Solutions to Quiz 4

Solution 1:

(a) (i) and (iv) are true. Path compression guarantees that long sequences of trivial splits do not occur, which implies that the tree has $O(n)$ nodes. But since the tree need not be balanced, its height may be as large as $O(n)$.

(a) The size of an s-WSPD for n points in $\mathbb{R}^d$ is $O(s^d n)$.

(b) By the WSPD Utility Lemma and the fact that x and x' could be equal,

$$0 \;\leq\; \frac{\|x - x'\|}{\|x - y\|} \;\leq\; \frac{2}{s}.$$

(c) By applying the WSPD Utility Lemma to both the ratio and its reciprocal, we have,

$$\frac{1}{1 + 4/s} \;\leq\; \frac{\|x' - y'\|}{\|x - y\|} \;\leq\; 1 + \frac{4}{s}.$$

(d) (i) By the Chain Property of kernels, X is a $(\varepsilon + \varepsilon')$-kernel of Z.

(e) Such a set is just an ε-net for $\varepsilon = m/n$. By the ε-Net Lemma, with constant probability, a random sample is an ε-net if it contains at least $O((d/\varepsilon)\log(1/\varepsilon))$ points. So the size of the random sample is $O(d(n/m)\log(n/m))$.

Solution 2: This solution is based on spanners, but it can be expressed in terms of WSPDs as well. We begin by computing a t-spanner for P, where $t = 1 + \varepsilon$. Recall that this can be done in $O(n \log n + (1/\varepsilon)^d n)$ time. For each point p, we consider the neighbors of p in this graph, and select the point q among these neighbors that is closest to p. This can be done in time proportional to the total size of the spanner, which is $O((1/\varepsilon)^d n)$.

To establish the correctness of this algorithm, consider any point p and let p^* be p's true nearest neighbor. We know that from spanner properties, there exists a path of length at most $t\|p - p^*\| = (1 + \varepsilon)\|p - p^*\|$. Let p' be the first vertex on the spanner path from p to p^*. If $p' = p^*$, then the above algorithm returns the true nearest neighbor to p. Otherwise, (by the triangle inequality) $\|p - p'\| \leq (1 + \varepsilon)\|p - p^*\|$, implying that p' is an ε-approximate nearest neighbor to p. Because the algorithm considers p' among the possible candidates for nearest neighbors, the point that it outputs with p will be at least as close, and therefore it is a valid ε-nearest neighbor.

Here is an alternative construction based directly on the WSPD. Use a separation factor of at least $s = 2/\varepsilon$. For each WSP $(P(u), P(v))$, let $\text{rep}(u)$ and $\text{rep}(v)$ denote the associated representatives. For each point $p \in P$, consider all the pairs $(P(u), P(v))$ where p is the representative point on either side of the pair. Define $N(p)$ to be the set of representative points on the other sides of these pairs. Select the point $q \in N(P)$ that is closest to p as the approximate nearest neighbor.

We assert that q is the desired approximate nearest neighbor of p. To see why, let p^* be the true nearest neighbor of p, and let $(P(u), P(v))$ be the pair such that $p \in P(u)$ and $p^* \in P(v)$.

It is easy to see that p is the only point of $P(u)$ (since any other point would have been even closer than p^*) and hence $p = \text{rep}(u)$. By the WSPD Utility Lemma, $\text{rep}(v)$ is within distance $(1 + 2/s)\|p - p^*\| = (1 + \varepsilon)\|p - p^*\|$, implying that $N(p)$ contains an approximate nearest neighbor of p. Because there are $O(n/\varepsilon^d)$ well-separated pairs, the total number size of the sets $N(P)$ is also $O(n/\varepsilon^d)$.

Solution 3:

(a) Consider the points $\{a, b, c, d\} = \{(0, 1), (1, 0), (1, 2), (2, 1)\}$, and consider the right triangle whose left and bottom sides pass through a and b, respectively, and whose hypotenuse passes through c and d (see Fig. 1(a)). By "wobbling" the sides of this triangle, we can either include or exclude any combination we like of the four points $\{a, b, c, d\}$.

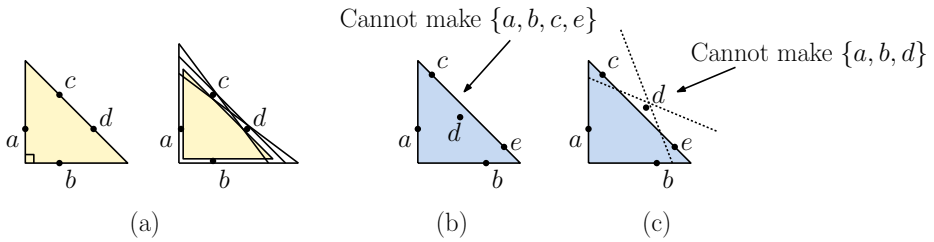

Figure 1: Shattering using axis-aligned right triangles.

(b) Consider any set of five points $\{a, b, c, d, e\} \in \mathbb{R}^2$. Let us label the points so that a is the leftmost point, b is the bottommost point. The inclusion of the remaining points $\{c, d, e\}$ depends entirely on the hypotenuse, and in particular, those points that lie in the lower halfplane defined by the hypotenuse are in the range.

