Lecture 5

TTIC 41000: Algorithms for Massive Data

Toyota Technological Institute at Chicago

Spring 2021

Instructor: Sepideh Mahabadi

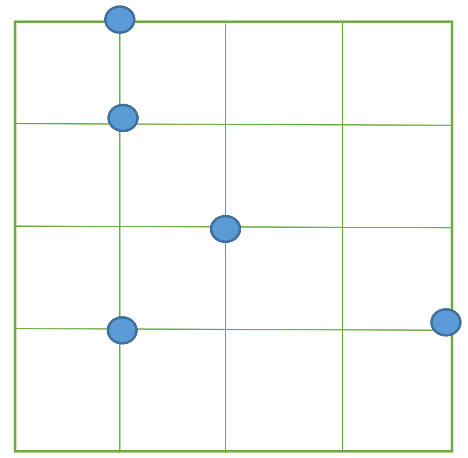
This Lecture

- Geometric Problems in the Stream

Geometric MST

Model

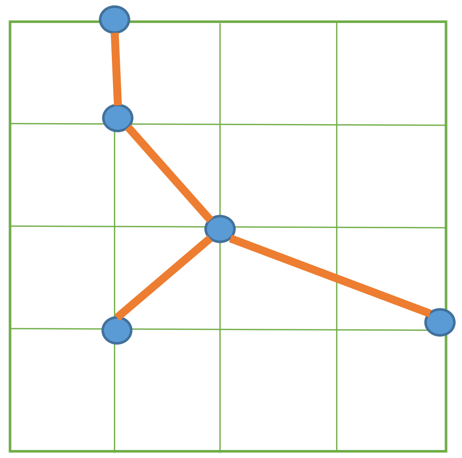
- Input: Points in $\{1, \ldots, \Delta\}^d$ are coming in a stream



Model

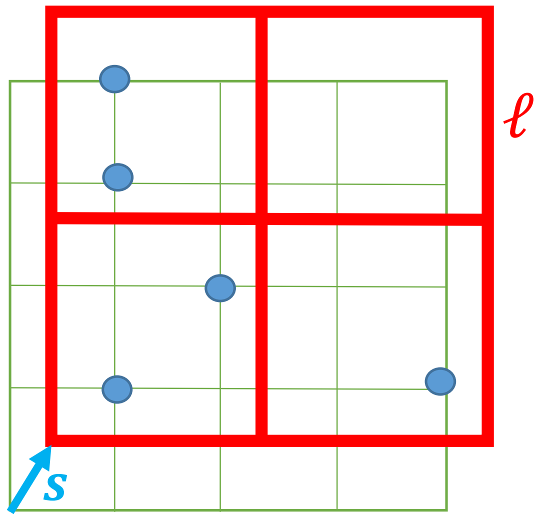
- Input: Points in $\{\mathbf{1}, \ldots, \boldsymbol{\Delta}\}^{\boldsymbol{d}}$ are coming in a stream
- Goal: Estimate the cost of the MST using small space

- Dynamic setting (the points might get deleted too)
- Approximation factor $O(d \cdot \log \Delta)$



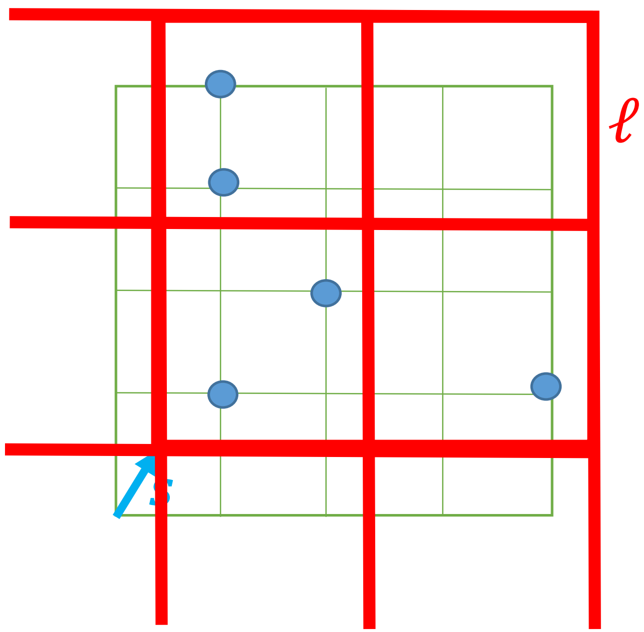
Randomly Shifted Grid

- Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)
  - Let the cell side be of length ℓ



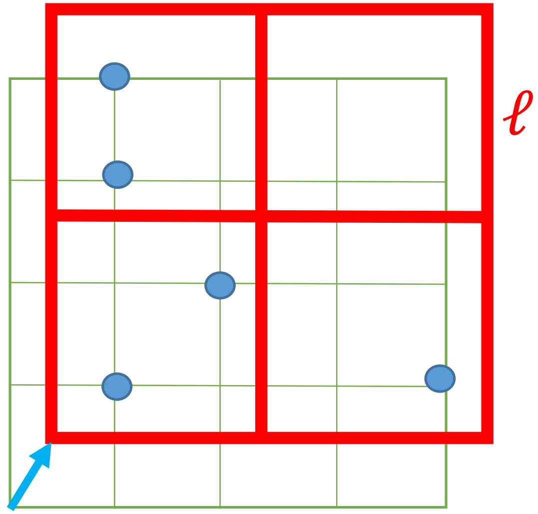
Randomly Shifted Grid

- ❑ Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)
  - Let the cell side be of length ℓ



Randomly Shifted Grid

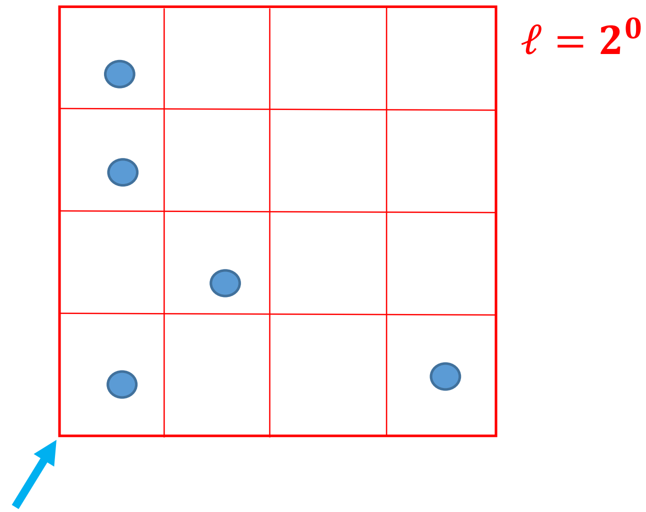
- Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)
  - Let the cell side be of length ℓ
  - Consider two points a and b
  - They are separated w.p $\|a-b\|_\infty/\ell \le p \le \|a-b\|_1/\ell$
  - E.g., if $\|a-b\|_\infty \ge \ell$ then they will be separated anyways
  - $\|a-b\|_\infty \le \|a-b\|_2 \le \|a-b\|_1 \le \sqrt{d}\|a-b\|_2 \le d\|a-b\|_\infty$



