Probabilistic analysis of vantage point trees

Vladyslav Bohun
vladyslavbogun@gmail.com

Taras Shevchenko National University of Kyiv

Vantage point tree

This talk is based on a recent study by Bohun (2021+).

Vantage point tree

This talk is based on a recent study by Bohun (2021+).

A vantage point tree (VP-tree) is a data structure for fast executing of nearest neighbor search queries in a given metric space first introduced by Yianilos (1993).

Vantage point tree

This talk is based on a recent study by Bohun (2021+).

A vantage point tree (VP-tree) is a data structure for fast executing of nearest neighbor search queries in a given metric space first introduced by Yianilos (1993).

Consider a metric space $([-1,1]^d, \mathsf{dist})$, $d > 1$, with the ℓ_∞-distance function

$$\mathsf{dist}(x,y) = \max_{1 \le j \le d} |x_j - y_j|.$$

Vantage point tree

This talk is based on a recent study by Bohun (2021+).

A vantage point tree (VP-tree) is a data structure for fast executing of nearest neighbor search queries in a given metric space first introduced by Yianilos (1993).

Consider a metric space $([-1,1]^d, \mathrm{dist})$, $d > 1$, with the ℓ_∞-distance function

$$\mathrm{dist}(x,y) = \max_{1 \le j \le d} |x_j - y_j|.$$

We build a VP-tree on a sequence $(x_i)_{i \ge 1}$ of independent uniformly distributed points in $[-1,1]^d$ by consistently inserting points into the tree.

Insert procedure

Fix a parameter $\tau \in (0, 1)$. Inserting of a new point into a VP-tree is executed in the following recursive manner.

Insert procedure

Fix a parameter $\tau \in (0, 1)$. Inserting of a new point into a VP-tree is executed in the following recursive manner.

- At the beginning the tree is always empty and the first point x_1 is always stored in its root.

Insert procedure

Fix a parameter $\tau \in (0,1)$. Inserting of a new point into a VP-tree is executed in the following recursive manner.

- At the beginning the tree is always empty and the first point x_1 is always stored in its root.
- Consider adding n-th element x_n. We start operation from node T which the root of the tree.

Insert procedure

Fix a parameter $\tau \in (0,1)$. Inserting of a new point into a VP-tree is executed in the following recursive manner.

- At the beginning the tree is always empty and the first point x_1 is always stored in its root.
- Consider adding n-th element x_n. We start operation from node T which the root of the tree.
- Let the current node T be non-empty and x_T be a point stored in it, we also put r is the distance from T to the root. If $\text{dist}(x_T, x_n) \leq \tau^{r+1}$ then we proceed operation from T-th left son, otherwise we proceed into the right son.

Insert procedure

Fix a parameter $\tau \in (0, 1)$. Inserting of a new point into a VP-tree is executed in the following recursive manner.

- At the beginning the tree is always empty and the first point x_1 is always stored in its root.
- Consider adding n-th element x_n. We start operation from node T which the root of the tree.
- Let the current node T be non-empty and x_T be a point stored in it, we also put r is the distance from T to the root. If $\text{dist}(x_T, x_n) \leq \tau^{r+1}$ then we proceed operation from T-th left son, otherwise we proceed into the right son.
- If the current node T is empty then the value x_n is stored in it.

Partitions over space

Consider the *leftmost path* of a VP-tree, that is, a path which starts in the root and always goes into the left son of the current node till reaching a leaf.

Partitions over space

Consider the *leftmost path* of a VP-tree, that is, a path which starts in the root and always goes into the left son of the current node till reaching a leaf.

The object in focus is the evolution of sets $(I_h)_{h \geq 1}$, where I_h consists of all points of $[-1, 1]^d$ which, if added to the tree, are attached to the left subtree of h-th node in the leftmost path.

Partitions over space

Consider the *leftmost path* of a VP-tree, that is, a path which starts in the root and always goes into the left son of the current node till reaching a leaf.

The object in focus is the evolution of sets $(I_h)_{h\geq 1}$, where I_h consists of all points of $[-1,1]^d$ which, if added to the tree, are attached to the left subtree of h-th node in the leftmost path.

$Y_h = I_h/\tau^h$ is a normalized sequence. It satisfies

$$Y_h = \frac{Y_{h-1} - u_h}{\tau} \bigcap [-1,1]^d,$$

where u_h is uniformly distributed on Y_{h-1} and $Y_0 = [-1,1]^d$.

Plot of a VP-tree

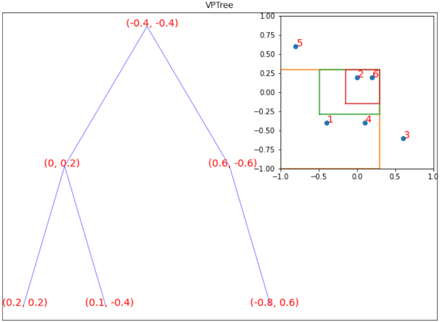

Figure: Plot of a VP-tree obtained after inserting 6 points with $\tau = 0.7$, as well as partitions I_h 1st, 2nd and 6th create as part of the leftmost path.

Convergence of normalized partitions

Every Y_h is a d-dimensional rectangle with edges being parallel to coordinate axes.

Convergence of normalized partitions

Every Y_h is a d-dimensional rectangle with edges being parallel to coordinate axes.

Let $X_{j,h}$ be the length of the projection of Y_h onto j-th coordinate, $1 \leq j \leq d$. Values over different dimensions are independent identically distributed random variables on $[1, 2]$.

