k-means++ Algorithm for Clustering[1]

- **Clustering Problems.** In this lecture we will see the use of randomization for an approximation algorithm in *clustering*. Clustering problems form a very important class of problems both in machine learning where one clusters data to understand patterns, and also in operations research where they appear as "facility location" problems which try to figure out where to open service facilities which maximizes certain happiness of the clients they serve. There are gazillions of types of clustering problems, and numerous kinds of algorithms to solve them (and just clustering would be a full term course). In this lecture or two, we are going to look at a particular, but very widely used, formulation *and* algorithm.

- **The Setting.** We assume P is a set of points in d-dimensional space, that is, $\mathbb{R}^d$. We also assume we are given a positive integer k. Our goal is to *partition* the point set into k parts. That is, we need to find $\Pi := (P_1, P_2, \ldots, P_k)$ which minimizes a certain cost function. Our cost function is going to be *additive* over the parts, that is, the cost of Π will be the sum of costs of the singleton cluster P_i, $1 \leq i \leq k$. We now explain what is the cost of a single subset of points.

  The k-means cost of any subset $A \subseteq P$ is defined as follows: $\min_{x\in\mathbb{R}^d} \sum_{p\in A} ||p - x||_2^2$. That is, we find the best *center* x, which can be any point in $\mathbb{R}^d$ (and not necessarily in P), that minimizes the squared sum of Euclidean distances of every point in P to this center x. A priori, this itself may seem like a difficult problem. This is where the particular choice of the objective function comes in (and probably explains its popularity in practice): the point x is simply the average of the points in A.

  Fact 1. For any subset $A \subseteq \mathbb{R}^d$, the $x \in \mathbb{R}^d$ minimizing $\sum_{p\in A} ||p - x||_2^2$ is $\overline{p} := \frac{1}{|A|}\sum_{p\in A} p$.

  Exercise: *Prove the above fact.*

- **Relaxed Triangle Inequality.** Fact 1 notwithstanding, we are going to look at things a bit more abstractly. We are going to assume the "*distance/cost*" between any two points p, q is some non-negative quantity $d(p, q)$. In the above setting, $d(p, q) = ||p - q||_2^2$, the squared Euclidean distance. The only property of these distances we will use is *symmetry*, that is $d(p, q) = d(q, p)$ and the following *relaxed triangle inequality*.

  $$\text{For any three points } p, q, r, \text{ we have,} \quad d(p, r) \leq 2\left(d(p, q) + d(q, r)\right) \tag{1}$$

  We will *not* assume anything else, in particular, we will not assume Fact 1 for this exposition. As a result, when we specify our solution $\Pi = (P_1, P_2, \ldots, P_k)$, for each P_i we will also specify the center a_i (which may not be in P). And once we do so, the cost of our solution will be

  $$\mathsf{cost}(\Pi) := \sum_{i=1}^{k} \mathsf{cost}(P_i; a_i) \quad \text{where} \quad \mathsf{cost}(P_i; a_i) := \sum_{p\in P_i} d(p, a_i)$$

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 25th April, 2021*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

- **The Cluster Centers Suffice.** In fact, the next observation is that one doesn't need to specify the parts P_i's at all; just specifying the centers are enough. In particular, if we select $C := \{a_1, \ldots, a_k\}$ as our potential centers, then that defines the parts P_i's as well. Indeed, $P_i \subseteq P$ will be the set of points p such that $d(p, a_i)$ is at most $d(p, a_j)$ for any $j \neq i$. Indeed, this is an important notation we will use.

  Definition 1. Given any subset C of centers, and any subset of points $A \subseteq P$, we define the cost of connecting A to C as follows. First, for any point p we define

  $$d(p, C) := \min_{c \in C} d(p, c) \quad \text{and} \quad \mathsf{cost}(A; C) := \sum_{p \in A} d(p, C)$$

  Given any subset of centers C, let Π be the partition of P defined by C. In particular, if $C = \{a_1, \ldots, a_k\}$, then $\Pi = (P_1, \ldots, P_k)$ where $P_i := \{p \in P : d(p, a_i) \leq d(p, a_j), \;\; \forall j \in [k]\}$. Then, the clustering cost of the centers is

  $$\mathsf{cost}(\Pi) = \sum_{i=1}^{k} \sum_{p \in P_i} d(p, a_i) \qquad \underbrace{=}_{\text{convince yourself}} \qquad \mathsf{cost}(P; C)$$

  We will use $\Pi^* = (P_1^*, \ldots, P_k^*)$ as the optimal clustering defined by the optimal cluster centers C^*.

- **The k-means++ Algorithm.** Without further ado, we discuss the simple k-means++ algorithm. It is an *importance sampling* algorithm which is extremely easy to state and implement, but whose analysis is a bit tricky.