Randomly Shifted Grid

- ❑ Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)
  - Let the cell side be of length ℓ
  - Consider two points a and b
  - They are separated w.p $\|a - b\|_\infty / \ell \leq p \leq \|a - b\|_1 / \ell$

- Consider nested grids of side lengths $\ell_i \in \{2^i | 0 \leq i \leq \log \Delta\}$ (all shifted using the same s)

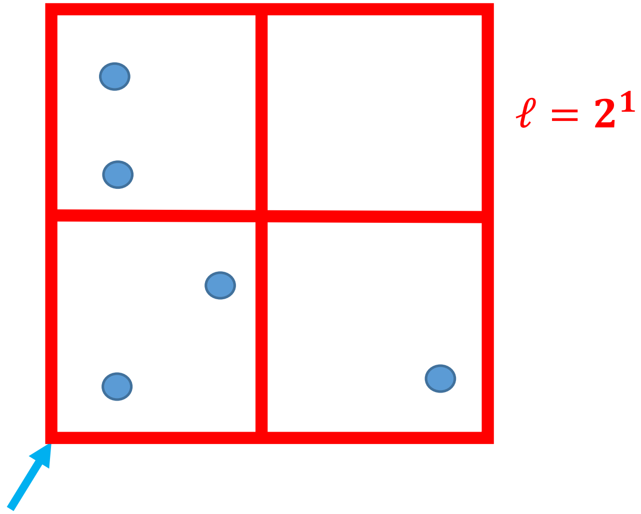


Randomly Shifted Grid

❑ Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)

- Let the cell side be of length ℓ
- Consider two points a and b
- They are separated w.p $\|a - b\|_\infty / \ell \leq p \leq \|a - b\|_1 / \ell$

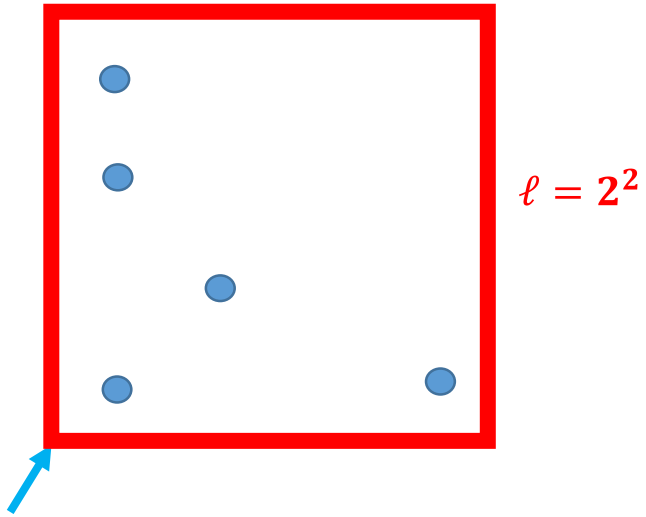
- Consider nested grids of side lengths $\ell_i \in \{2^i | 0 \leq i \leq \log \Delta\}$ (all shifted using the same s)



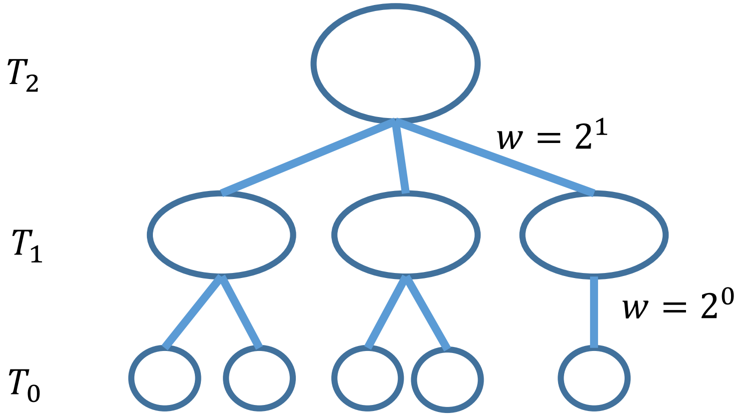
Randomly Shifted Grid

- ❑ Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)
  - Let the cell side be of length ℓ
  - Consider two points a and b
  - They are separated w.p $\|a - b\|_\infty / \ell \leq p \leq \|a - b\|_1 / \ell$

- Consider nested grids of side lengths $\ell_i \in \{2^i | 0 \leq i \leq \log \Delta\}$ (all shifted using the same s)

