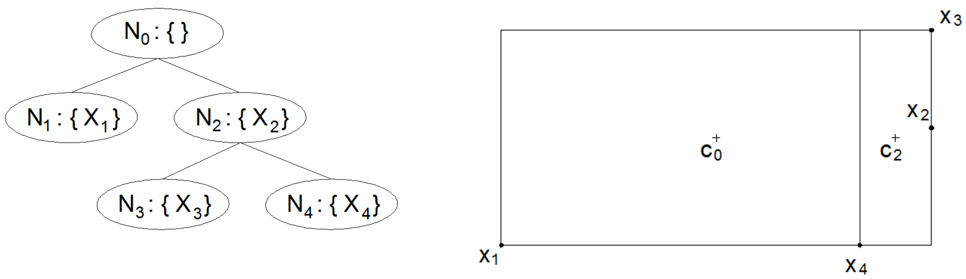
Neighbor Search's Bounds

Marcos Pividori

While the previous proof works for ball trees, it doesn't seem to be correct for different bounds.
The problem is in the assumption:
"The ball of radius $\lambda(N_c)$ centered at the center of the node N_c lies entirely within the ball of radius $\lambda(N_q)$ centered at the center of the node N_q."
This is not always true for some bounds such as hyperrectangles used in KDTrees. Let's see an example:
Let consider a four-point dataset $\{x_1, x_2, x_3, x_4\} \subseteq \mathbb{R}^2$.

$x_1 = (0, 0) \qquad x_2 = (12, \frac{36}{11}) \qquad x_3 = (12, 6) \qquad x_4 = (10, 0)$

(x_2 particularly chosen to be aligned with c_2 and x_1, making the proof simpler)
The abstract representation of the space tree is shown in the Figure 1a, and a $\mathbb{R}^2$ representation including convex subsets can be seen in Figure 1b. c_0 and c_2 represent the centroids of the nodes N_0 and N_2 respectively.



(a) Abstract representation.

(b) $\mathbb{R}^2$ representation.

As can be seen in Figure 2, the ball B_0 (ball of radius $\lambda(N_0)$ centered at the center of node N_0) doesn't completely include the ball B_2 (ball of radius $\lambda(N_2)$ centered at the center of node N_2).
If we consider the point x_1 in the figure, we can prove that the distance between x_1 and the closest point $y \in B_2$ is greater than $2\,\lambda(N_0) - 2\,\lambda(N_2)$. In contradiction to what was mentioned in the previous proof ("The furthest possible distance between any $p_q \in B_q$ and the closest $p_c \in B_c$ is $2\,\lambda(N_q) - 2\,\lambda(N_c)$").

$$dist(x_1, y) > 2\,\lambda(N_0) - 2\,\lambda(N_2) \tag{1}$$

This is easy to see from the figure, a proof could be provided if necessary.

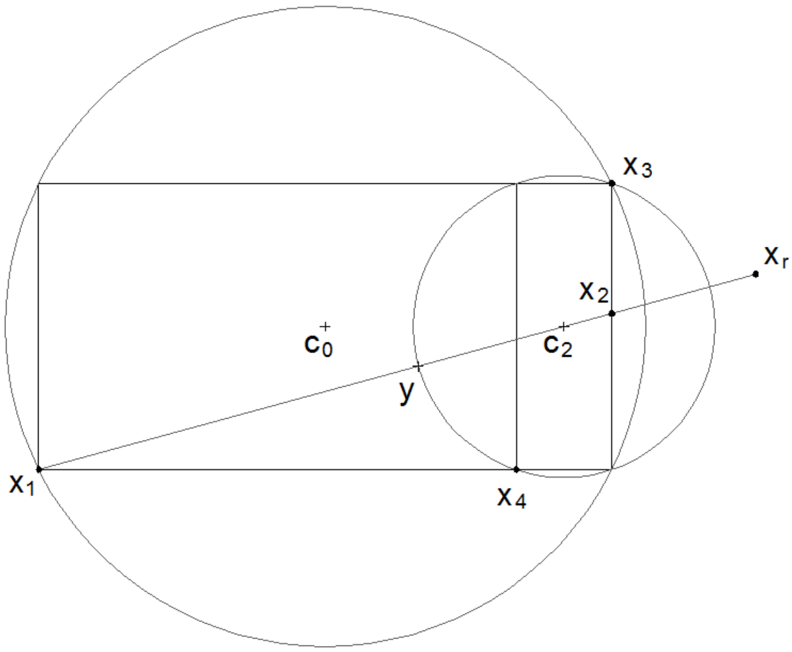


Figure 2: $\mathbb{R}^2$ representation incliuding B_2 and B_0

Let's consider a reference dataset $\{x_r\} \subseteq \mathbb{R}^2$, $x_r = (15, \frac{45}{11})$
(x_r particularly chosen to be aligned with x_1, c_2 and x_2, making the proof simpler)
It is easy to see from the figure that:

$$dist(x_2, x_r) < dist(x_3, x_r) \tag{2}$$
$$dist(x_2, x_r) < dist(x_4, x_r) \tag{3}$$

