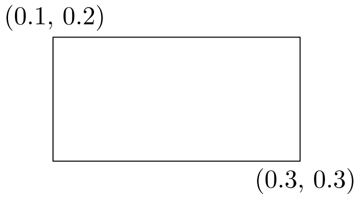
INFS 4205/7205: Exercise Set 2

Prepared by Yufei Tao and Junhao Gan

In all the following problems, a d-dimensional data space is defined to be $[0, 1]^d$, namely, each dimension has a domain from 0 to 1.

Problem 1. Fix a rectangle r whose top-left and bottom-right corners have coordinates $(0.1, 0.2)$ and $(0.3, 0.3)$, respectively (see the figure below). Let q be a 2D square with side length 0.3, whose location is uniformly distributed on condition that q intersects the data space. What is the probability that q intersects r?



Solution. The rectangle has side lengths 0.2 and 0.1. Hence, the probability is $\frac{0.2+0.3}{1+0.3} \cdot \frac{0.1+0.3}{1+0.3} = 0.118$.

Problem 2. In the class, we considered that the query rectangle is uniformly distributed "on condition that" it intersects the data space. The instructor mentioned that this allowed us to focus on the core of the analysis without unnecessary details. In this problem, we will understand why—for which purpose, it suffices to look at the 1D space.

- Fix an interval $I = [0.1, 0.3]$. Let q be an interval of length 0.4, whose location is uniformly distributed on condition that q is *within* the data space (i.e., range $[0, 1]$). What is the probability that q intersects I?
- Now consider the general form of the previous question. Fix an interval $I = [a, b]$. Let q be an interval of length $\ell < 1$, whose location is uniformly distributed on condition that q is *within* the data space. What is the probability that q intersects I?

Solution.

- The center of q ranges in the interval $[0.2, 0.8]$. For q to intersect I, the center of q must be in $[0.2, 0.5]$. Therefore, the probability that q and I intersects is $\frac{0.5-0.2}{0.8-0.2} = 0.5$.
- The center of q ranges in the interval $[\ell/2, 1 - \ell/2]$. For q to intersect I, the center of q must be in $[\max\{a - \ell/2, \ell/2\}, \min\{b + \ell/2, 1 - \ell/2\}]$. Therefore, the probability equals

$$\frac{\min\{b + \ell/2, 1 - \ell/2\} - \max\{a - \ell/2, \ell/2\}}{1 - \ell}.$$

Problem 3. Define $Q(\ell)$ as the set of queries such that their search regions (i) are squares of side length ℓ, and (ii) intersect the data space. In the class, we have seen that, in 2D space, as long as we remember 3 values VolSum, PeriSum, and NumNodes, it is possible to obtain the expected number of node accesses for a query chosen uniformly at random from $Q(\ell)$. Furthermore, this is true without having to know the value of ℓ in advance.

Define $Q_\rho(\ell)$ as the set of queries such that their search regions (i) are rectangles of side lengths ℓ and $\ell \cdot \rho$, respectively, and (ii) intersect the data space. Prove a similar conclusion on $Q_\rho(\ell)$, namely, as long as we remember 3 values and target a specific value of ρ, we can obtain the expected number of node accesses for a query chosen uniformly at random from $Q_\rho(\ell)$; furthermore, this is true without having to know the value of ℓ in advance.

Solution. As discussed in the lecture, the expected number of node accesses is:

$$\begin{aligned}
& \sum_u \frac{(s_1(u)+\ell_1)(s_2(u)+\ell_2)}{(1+\ell_1)(1+\ell_2)} \\
= & \frac{1}{(1+\ell_1)(1+\ell_2)}\left(\sum_u s_1(u)s_2(u) + \sum_u(\ell_2 \cdot s_1(u) + \ell_1 \cdot s_2(u)) + \sum_u \ell_1\ell_2\right) \\
= & \frac{1}{(1+\ell)(1+\rho\ell)}\left(\text{VolSum} + \ell \cdot \sum_u (\rho \cdot s_1(u) + s_2(u)) + \text{NumNodes} \cdot \rho\ell^2\right)
\end{aligned}$$

Therefore, for a specific ρ, it is sufficient to record only 3 values: VolSum, NumNodes and $\sum_u(\rho \cdot s_1(u) + s_2(u))$.

Problem 4. Consider now 3D space. Define $Q(\ell)$ as the set of queries such that their search regions (i) are (3D) hyper-squares of side length ℓ, and (ii) intersect the data space. Prove that, as long as we remember 4 values, it is possible obtain the expected number of node accesses for a query chosen uniformly at random from $Q(\ell)$. Furthermore, this is true without having to know the value of ℓ in advance.

Solution. For a query q chosen uniformly at random from $Q(\ell)$, the expected number of node accesses for q is as follows:

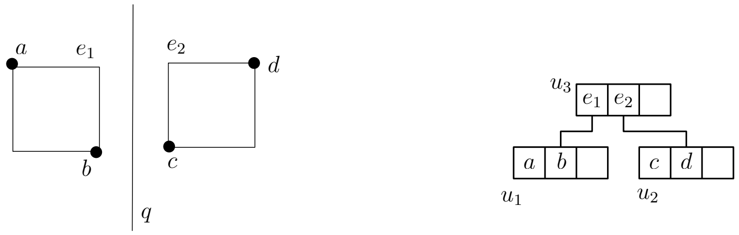
$$\begin{aligned}
& \sum_u \frac{(s_1(u)+\ell)(s_2(u)+\ell)(s_3(u)+\ell)}{(1+\ell)^3} \\
= & \frac{1}{(1+\ell)^3}\sum_u (s_1(u)+\ell)(s_2(u)+\ell)(s_3(u)+\ell) \\
= & \frac{1}{(1+\ell)^3}\sum_u \Big(s_1(u)s_2(u)s_3(u) + \ell\,(s_1(u)s_3(u) + s_2(u)s_3(u) + s_1(u)s_2(u)) + \\
& \ell^2(s_1(u)+s_2(u)+s_3(u)) + \ell^3\Big) \\
= & \frac{1}{(1+\ell)^3}\Big(\text{VolSum} + \ell \cdot \sum_u (s_1(u)s_3(u) + s_2(u)s_3(u) + s_1(u)s_2(u)) + \\
& \ell^2 \cdot \text{PeriSum}/4 + \ell^3 \cdot \text{NumNodes}\Big)
\end{aligned}$$

Therefore, it is sufficient to record only 4 values: VolSum, PeriSum, NumNodes and $\sum_u(s_1(u)s_3(u)+ s_2(u)s_3(u) + s_1(u)s_2(u))$.

Problem 5. Let us define the *level* of a node as follows. The level of a leaf is 0. Inductively, if the level of a node is i, then its parent is at level $i+1$. Prove or disprove:

- Any range query must access at least as many level-1 nodes as leaf nodes.
- Any range query must access at least as many leaf nodes as level-1 nodes.

Solution. Both statements are false. Consider the R-tree as shown below. To disprove the first statement, simply take the query whose search region covers the whole data space. To disprove the second, use the query whose search region is the line q in the figure.



a
e_1
e_2
d
b
c
q
u_3
e_1
e_2
a
b
c
d
u_1
u_2