However, the range space of lower halfplanes has VC-dimension 2, and hence it cannot shatter any 3-element set. To see why, sort $\{c, d, e\}$ from left to right. Observe that if d lies below $\overline{ce}$, then any lower halfspace that contains c and e, must contain d, therefore $\{c, e\}$ cannot be generated. On the other hand, if d lies above $\overline{ce}$, then any lower halfplane that contains d must contain either c or e (or both), and therefore the range $\{d\}$ cannot be generated. (see Fig. 1(c)).

Solution 4: Let C^* denote the optimal set of centers, and let $r^* = r^*(P, k)$. Let C' denote the crude approximation, and let r' denote its covering radius. As observed in the problem statement, we have $r^* \leq r' \leq 3r^*$. As given in the problem statement, this approximation can be computed in $O(n \log n)$ time.

To construct the coreset, build a hypercube grid of side length $\varepsilon r'/(3\sqrt{d})$. The cells of this grid each have diameter $\delta = \varepsilon r'/3$. Hash all the points of P into these grid squares and select one representative point from each nonempty grid square. Under the assumption that hashing takes $O(1)$ time per point, this can be done in $O(n)$ time. We assert that the resulting set of representatives, denoted R, is the desired coreset (see Fig. 2(a)).

We first observe that $|R| = O(k/\varepsilon^d)$. To see this, observe that every point of P lies within distance r' of one of the k points of C'. This implies that every point lies within a hypercube of side length $2r'$ centered about each of the points of C' (see Fig. 2(b)). Given our assumption that d

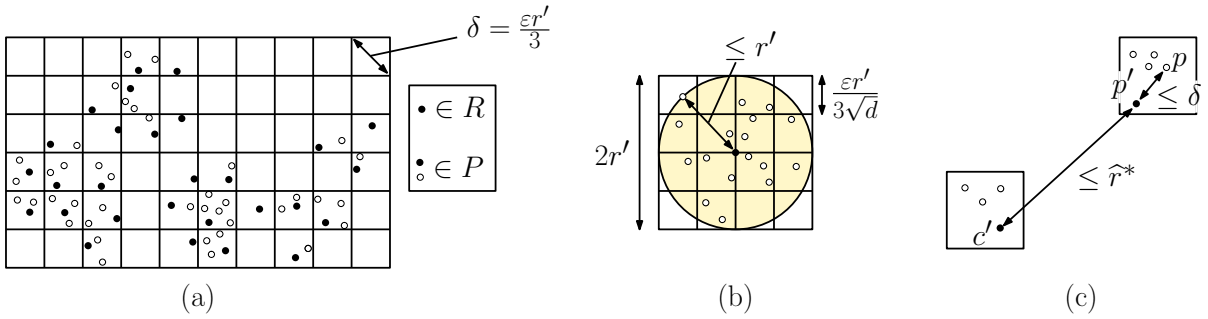

Figure 2: Euclidean k-center coreset.

is constant, it follows that (up to constant factors), the number of grid cells of side length $\varepsilon r'/2\sqrt{d}$ overlapping such a hypercube is at most

$$\left(\frac{2r'}{\varepsilon r'/3\sqrt{d}}\right)^d = \left(\frac{6\sqrt{d}}{\varepsilon}\right)^d = O\left(\left(\frac{1}{\varepsilon}\right)^d\right).$$

Taking the union over all the k centers of $\widehat{C}$, we have a total of at most $O(k/\varepsilon^d)$ nonempty cells, and hence $|R| = O(k/\varepsilon^d)$. The overall running time is $O(n \log n + n + |R|) = O(n \log n)$.

To establish correctness, we will show that R is an ε-coreset. Let $\widehat{r}^* = r^*(R, k)$ denote the optimal covering radius for R, and let $\widehat{C}^*$ denote the set of centers for this solution. It suffices to show that $(1-\varepsilon)r^* \leq \widehat{r}^* \leq r^*$. The second inequality holds trivially because $R \subseteq P$, and hence any set of disks that covers P also covers R.

We first prove that $r^* \leq \widehat{r}^* + \delta$. It suffices to show that every point of P lies within distance $\widehat{r}^* + \delta$ of some point of $\widehat{C}^*$. (Since this holds for $\widehat{C}^*$, it certainly holds for the optimal centers.) To see why, consider any $p \in P$. Let p' denote the representative point from p's grid cell (see Fig. 2(c)). Because R has a covering of radius $\widehat{r}^*$, there exists a center $c' \in \widehat{C}^*$ such that $\|p' - c'\| \leq \widehat{r}^*$. Because the diameter of the cell is δ, we have $\|p - p'\| \leq \delta$, and hence by the triangle inequality,

$$\|p - c'\| \;\leq\; \|p - p'\| + \|p' - c'\| \;\leq\; \delta + \widehat{r}^*,$$

as desired. Equivalently, $r^* - \delta \leq \widehat{r}^*$. By our choice of δ and the fact that $\widehat{r} \leq 3r^*$, we have $\delta = \varepsilon\widehat{r}/3 \leq \varepsilon r^*$. Thus, $\delta \leq \varepsilon r^*$, which implies that $(1-\varepsilon)r^* \leq r^* - \delta$. Therefore, we have

$$(1-\varepsilon)r^* \;\leq\; \widehat{r}^* \;\leq\; r^*,$$

implying that R is an ε-coreset.