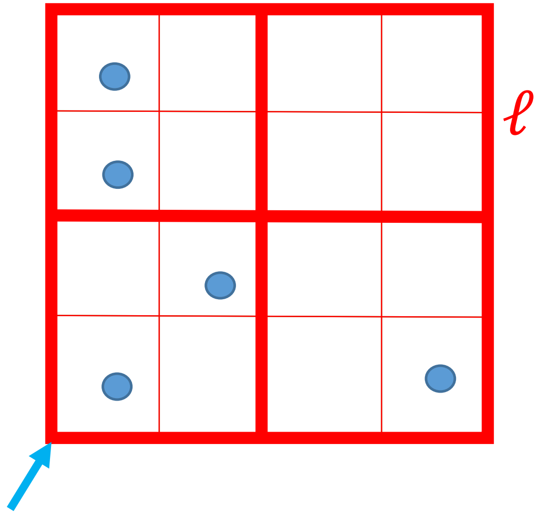


Randomly Shifted Grid

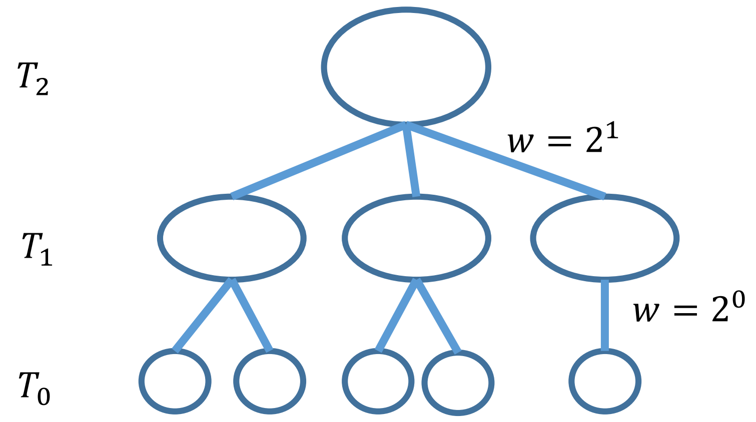


- Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)
  - Let the cell side be of length ℓ
  - Consider two points a and b
  - They are separated w.p $\|a - b\|_\infty / \ell \leq p \leq \|a - b\|_1 / \ell$

- Consider nested grids of side lengths $\ell_i \in \{2^i | 0 \leq i \leq \log \Delta\}$ (all shifted using the same s)
  - Build a tree for them
    - The nodes corresponds to the cells
    - The children of each cell are the cells it contains.
    - Call the nodes at height i (corresponding to non-empty cells in the i-th grid) as $T_i \subseteq G_i$
    - The weight of the edges from T_i to their parents is 2^i for $0 \leq i < \log \Delta$



Randomly Shifted Grid



- ❑ Impose a randomly shifted grid (shifted by a vector $s \in [0, \Delta]^d$)
  - Let the cell side be of length ℓ
  - Consider two points a and b
  - They are separated w.p $\|a - b\|_\infty / \ell \leq p \leq \|a - b\|_1 / \ell$

- Consider nested grids of side lengths $\ell_i \in \{2^i | 0 \leq i \leq \log \Delta\}$ (all shifted using the same s)
  - Build a tree for them
    - The nodes corresponds to the cells
    - The children of each cell are the cells it contains.
    - Call the nodes at height i (corresponding to non-empty cells in the i-th grid) as $T_i \subseteq G_i$
    - The weight of the edges from T_i to their parents is 2^i for $0 \leq i < \log \Delta$
  - This is a 2-HST (Hierarchically Well-separated Tree)
    - Distance from each node to all children are equal
    - Weights on each path down the tree decreases by a factor of at least 2, in each level

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right) \approx 2^{i+2}$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)

- They are separated w.p $\|a-b\|_\infty / \ell \leq p \leq \|a-b\|_1 / \ell$
- $\|a-b\|_\infty \leq \|a-b\|_2 \leq \|a-b\|_1 \leq \sqrt{d}\|a-b\|_2 \leq d\|a-b\|_\infty$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d\log\Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d\log\Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right)\approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2\geq\sqrt{d}\cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i}\geq\frac{\|a-b\|_2}{\sqrt{d}\cdot 2^i}\geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2\leq\frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i}\leq\frac{\sqrt{d}\cdot\|a-b\|_2}{2^i}$

- They are separated w.p $\|a-b\|_\infty/\ell\leq p\leq\|a-b\|_1/\ell$
- $\|a-b\|_\infty\leq\|a-b\|_2\leq\|a-b\|_1\leq\sqrt{d}\|a-b\|_2\leq d\|a-b\|_\infty$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2 \leq \frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i} \leq \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i}$
- Otherwise, we have $\frac{2^i}{\sqrt{d}} \leq \|a-b\|_2 \leq \sqrt{d} \cdot 2^i$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2(2^{i+1}-1) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2 \leq \frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i} \leq \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i}$
- Otherwise, we have $\frac{2^i}{\sqrt{d}} \leq \|a-b\|_2 \leq \sqrt{d} \cdot 2^i$
- $\frac{\|a-b\|_2}{\sqrt{d}} \leq d_T(a,b)$ (consider the grid G_i where $2^i \leq \|a-b\|_2/\sqrt{d}$)

Randomly Shifted Grid

- ❑ This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2 \leq \frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i} \leq \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i}$
- Otherwise, we have $\frac{2^i}{\sqrt{d}} \leq \|a-b\|_2 \leq \sqrt{d} \cdot 2^i$
- $\frac{\|a-b\|_2}{\sqrt{d}} \leq d_T(a,b)$
- $\mathbb{E}[d_T(a,b)] \leq \sum_{i=\log \Delta}^{0} \Pr[a, b \text{ are cut by } G\_i | \text{they are not cut by } G_{j>i}] \cdot 2^{i+2} \leq \sqrt{d} \cdot \|a-b\|_2 +$

  $\sum_{i=\log \Delta}^{\log \sqrt{d} \cdot \|a-b\|_2} \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i} \cdot 2^{i+2} \leq \log \Delta \cdot \sqrt{d} \cdot \|a-b\|_2$

- $2^i \leq \sqrt{d} \cdot \|a-b\|_2$
- $2^i \geq \sqrt{d} \cdot \|a-b\|_2$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2(2^{i+1}-1) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2 \leq \frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i} \leq \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i}$
- Otherwise, we have $\frac{2^i}{\sqrt{d}} \leq \|a-b\|_2 \leq \sqrt{d} \cdot 2^i$
- $\frac{\|a-b\|_2}{\sqrt{d}} \leq d_T(a,b)$
- $\mathbb{E}[d_T(a,b)] \leq \sum_{i=\log\Delta}^{0} \Pr[a, b \textit{ are cut by } G\_i | \textit{they are not cut by } G_{j>i}] \cdot 2^{i+2} \leq \sqrt{d} \cdot \|a-b\|_2 +$

  $\sum_{i=\log\Delta}^{\log \sqrt{d} \cdot \|a-b\|_2} \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i} \cdot 2^{i+2} \leq \log \Delta \cdot \sqrt{d} \cdot \|a-b\|_2$