Convergence of normalized partitions

Every Y_h is a d-dimensional rectangle with edges being parallel to coordinate axes.

Let $X_{j,h}$ be the length of the projection of Y_h onto j-th coordinate, $1 \le j \le d$. Values over different dimensions are independent identically distributed random variables on $[1, 2]$.

Theorem

As $h \to \infty$, $X_{1,h} \xrightarrow{d} X_\infty$, where X_∞ is the unique solution of the stochastic fixed-point equation

$$X_\infty \stackrel{d}{=} \min\left\{\tau^{-1} U X_\infty, 1\right\} + \min\left\{\tau^{-1}(1-U)X_\infty, 1\right\},$$

where U is uniformly distributed on $[0, 1]$ and independent of X_∞.

Example 1

The value X_∞ can be found explicitly from the corresponding stochastic fixed-point equation, but the solution depends heavily on τ.

Example 1

The value X_∞ can be found explicitly from the corresponding stochastic fixed-point equation, but the solution depends heavily on τ.

Let $\tau \leq 1/2$ then $X_\infty = p(1+U) + (1-p)\delta_2$, where δ_2 is a Dirac measure at 2, U is uniformly distributed on $[0,1]$ and $p = \frac{1}{1 - 2\log 2 + \frac{1}{\tau}}$.

Example 1

The value X_∞ can be found explicitly from the corresponding stochastic fixed-point equation, but the solution depends heavily on τ.

Let $\tau \leq 1/2$ then $X_\infty = p(1+U) + (1-p)\delta_2$, where δ_2 is a Dirac measure at 2, U is uniformly distributed on $[0,1]$ and $p = \frac{1}{1-2\log 2 + \frac{1}{\tau}}$.

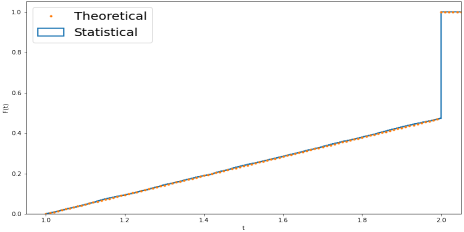

Figure: Theoretical and statistically computed values of CDF of X_∞ with $\tau = 0.4$.

Example 2

Consider a more complex example. Let $1/2 < \tau \leq 1/\sqrt{2}$.

Example 2

Consider a more complex example. Let $1/2 < \tau \leq 1/\sqrt{2}$.

For $1 \leq t \leq 1/\tau$

$$\mathbb{P}\{X_\infty < t\} = 2\tau p(t-1).$$

Example 2

Consider a more complex example. Let $1/2 < \tau \leq 1/\sqrt{2}$.

For $1 \leq t \leq 1/\tau$

$$\mathbb{P}\{X_\infty < t\} = 2\tau p(t-1).$$

For $1/\tau < t \leq 2$

$$\mathbb{P}\{X_\infty < t\} = 2\tau p(t + \tau t - 2\tau t \log(\tau t) + 2\tau \log(\tau t) - 2).$$

Example 2

Consider a more complex example. Let $1/2 < \tau \le 1/\sqrt{2}$.

For $1 \le t \le 1/\tau$

$$\mathbb{P}\{X_\infty < t\} = 2\tau p(t-1).$$

For $1/\tau < t \le 2$

$$\mathbb{P}\{X_\infty < t\} = 2\tau p(t + \tau t - 2\tau t \log(\tau t) + 2\tau \log(\tau t) - 2).$$

There is also a special value $X_\infty = 2$:

$$\mathbb{P}\{X_\infty = 2\} = 1 - 4\tau^2 p(1 - \log(2\tau)).$$

Example 2

Consider a more complex example. Let $1/2 < \tau \leq 1/\sqrt{2}$.

For $1 \leq t \leq 1/\tau$

$$\mathbb{P}\{X_\infty < t\} = 2\tau p(t-1).$$

For $1/\tau < t \leq 2$

$$\mathbb{P}\{X_\infty < t\} = 2\tau p(t + \tau t - 2\tau t \log(\tau t) + 2\tau \log(\tau t) - 2).$$

There is also a special value $X_\infty = 2$:

$$\mathbb{P}\{X_\infty = 2\} = 1 - 4\tau^2 p(1 - \log(2\tau)).$$

And in this example the parameter p is the following

$$p = \frac{1}{2(1 - 4\tau^3 + 4\tau^2 + 2\tau^2 \log^2(2\tau) + 2\tau \log(\tau) - 2\tau \log(2\tau))}.$$

Example 2

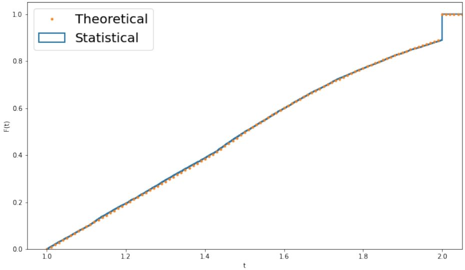

Figure: Theoretical and statistically computed values of CDF of X_∞ with $\tau = 0.7$.

Thank you for attention!

Bohun, V., *Probabilistic Analysis of Vantage Point Trees*. 2021+. Preprint is available at `http://do.unicyb.kiev.ua/images/publications/bohun/bohun_vp_trees.pdf`.

Yianilos, P., *Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces*. Fourth annual ACM-SIAM symposium on Discrete algorithms, 1993, 311-321.