A quick note on $B_2(\cdot)$

Ryan R. Curtin

May 25, 2016

Our goal here is to show that $B_2(\cdot)$ from [1] is a correct bound for nearest neighbor search. We can prove this, but first let's rewrite the bound function itself:

$$B_2(\mathscr{N}_q) = \min\{\min_{p \in \mathscr{P}_q} \left(D_p[k] + \rho(\mathscr{N}_q) + \lambda(\mathscr{N}_q)\right), \min_{\mathscr{N}_c \in \mathscr{C}_q} \left(B_2(\mathscr{N}_c) + 2(\lambda(\mathscr{N}_q) - \lambda(\mathscr{N}_c))\right)\}. \quad (1)$$

Theorem 1. *$B_2(\mathscr{N}_q)$ gives, for any $\mathscr{N}_q$, an upper bound on the distance between any descendant point of $\mathscr{N}_q$ and its k-nearest neighbor.*

Proof. To prove the correctness of $B_2(\mathscr{N}_q)$, we have to consider two cases: when $\mathscr{N}_q$ is a leaf (has no children), and when $\mathscr{N}_q$ is not a leaf. This strategy resembles induction, where the base case is a leaf.

First, consider when $\mathscr{N}_q$ is a leaf. In this setting, the second min in Equation 1 does not evaluate since $|\mathscr{C}(\mathscr{N}_q)| = 0$. So we only need to consider the first term. Also, when $\mathscr{N}_q$ is a leaf, $\lambda(\mathscr{N}_q) = \rho(\mathscr{N}_q)$ because $\mathscr{P}_q = \mathscr{D}^p_q$ (that is, the set of points held in $\mathscr{N}_q$ is the same as the set of descendant points of $\mathscr{N}_q$). Thus in this case,

$$B_2(\mathscr{N}_q) = \min_{p \in \mathscr{P}_q} D_p[k] + 2\lambda(\mathscr{N}_q). \quad (2)$$

We can show the correctness here using the triangle inequality. Any point in $\mathscr{P}_q$ is separated from any other points in $\mathscr{P}_q$ by a maximum of $2\lambda(\mathscr{N}_q)$. Thus, if there exists some point p with k-furthest neighbor candidate distance $D_p[k]$, then for any other point p_i in $\mathscr{P}_q$, then

$$\begin{aligned} D_{p_i}[k] &\leq D_p[k] + d(p, p_i) & (3)\\ &\leq D_p[k] + 2\lambda(\mathscr{N}_q). & (4) \end{aligned}$$

Thus, $B_2(\mathscr{N}_q)$ is correct when $\mathscr{N}_q$ is a leaf. Now, let us consider the other case, where $\mathscr{N}_q$ is not a leaf. Here we must prove that both sides of Equation 1 are correct. We will consider the first side first, with a similar argument.

Since $\mathscr{N}_q$ is not a leaf, then $\rho(\mathscr{N}_q) \leq \lambda(\mathscr{N}_q)$ (that is, we do not have strict equality). We know that any point in $\mathscr{P}_q$ (any point held in $\mathscr{N}_q$) is separated from $\mathscr{D}^p_q$ (any descendant point of $\mathscr{N}_q$) by at most $\rho(\mathscr{N}_q) + \lambda(\mathscr{N}_q)$. Thus, if there exists some point $p \in \mathscr{P}_q$ with k-furthest neighbor candidate distance $D_p[k]$, then for any descendant point $p_i \in \mathscr{D}^p_q$, then

$$\begin{aligned} D_{p_i}[k] &\leq D_p[k] + d(p, p_i) & (5)\\ &\leq D_p[k] + \rho(\mathscr{N}_q) + \lambda(\mathscr{N}_q). & (6) \end{aligned}$$

Now we may turn to proving the correctness of the second side of Equation 1. Assume that $\mathscr{B}_2(\mathscr{N}_c)$ is valid for each child $\mathscr{N}_c$ of $\mathscr{N}_q$ (that is, it satisfies the statement of the theorem). This means that $\mathscr{B}_2(\mathscr{N}_c)$ is a valid upper bound on the distance between any descendant point of $\mathscr{N}_c$ and its k-nearest neighbor. But we can actually say something slightly stricter due to the way $B_2(\mathscr{N}_c)$ is constructed: $B_2(\mathscr{N}_c)$ is a valid upper bound on the distance between *any point that falls into the ball of radius $\lambda(\mathscr{N}_c)$ centered at the center of the node $\mathscr{N}_c$* and its k-nearest neighbor.

The ball of radius $\lambda(\mathscr{N}_c)$ centered at the center of the node $\mathscr{N}_c$ lies entirely within the ball of radius $\lambda(\mathscr{N}_q)$ centered at the center of the node $\mathscr{N}_q$. For simplicity for what I'm about to write, call B_i the ball of radius λ_i centered at the center of node $\mathscr{N}_i$.

Then, for any point $p_q \in B_q$ and any point $p_c \in B_c$, we may construct a valid upper bound u_q on the k-nearest neighbor of p_q:

$$u_q = D_{p_c}[k] + d(p_q, p_c). \tag{7}$$

If $p_q \in B_c$ (that is, p_q not only is contained in the ball B_q but also in B_c) then we may simply pick $p_q = p_c$ so $d(p_q, p_c) = 0$. And if $p_q \notin B_c$, we can pick the closest point in B_c to p_q. The furthest possible distance between any $p_q \in B_q$ and the closest $p_c \in B_c$ is $2\lambda(\mathscr{N}_q) - 2\lambda(\mathscr{N}_c)$. (Maybe it is easiest to see this geometrically, but I don't feel like drawing out the figure for this 'short' response.)

Thus we can conclude that in any situation, $d(p_q, p_c) \leq 2(\lambda(\mathscr{N}_q) - \lambda(\mathscr{N}_c)$. Therefore

$$u_q = D_{p_c}[k] + 2(\lambda(\mathscr{N}_q) + \lambda(\mathscr{N}_c)) \tag{8}$$

and since $B_2(\mathscr{N}_q)$ is a valid upper bound for any point $p_c \in B_c$, we may simplify to

$$u_q = B_2(\mathscr{N}_q) + 2(\lambda(\mathscr{N}_q) + \lambda(\mathscr{N}_c)). \tag{9}$$

We know that u_q is valid for any $p_q \in B_q$; thus, we can conclude that the second term in Equation 1 is a valid upper bound on the k-nearest neighbor for any p_q that is a descendant point of $\mathscr{N}_q$.

Combining upper bounds via min still gives valid upper bounds, so the statement of the theorem holds. $\square$

References

[1] Ryan R. Curtin, William B. March, Parikshit Ram, David V. Anderson, Alexander G. Gray, and Charles L. Isbell Jr. Tree-independent dual-tree algorithms. In *Proceedings of The 30th International Conference on Machine Learning (ICML '13)*, pages 1435–1443, 2013.