- $2^i \leq \sqrt{d} \cdot \|a-b\|_2$
- $2^i \geq \sqrt{d} \cdot \|a-b\|_2$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2 \leq \frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i} \leq \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i}$
- Otherwise, we have $\frac{2^i}{\sqrt{d}} \leq \|a-b\|_2 \leq \sqrt{d} \cdot 2^i$
- $\frac{\|a-b\|_2}{\sqrt{d}} \leq d_T(a,b)$
- $\mathbb{E}[d_T(a,b)] \leq \sum_{i=\log\Delta}^{0} \Pr[a, b\ are\ cut\ by\ G\_i | they\ are\ not\ cut\ by\ G_{j>i}] \cdot 2^{i+2} \leq \sqrt{d} \cdot \|a-b\|_2 +$

$\sum_{i=\log\Delta}^{\log \sqrt{d} \cdot \|a-b\|_2} \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i} \cdot 2^{i+2}$

- $2^i \leq \sqrt{d} \cdot \|a-b\|_2$
- $2^i \geq \sqrt{d} \cdot \|a-b\|_2$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
    - The distance between any two points a and b never decreases,
    - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2 \leq \frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i} \leq \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i}$
- Otherwise, we have $\frac{2^i}{\sqrt{d}} \leq \|a-b\|_2 \leq \sqrt{d} \cdot 2^i$
- $\frac{\|a-b\|_2}{\sqrt{d}} \leq d_T(a,b)$
- $\mathbb{E}[d_T(a,b)] \leq \sum_{i=\log\Delta}^{0} \Pr[a, b\ are\ cut\ by\ G\_i | they\ are\ not\ cut\ by\ G_{j>i}] \cdot 2^{i+2} \leq \sqrt{d} \cdot \|a-b\|_2 +$

  $\sum_{i=\log\Delta}^{\log \sqrt{d} \cdot \|a-b\|_2} \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i} \cdot 2^{i+2} \leq \log\Delta \cdot \sqrt{d} \cdot \|a-b\|_2$

Randomly Shifted Grid

- This is a probabilistic embedding of the metric into a collection of trees with distortion of $O(d \log \Delta)$.
  - The distance between any two points a and b never decreases,
  - In expectation, each distance does not increase by more than a factor of $O(d \log \Delta)$.
- If a and b are cut by G_i, i.e., their least common ancestor is at level $i+1$, then their distance on the tree is $2\left(2^{i+1}-1\right) \approx 2^{i+2}$
- If two points are at distance $\|a-b\|_2 \geq \sqrt{d} \cdot 2^i$, then the probability that they are cut by G_i is at least $\frac{\|a-b\|_\infty}{2^i} \geq \frac{\|a-b\|_2}{\sqrt{d} \cdot 2^i} \geq 1$ (so they will be cut)
- Otherwise if $\|a-b\|_2 \leq \frac{2^i}{\sqrt{d}}$, then the probability that they are cut by G_i is at most $\frac{\|a-b\|_1}{2^i} \leq \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i}$
- Otherwise, we have $\frac{2^i}{\sqrt{d}} \leq \|a-b\|_2 \leq \sqrt{d} \cdot 2^i$
- $\frac{\|a-b\|_2}{\sqrt{d}} \leq d_T(a,b)$
- $\mathbb{E}[d_T(a,b)] \leq \sum_{i=\log\Delta}^{0} \Pr[a, b\ are\ cut\ by\ G\_i | they\ are\ not\ cut\ by\ G_{j>i}] \cdot 2^{i+2} \leq \sqrt{d} \cdot \|a-b\|_2 +$

  $\sum_{i=\log\Delta}^{\log \sqrt{d} \cdot \|a-b\|_2} \frac{\sqrt{d} \cdot \|a-b\|_2}{2^i} \cdot 2^{i+2} \leq \log\Delta \cdot \sqrt{d} \cdot \|a-b\|_2$
- In general, this is a very useful technique! (we now need to solve the problem over a tree).

Cost of MST

- Goal: find MST on the tree
- Exercise: Cost of MST on the point set $\leq O(d \cdot \log \Delta) \cdot$ Cost of MST on the HST
- Exercise: Cost of MST can be approximated by $n_i \cdot 2^i$ where n_i is the number of non-empty cells in G_i

- Algorithms: estimate n_i for each i throughout the stream
- Equivalent to the #distinct elements in the stream!

- Total space: $\tilde{O}(1)$

Cost of Minimum Weight Matching

- Goal: estimate the cost of minimum weight matching on the tree
- Intuition: match the points as much as possible inside the cells.
- Exercise: cost of MST can be approximated by $n + \sum_i m_i 2^i$ where m_i is the number of cells in G_i with an odd number of points in them.

- Algorithms: estimate m_i for each i throughout the stream
- Solve the decision version: for a threshold T, decide whether $m_i \leq T/10$ or $m_i \geq T$
- Sample the cells w.p. $1/T$ and keep a single bit which is the sum of #point in the sampled cells of G_i mod 2.
- Exercise: the above algorithm has a constant prob of success.

Approximating SVM cost

Streaming Algorithms for SVM

Input: a stream of n labeled data points $P = \{(p_i, y_i)\}$ where $p_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$

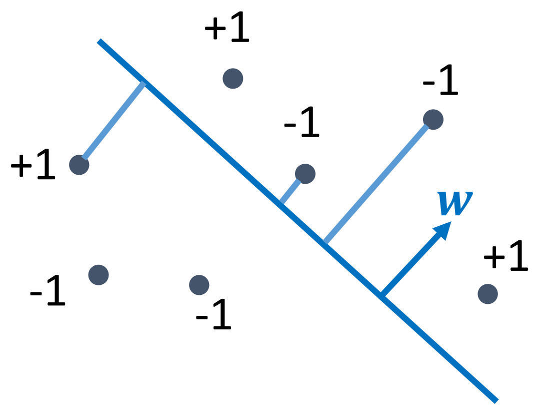
Streaming Algorithms for SVM

Input: a stream of n labeled data points $P = \{(p_i, y_i)\}$ where $p_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$

Goal: Find a sketch so that the SVM cost of any hyper-plane can be computed on the fly, i.e.,

