

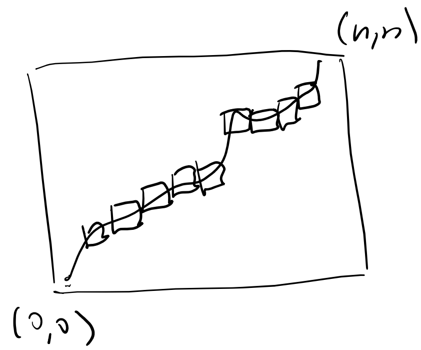
$O(1)$ depends doubly exponentially on $1/\varepsilon$

Alg. **Idea:** want to test whether edit distance is $< \varepsilon n$ or $\geq C \cdot \varepsilon n$ for some fixed constant C (approx factor)

$\to$ try $\varepsilon = 2^{-i}$ $\quad i = 0, \dots \log n^{1/10}$.

For $\varepsilon \leq \frac{1}{n^{1/10}}$ $\to$ compute E.D. exactly in time $\left(n^{9/10}\right)^2 = n^{1.8}$ using $\tilde{O}(n+k^2)$ alg.

set $\theta = \frac{1}{n^{1/10}}$, consider the E.D. graph for x & y

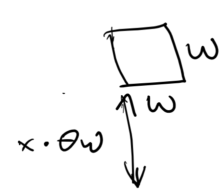


goal: cover the shortest path from $(0,0) \to (n,n)$ by boxes of size $w_1 \times w_1$ or $w_2 \times w_2$

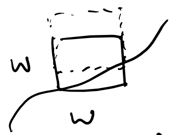
for $w_1 = n^{1/10}$ & $w_2 = n^{3/10}$

(WLOG w_1, w_2, n & $1/\theta$ are powers of 2)

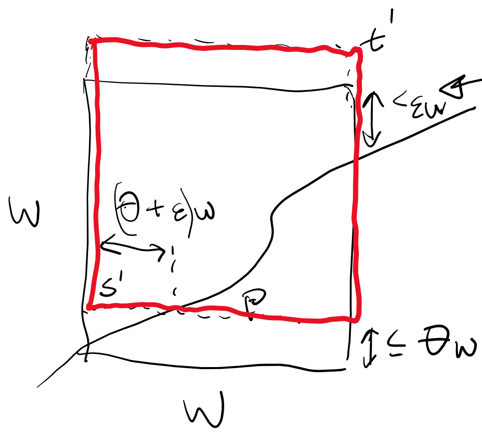
we say a $w \times w$ box is θ-aligned if it starts at a horizontal line that is a multiple of θw:



θ - aligned



- replacing a w×w box by a θ-aligned w×w box increases its E.D cost by a constant factor + 2θw.



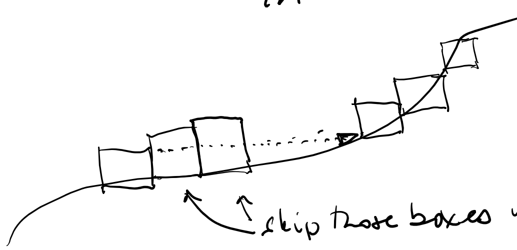
cost(P) ≤ εw

cost(s'→t') ≤ 3εw + 2θw

- a path of cost εn can be approximated using w×(1-2ε)w θ-aligned boxes of cost 5εn

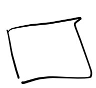
θ < ε

εn → 5εn



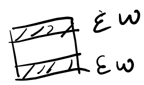
skip those boxes using a horizontal edges

θ-aligned w×w



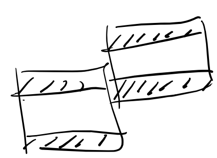
of cost εw

shrink →



w×(1-2ε)w

so they could connect to each other



Fact: Given a set B of w×w boxes, each with an estimate of its edit distance, we can find in time $\tilde{O}(|B|)$ the best path consisting of the boxes & horizontal & vertical edges.

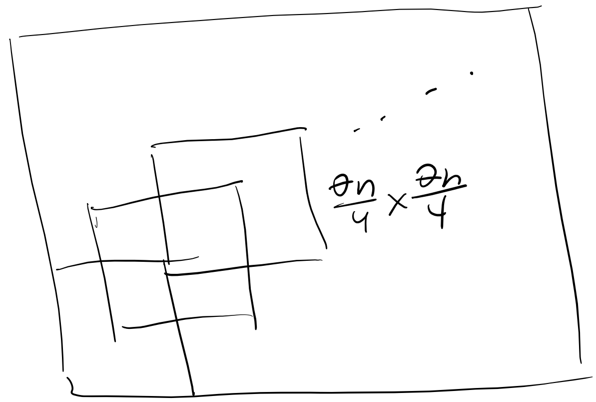
connecting (0,0) to (n,n)

→ We give an algorithm in time $\tilde{O}(n^{1.8})$ that identifies a set of $\le n^{1.8}$ w×w

... (0,0) to (n,n)

→ We give an algorithm in time $\tilde{O}(n^{1.8})$ that identifies a set of $\le n^{1.8}$ $w \times w$ boxes which well approximate $(0,0) \to (n,n)$ path.