Assuming *1-nearest neighbor search* ($k = 1$), let's analyze the value of B_2 bound after traversing the space tree considering the reference point x_r:

$$B_2(N_4) = dist(x_4, x_r)$$

$$B_2(N_3) = dist(x_3, x_r)$$

$$\begin{aligned} B_2(N_2) &= min\{min_{p \in \mathscr{P}_2}(Dp[1] + \rho(N_2) + \lambda(N_2))\,,\, min_{N_c \in \mathscr{C}_2}(B_2(N_c) + 2\,(\lambda(N_2) - \lambda(N_c)))\} \\ &= min\{dist(x_2, x_r) + dist(c_2, x_2) + \lambda(N_2)\,,\, B_2(N_3) + 2\,\lambda(N_2)\,,\, B_2(N_4) + 2\,\lambda(N_2)\} \\ &= min\{dist(x_2, x_r) + dist(c_2, x_2) + \lambda(N_2)\,,\, dist(x_3, x_r) + 2\,\lambda(N_2)\,,\, dist(x_4, x_r) + 2\,\lambda(N_2)\} \\ &= min\{dist(x_2, x_r) + dist(c_2, x_2) + dist(c_2, y)\,,\, dist(x_3, x_r) + 2\,dist(c_2, y) \\ &\qquad ,dist(x_4, x_r) + 2\,dist(c_2, y)\} \end{aligned}$$

Since (2), (3) and $dist(c_2, x_2) < dist(c_2, y)$, results:

$$dist(x_2, x_r) + dist(c_2, x_2) + dist(c_2, y) < dist(x_3, x_r) + 2\,dist(c_2, y)$$
$$dist(x_2, x_r) + dist(c_2, x_2) + dist(c_2, y) < dist(x_4, x_r) + 2\,dist(c_2, y)$$

So, therefore:

$$
\begin{aligned}
B_2(N_2) &= dist(x_2, x_r) + dist(c_2, x_2) + dist(c_2, y) \\
&= dist(y, x_r)
\end{aligned}
$$

$$
B_2(N_1) = dist(x_1, x_r)
$$

$$
\begin{aligned}
B_2(N_0) &= min\{min_{p \in \mathscr{P}_0}(Dp[1] + \rho(N_0) + \lambda(N_0)), min_{N_c \in \mathscr{C}_0}(B_2(N_c) + 2\,(\lambda(N_0) - \lambda(N_c)))\} \\
&= min\{B_2(N_1) + 2\,(\lambda(N_0) - \lambda(N_1)), B_2(N_2) + 2\,(\lambda(N_0) - \lambda(N_2))\} \\
&= min\{dist(x_1, x_r) + 2\,(\lambda(N_0) - \lambda(N_1)), dist(y, x_r) + 2\,(\lambda(N_0) - \lambda(N_2))\} \\
&= min\{dist(x_1, x_r) + 2\,\lambda(N_0), dist(y, x_r) + 2\,(\lambda(N_0) - \lambda(N_2))\} \\
&= dist(y, x_r) + 2\,(\lambda(N_0) - \lambda(N_2))
\end{aligned}
$$

As mentioned at the beginning (1): $dist(x_1, y) > 2\,\lambda(N_0) - 2\,\lambda(N_2)$
Therefore: $dist(x_1, y) + dist(y, x_r) > 2\,\lambda(N_0) - 2\,\lambda(N_2) + dist(y, x_r)$
Resulting in: $dist(x_1, x_r) > B_2(N_0)$
So, $B_2(N_0)$ **is not** an upper bound on the distance between any descendant point of N_0 and its *1-nearest neighbor*.
We could make errors when prunning, considering actual B_2 bound definition.
For example, if we increase the reference dataset with another point x'_r (Figure 3), included in a leaf node N'_r, x'_r aligned with c_0 and x_1, and at the fixed distance:

$dist(x_1, x'_r) = B_2(N_0) + (dist(x_1, x_r) - B_2(N_0))/2$

Clearly $dist(x_1, x'_r) < dist(x_1, x_r)$, but the node combination: (N_0, N'_r) will be pruned because:

$d_{min}(N_0, N'_r) = dist(x_1, x'_r) > B_2(N_0)$

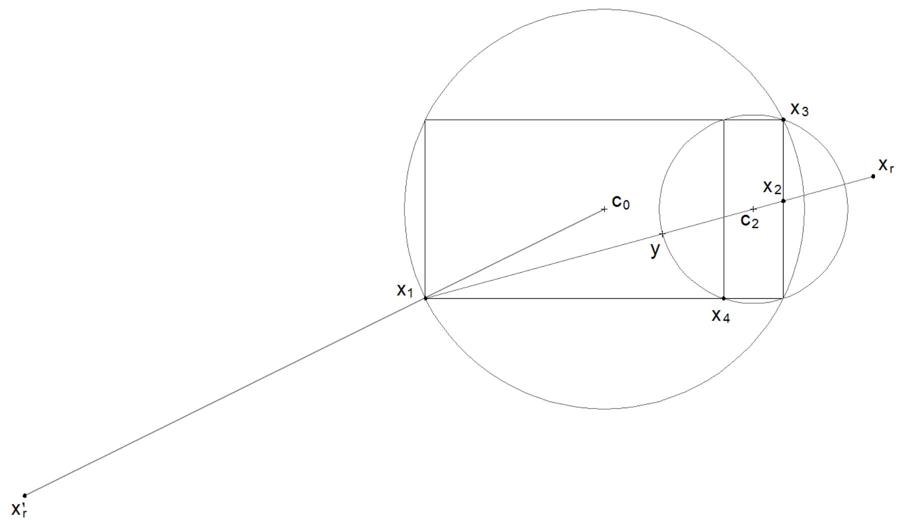


Figure 3