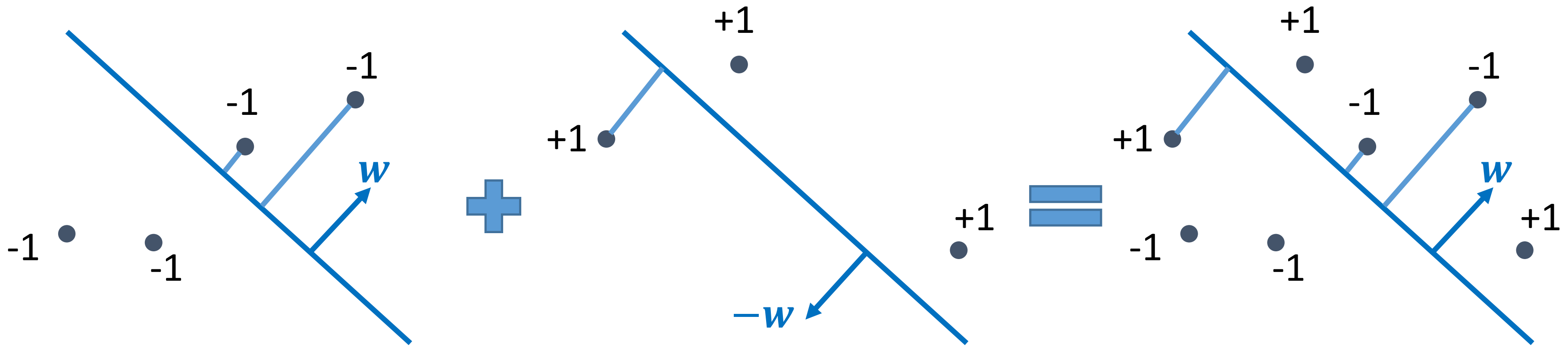
- **Given:** $w \in \mathbb{R}^d, b \in \mathbb{R}$, compute

$$\sum_{p_i \in P} \max(0, -y_i[\langle p_i, w\rangle - b])$$



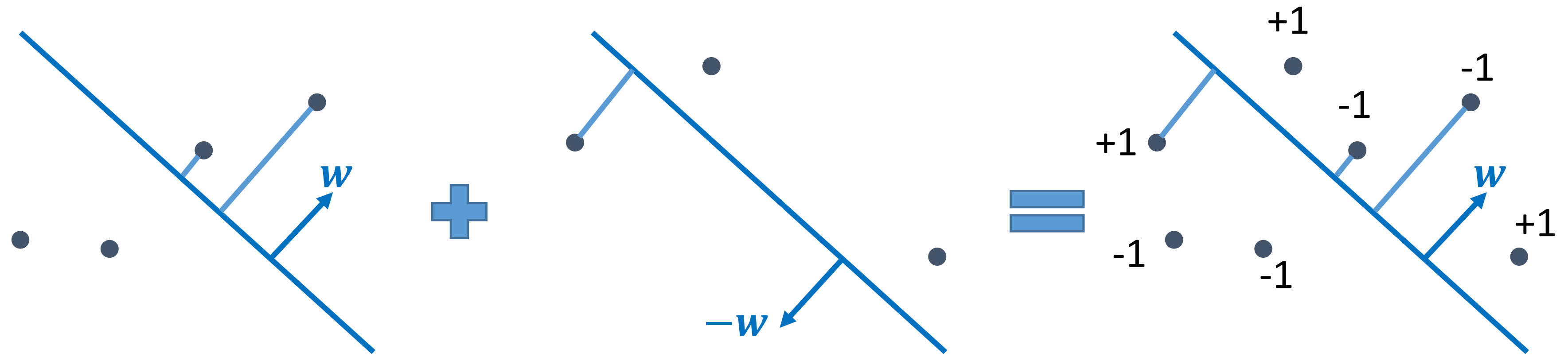
Remove the labels

- A Data structure for each of +1 and -1 labels separately



Remove the labels

- A Data structure for each of +1 and -1 labels separately

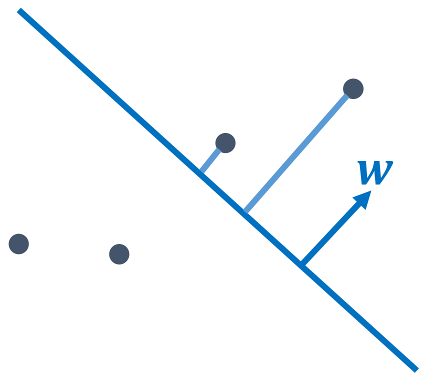


Remove the labels

- A Data structure for each of +1 and -1 labels separately

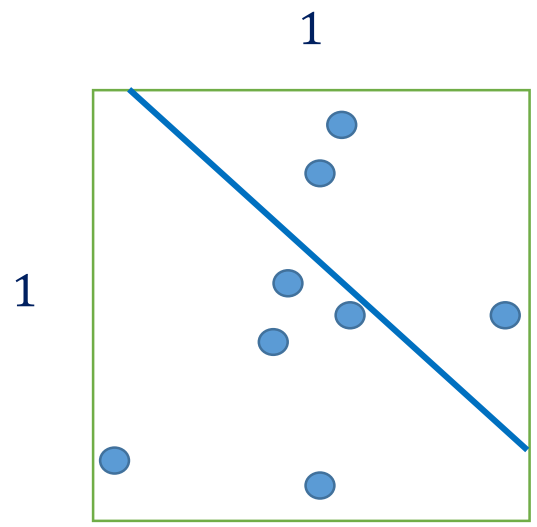
New Problem: Given a stream of points $p_i \in \mathbb{R}^d$, process them such that given a query hyperplane denoted by w, b , computes

$$\sum_{p \in P} \max\{\langle w, p \rangle - b, 0\}$$



Algorithm for 2-dimensions: Naïve algorithm

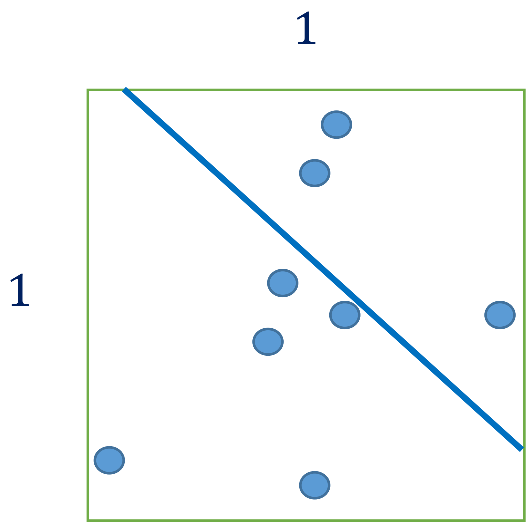
Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ



Algorithm for 2-dimensions: Naïve algorithm

Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ

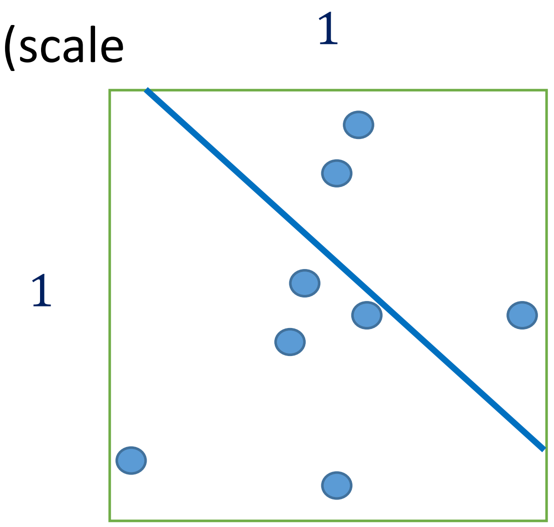
- Approximate $\sum_i Z_i$ where $0 \leq Z_i \leq \sqrt{2}$
- Sample one of Z_i and report $Z' = n \cdot Z_i$
  - $\mathbb{E}[Z'] = \sum_i \left(\frac{1}{n}\right) \cdot (n \cdot Z_i) = \sum_i Z_i$
  - $Var(Z') \leq \mathbb{E}[Z'^2] = \sum_i \left(\frac{1}{n}\right) \cdot (n \cdot Z_i)^2 \leq 2n^2$
- Sample t of them and report the average times n,
  - Unbiased estimator
  - Variance $\left(\frac{1}{t}\right) \cdot 2n^2$
  - $\Pr[|v' - v| \geq \epsilon n] \leq \frac{\frac{1}{t} \cdot n^2 \cdot 2}{(\epsilon n)^2} = \frac{2}{\epsilon^2 t}$
- We need to sample $\Omega(1/\epsilon^2)$
- The error is additive $\epsilon \Delta n$



Algorithm for 2-dimensions: Naïve algorithm

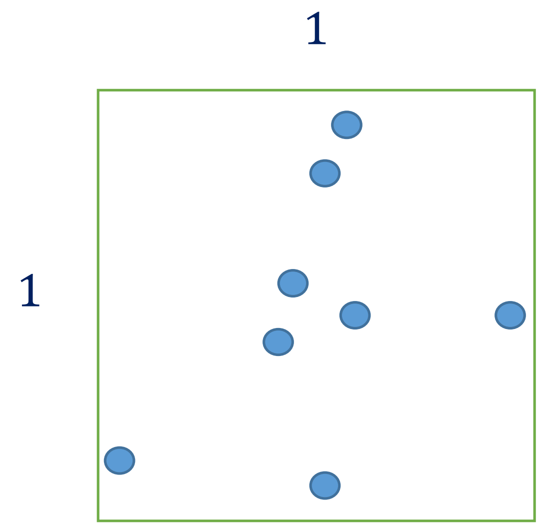
Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ

- Sketch: Sample and keep $\Omega(1/\epsilon^2)$ points
- Algorithm: Estimate the cost using sampled points (scale them accordingly by n/t)



Can we improve the space complexity over ϵ^{-2}?

Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ



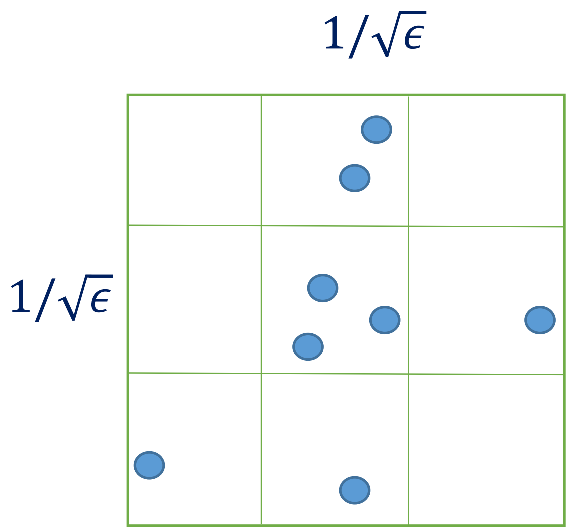
Algorithm for 2-dimensions

Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ

Natural Idea:

- Partition to a grid of side length $\sqrt{\epsilon}$
- Keep the mean in each grid and the number of points

Memory is $O(1/\epsilon)$



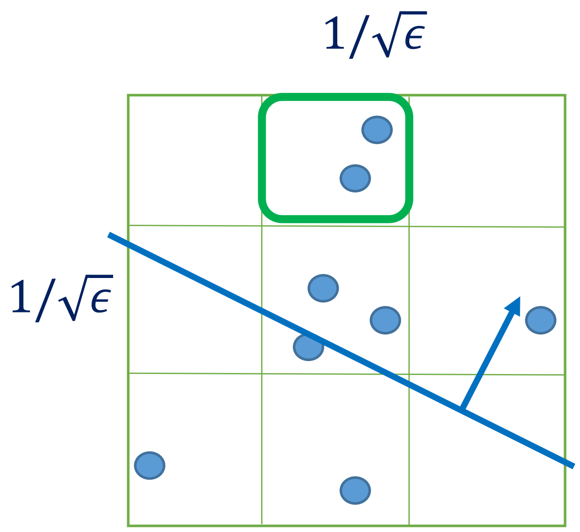
Algorithm for 2-dimensions

Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ

Natural Idea:

- Partition to a grid of side length $\sqrt{\epsilon}$
- Keep the mean in each grid and the number of points
- For cells far from the line compute the distance exactly
  - $\langle w, p_1 \rangle - b + \langle w, p_2 \rangle - b = \langle w, p_1 + p_2 \rangle - 2b$

Memory is $O(1/\epsilon)$



Algorithm for 2-dimensions

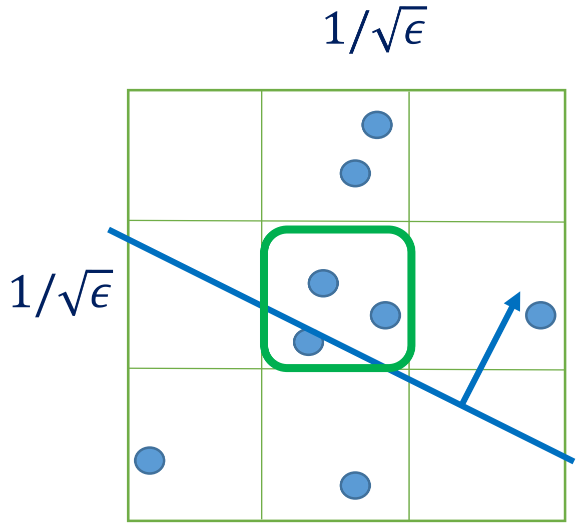
Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ