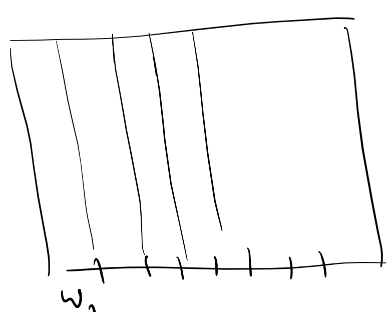
Two phases:
1) dense column analysis
2) sparse column diagonal extension



Cover the main diagonal by $\frac{\theta n}{4} \times \frac{\theta n}{4}$ large boxes & process each box in turn to cover it by small boxes.

processing one $\frac{\theta n}{4} \times \frac{\theta n}{4}$ box:

phase 1: for $\varepsilon = 2^{-i}$, $i = 0 \dots \lg \theta$



$d = n^{2/10}$

divide the box into columns of width $w_1 = n^{1/10}$. Out of $\frac{\theta n}{4} / \theta w_1 = \frac{n}{4 w_1}$ θ-aligned $w_1 \times w_1$ boxes in each column, sample at random $\frac{c \lg n}{d}$ - fraction of these boxes & compute their edit distance.

E.D.

→ time per column $\frac{n}{w_1} \times \frac{1}{d} \times w_1^2 = \frac{n \cdot w_1}{d}$

→ total time $\frac{n}{w_1} \times \frac{n \cdot w_1}{d} = \frac{n^2}{d} = n^{1.8}$

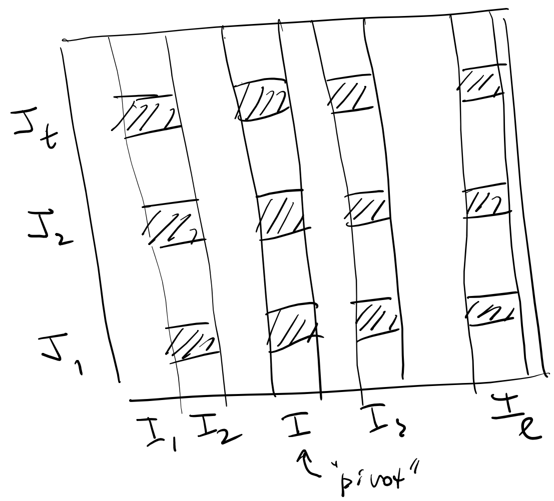
total # of columns

(ignoring $\tilde{O}(1)$ factors in time)

$\frac{c \lg n}{d}$

(ignoring $\hat{O}(1)$ factors in time)

Call a column (ε, d)-dense of more than $\frac{c \lg n}{2}$ of sampled boxes have $E.D. \le \varepsilon w_1$.



process columns I one by one.
If (ε, d)-dense then find all I_i s.t. $ED(x_I, x_{I_i}) \le 3\varepsilon w_1$ & find all J_j s.t. $ED(x_I, y_{J_j}) \le 2\varepsilon w_1$

→ declare all boxes $x_{I_i} \times y_{J_j}$ as having edit distance $\le 5\varepsilon$.

→ remove columns $I_1, \dots I_\ell$ from further processing.

- Notice, if I & I' are pivots chosen during the process then $J_I = \{J, ED(x_I, y_J) \le \varepsilon w_1\}$
$J_{I'} = \{J, ED(x_{I'}, y_J) \le \varepsilon w_1\}$
are disjoint.
(If $J \in J_I \cap J_{I'}$ then $ED(x_I, x_{I'}) \le 2\varepsilon w_1$)
so both I & I' cannot be pivots