Natural Idea:

- Partition to a grid of side length $\sqrt{\epsilon}$
- Keep the mean in each grid and the number of points
- For cells far from the line compute the distance exactly
  - $\langle w, p_1 \rangle - b + \langle w, p_2 \rangle - b = \langle w, p_1 + p_2 \rangle - 2b$
- For intersecting cells, ignore

Memory is $O(1/\epsilon)$

Error is $O(n\sqrt{\epsilon})$

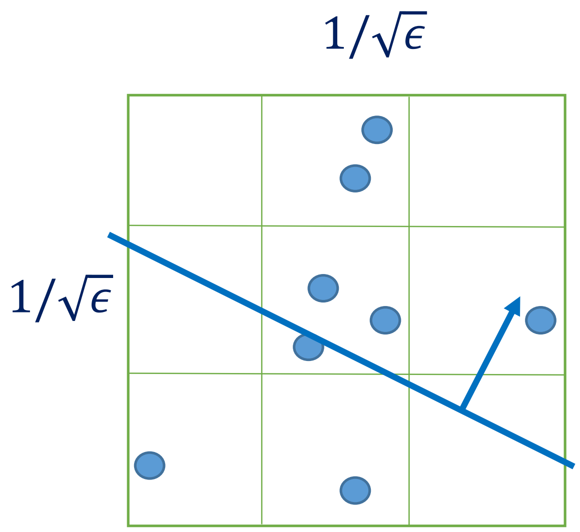


Algorithm for 2-dimensions

Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ

Natural Idea:

- Partition to a grid of side length $\sqrt{\epsilon}$
- Keep the mean in each grid and the number of points
- For cells far from the line compute the distance exactly
  - $\langle w, p_1 \rangle - b + \langle w, p_2 \rangle - b = \langle w, p_1 + p_2 \rangle - 2b$
- For intersecting cells, ignore



Memory is $O(1/\epsilon)$

Error is $O(n\sqrt{\epsilon})$

➢ No improvement yet

Algorithm for 2-dimensions

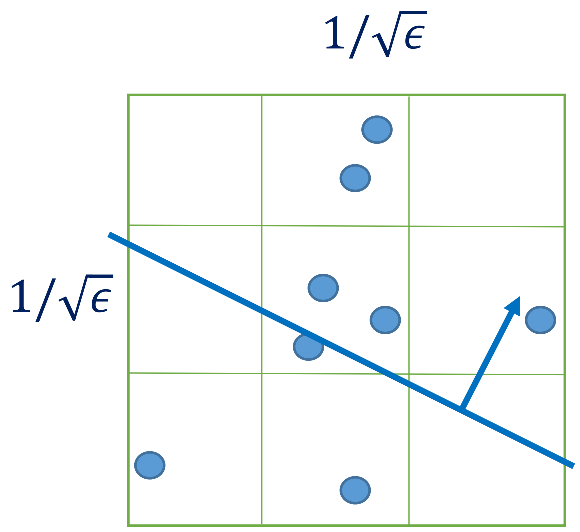
Given a set of n points on the $[0,1] \times [0,1]$ plane, process them to compute the cost for any line query ℓ

Natural Idea:

- Partition to a grid of side length $\sqrt{\epsilon}$
- Keep the mean in each grid and the number of points
- For cells far from the line compute the distance exactly
  - $\langle w, p_1 \rangle - b + \langle w, p_2 \rangle - b = \langle w, p_1 + p_2 \rangle - 2b$
- For intersecting cells, ignore

Memory is $O(1/\epsilon)$

Error is $O(n\sqrt{\epsilon})$

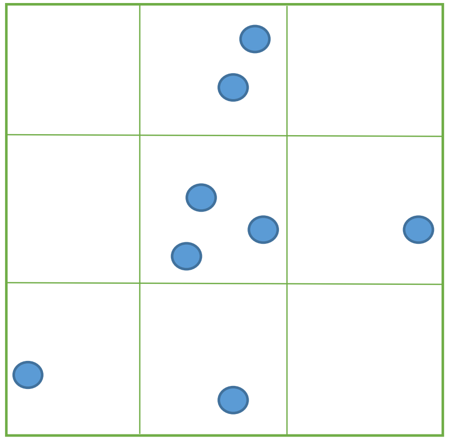


➢ If a cell has too many points partition further.

Quad Tree Approach

Data Structure

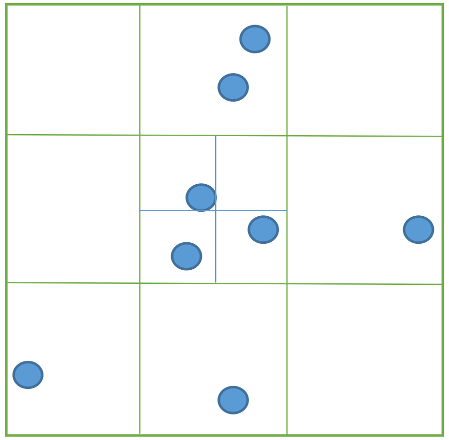
- Consider a Quad tree on the set of points of starting with a $\frac{1}{\sqrt{\epsilon}}$ by $\frac{1}{\sqrt{\epsilon}}$ grid



Quad Tree Approach

Data Structure

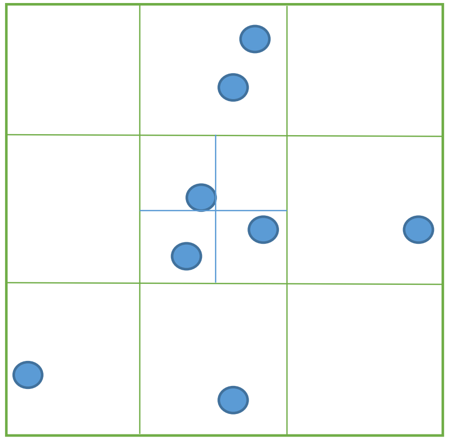
- Consider a Quad tree on the set of points of starting with a $\frac{1}{\sqrt{\epsilon}}$ by $\frac{1}{\sqrt{\epsilon}}$ grid
- If there are many $(\geq n\epsilon)$ points in a cell partition further, until for each cell either
  - Side length is at most $\boldsymbol{\epsilon}$
  - Number of points is at most $\boldsymbol{n\epsilon}$



Quad Tree Approach

Data Structure

- Consider a Quad tree on the set of points of starting with a $\frac{1}{\sqrt{\epsilon}}$ by $\frac{1}{\sqrt{\epsilon}}$ grid
- If there are many ($\geq n\epsilon$) points in a cell partition further, until for each cell either
  - Side length is at most $\boldsymbol{\epsilon}$
  - Number of points is at most $\boldsymbol{n\epsilon}$



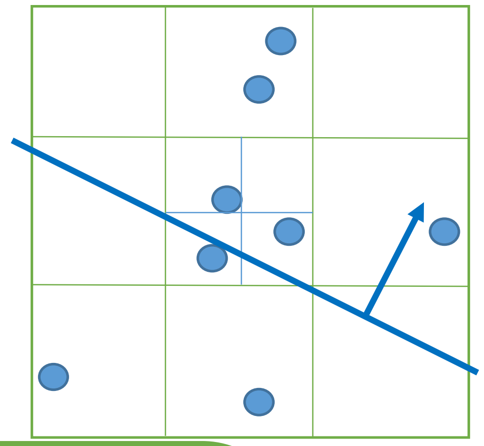
Memory:

- height of the tree is $\log 1/\epsilon$
- Number of Cells is $O\left(\frac{1}{\epsilon}\log 1/\epsilon\right) \approx \epsilon^{-1}$

Quad Tree Approach

Data Structure

- Consider a Quad tree on the set of points of starting with a $\frac{1}{\sqrt{\epsilon}}$ by $\frac{1}{\sqrt{\epsilon}}$ grid
- If there are many ($\geq n\epsilon$) points in a cell partition further, until for each cell either
  - Side length is at most $\boldsymbol{\epsilon}$
  - Number of points is at most $\boldsymbol{n\epsilon}$



Error (caused by intersecting cells)

- Cells with large number of points
  - Side Length is $\leq \epsilon$
  - Total Error is $\boldsymbol{n\epsilon}$
- Other cells:
  - Side length ℓ, total error at most $\boldsymbol{\ell \cdot n\epsilon \cdot (1/\ell) = n\epsilon}$
  - Sum over all ℓ, the error is $O(n\epsilon \log(1/\epsilon))$

So far

- First Approach (keep the number of points and the mean for each cell of a grid)

Memory is $O(1/\epsilon)$

Error is $O(n\sqrt{\epsilon})$

- Second Approach (keep the number of points and the mean for each cell of a quad tree)

Memory is $O(1/\epsilon)$

Error is $O(n\epsilon)$

- Third Approach

Memory is $O(1/\epsilon)$

Error is $O(n\epsilon^{5/4})$

To improve from ϵ^{-1} to $\epsilon^{-4/5}$

To improve from ϵ^{-1} to $\epsilon^{-4/5}$

- For each cell, also keep a **random point** from the cell, in case of intersection with the line.
  - Don't ignore the intersecting cells, instead use the random point to estimate the cost

- Why does it help?
  - In expectation we get the correct value for all (including intersecting) cells.

- By bounding the variance, we can show the improvement.
  - Over multiple cells, the over estimation and under estimations cancel out.

To improve from ϵ^{-1} to $\epsilon^{-4/5}$

- For each cell, also keep a **random point** from the cell, in case of intersection with the line.

- Take a cell c with side length ℓ that intersects with the line, and let n_c be the number of points in the cell. What is the variance in the cell?
- $Var[n_c \cdot \max\{0, D(r_c, L)\}] \leq \sum_{i=1}^{n_c} \left(\frac{1}{n_c}\right)(n_c \cdot \ell)^2 \leq (n_c \ell)^2 \leq (n\epsilon)^2 \ell^2$
- The variance over all cells with side length ℓ (there are $1/\ell$ of them) is at most $(n\epsilon)^2 \ell$ (since the samples are chosen independently).

To improve from ϵ^{-1} to $\epsilon^{-4/5}$

- For each cell, also keep a **random point** from the cell, in case of intersection with the line.

- Take a cell c with side length ℓ that intersects with the line, and let n_c be the number of points in the cell. What is the variance in the cell?
- $Var[n_c \cdot \max\{0, D(r_c, L)\}] \leq \sum_{i=1}^{n_c} \left(\frac{1}{n_c}\right)(n_c \cdot \ell)^2 \leq (n_c\ell)^2 \leq (n\epsilon)^2\ell^2$
- The variance over all cells with side length ℓ (there are $1/\ell$ of them) is at most $(n\epsilon)^2\ell$ (since the samples are chosen independently).
- Sum over different side lengths $\epsilon \leq \ell \leq \sqrt{\epsilon}$, the total variance is $\tilde{O}((n\epsilon)^2\sqrt{\epsilon})$

To improve from ϵ^{-1} to $\epsilon^{-4/5}$

- For each cell, also keep a **random point** from the cell, in case of intersection with the line.

- Take a cell c with side length ℓ that intersects with the line, and let n_c be the number of points in the cell. What is the variance in the cell?
- $Var[n_c \cdot \max\{0, D(r_c, L)\}] \leq \sum_{i=1}^{n_c} \left(\frac{1}{n_c}\right)(n_c \cdot \ell)^2 \leq (n_c \ell)^2 \leq (n\epsilon)^2 \ell^2$
- The variance over all cells with side length ℓ (there are $1/\ell$ of them) is at most $(n\epsilon)^2 \ell$ (since the samples are chosen independently).
- Sum over different side lengths $\epsilon \leq \ell \leq \sqrt{\epsilon}$, the total variance is $\tilde{O}((n\epsilon)^2\sqrt{\epsilon})$
- The standard Deviation is $\tilde{O}(n\epsilon^{\frac{5}{4}})$ which gives a better Chebyshev's inequality.

How to make it work in the streaming model

Challenge: The quad tree partitioning depends on all the data.

Solution: whenever a cell becomes too heavy, partition it further, but only the upcoming points will be assigned to children.