→ time to process one dense column

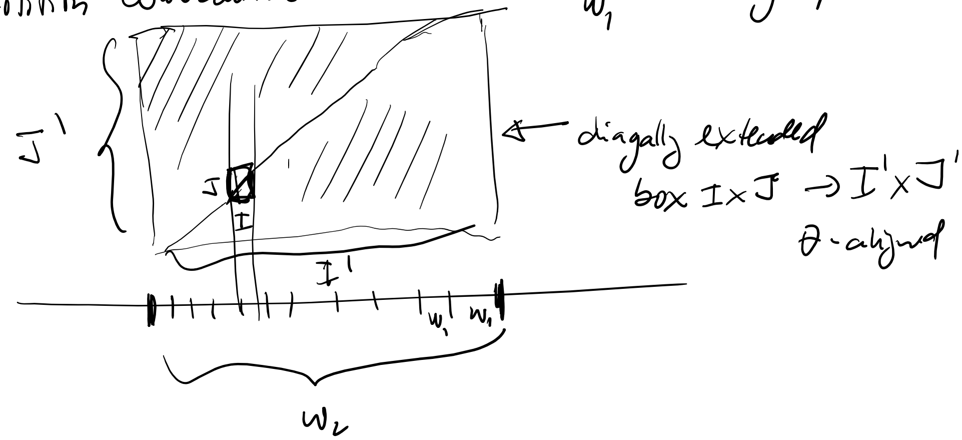
$$\frac{n}{w_1} \cdot w_1^2 = n \cdot w_1$$

→ time to process all dense columns

$$\underbrace{\frac{n}{w_1} \cdot \frac{1}{d}}_{\text{max \# of distinct pivots by claim}} \cdot n w_1 = \frac{n^2}{d} = n^{1.8}$$

$\underbrace{w_1 \quad 4}$ max # of distinct pivots by Claim

Phase 2: partition consecutive columns into $\frac{w_2}{w_1}$-size groups



process each group of columns one-by-one:

for each $\varepsilon = 2^i$, $i = 0, \dots \lg \theta$, "not dense"

sample $\frac{c \lg^2 n}{\theta}$ of (ε, d)-sparse columns within the group. In each sampled column I identify all $w_1 \times w_1$-boxes J where

$$ED(x_I, y_J) \le \varepsilon w_1,$$

and extend them to the whole group along its diagonal. Compute the edit distance of the $w_2 \times w_2$ box, output the $w_2 \times w_2$-box with its edit distance.

→ time to process all groups:

$$\underbrace{\frac{n}{w_2}}_{\#\text{groups}} \cdot \underbrace{\frac{1}{\theta}}_{\#\text{ of samples}} \cdot \underbrace{\left(\frac{n}{w_1} \cdot (w_1)^2 + d \cdot (w_2)^2\right)}_{\text{time to process each sample}} = \frac{n d w_2}{\theta} + \frac{n^2 \cdot w_1}{w_2 \cdot \theta}$$

(ignoring log-factors)

$1.6 \quad \, ^{1.9} = \tilde{O}(n^{1.9})$

#groups   #of samples   each sample
(ignoring log-factors)
$$= n^{1.6} + n^{1.9} = \tilde{O}(n^{1.9})$$

$w_1 = n^{1/10}$ $d = n^{2/10}$

$w_2 = n^{3/10}$ $\theta = \frac{1}{n^{1/10}}$

→ collect all $w_1 \times w_1$-boxes from phase 1 & $w_2 \times w_2$-boxes from phase 2 & stitch them together using the algorithm from the begining of the lecture.

$\# w_1 \times w_1$ boxes $\le \frac{n}{w_1} \cdot \frac{n}{w_1} = n^{1.8}$

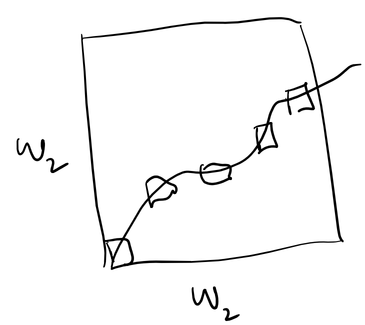
$\# w_2 \times w_2$ boxes $\le \frac{n}{w_2} \cdot \frac{n}{w_2} = n^{1.4}$

→ the stitching algorithm runs in time $\tilde{O}(n^{1.8})$.

Correctness:

1) With probability $\ge 1 - \frac{1}{n^3}$, we determine correctly for each (ε, d)-dense box that it is dense.

2) For a given $w_2 \times w_2$-box and path P through the box,



a) we found a cover of the path by $w_1 \times w_2$ boxes for all but $\theta \frac{w_2}{w_1}$ (sparse) columns. The path through that column will be counted as maximum thus adding to the total cost $\theta \frac{w_2}{w_1} \cdot w_1 = \theta w_2$.

thus adding to the total cost $\theta \frac{w_2}{w_1} \cdot w_1 = \theta w_2$.

b) Or a sparse column was sampled whose diagonal extention well approximates the path P.

(if # of (ε, d)-sparse columns, $\varepsilon = 2^{-d}$, that intersect path P in boxes of cost $\leq \varepsilon w_1$ is $\geq \frac{\theta}{\lg n} \cdot \frac{w_2}{w_1}$ then we are likely to sample such column & its diagonal extension well approximates the path.)