  The algorithm picks the k centers C in order. These centers, in fact, are a subset of the original points itself. The first center c_1 is picked uniformly at random among P, and $C \leftarrow c_1$. Henceforth, every point $p \in P$ maintains $d(p, C)$ as in Definition 1, and the next center is chosen with probability *proportional* to $d(p, C)$. The algorithm stops when k centers have been picked.

```
procedure k-MEANS++(P): ▷ This is also called D^2-sampling in the literature
    $C \leftarrow \emptyset$.
    Select $c_1$ uniformly at random among the points in $P$.
    for $i = 2$ to $k$ do:
        Select $c_i$ among $P$ with probability proportional to $d(p, C)$. More precisely,

            For any $p \in P$,  $\mathbf{Pr}[c_i = p] = \frac{d(p, C)}{\mathsf{cost}(P; C)}$

        Add $C \leftarrow C \cup \{c_i\}$.
```

Theorem 1 (D. Arthur, S. Vassilvitskii, 2007)**.** The expected cost of the final set returned in an $O(\log k)$ approximation. More precisely,

$$\mathbf{Exp}[\mathsf{cost}(P; C)] \leq 16 \cdot (1 + H_k)\mathsf{cost}(\Pi^*)$$

where H_k is the kth Harmonic number.

- **Baby Step 1 : Uniform Center in a Single Subset.** The analysis follows in many steps. The first observation is the following. Fix any subset $A \subseteq P$ of the points. Let x be the point which minimizes $\sum_{p \in A} d(p,x)$. That is, x is the optimal choice of center for the subset A resulting in a clustering cost of $\mathsf{opt}(A)$. Note that x may not lie in A at all. Indeed, when $d(\cdot,\cdot)$ is the squared Euclidean distance, then Fact 1 tells us that x is the mean of A. Nevertheless, the following lemma states that if we instead select a r *uniformly at random* from A as our center, then the expected cost of connecting all the points in A to r is at most a 4-factor of $\mathsf{opt}(A)$.

  Lemma 1. For any $A \subseteq P$, let r be a point uniformly drawn from A. Then,

$$\mathbf{Exp}[\mathsf{cost}(A;\{r\})] = \frac{1}{|A|}\sum_{r \in A}\sum_{p \in A} d(p,r) \quad \le 4 \cdot \mathsf{opt}(A) := 4 \cdot \min_{x \in \mathbb{R}^d}\left(\sum_{p \in A} d(p,x)\right)$$

  Proof. This is a simple consequence of the relaxed triangle inequality. Fix the optimal x which obtains the cost $\mathsf{opt}(A)$. For any r and p, we get

$$d(p,r) \le 2 \cdot (d(p,x) + d(x,r))$$

  And thus,

$$\begin{aligned}
\mathbf{Exp}[\mathsf{cost}(A;\{r\})] &= \frac{1}{|A|}\sum_{p,r \in A \times A} d(p,r) \\
&\le \frac{2}{|A|}\sum_{p,r \in A \times A} (d(p,x) + d(x,r)) \\
&= \frac{2}{|A|} \cdot \left(|A| \underbrace{\sum_{p \in A} d(p,x)}_{\mathsf{opt}(A)} + |A| \underbrace{\sum_{r \in A} d(r,x)}_{\mathsf{opt}(A)} \right) = 4\mathsf{opt}(A) \qquad \square
\end{aligned}$$

  Why is the above lemma useful for analyzing k-means++? We will soon see. But qualitatively, imagine a scenario where the clusters in Π^* are "far apart". So much so that once you select a center c_1 in P_1^* say, the distances $d(p,c_1) \approx 0$ for $p \in P_1^*$, and $d(p,c_1) \approx M$, for some large M, when $p \notin P_1^*$. Then, the next center c_2 will be roughly uniformly distributed among the points not in P_1^*; one will almost surely not pick anything from P_1^*, and the remaining points have roughly the same distance to c_1 and thus will be sampled roughly uniformly. And the above lemma states that in that case the cost incurred is within a constant. The next lemma makes this rigorous.

- **Baby Step 2 : Hitting a (Optimal) Cluster is Good.** Suppose at some point of time we have picked a collection of centers C. Fix a subset of points $A \subseteq P$. Now, we pick a new center c using the importance sampling process, that is, with probability proportional to the distance to the current set of centers C. And now *condition* on the event that $c \in A$. Note that c is **not** uniformly distributed among A: it is more likely we will sample points which are farther from the set C. The next lemma shows that nonetheless the expected cost of connecting A to $C \cup \{c\}$ is $\leq 16\mathsf{opt}(A)$. In plain English, for any subset of points A, if we ever pick a center in A, then in expectation, the connection costs of the points in A is not too far from the optimal solution.

Lemma 2. Let C be an arbitrary collection of centers and let A be an arbitrary subset of P. Let $\mathsf{opt}(A) := \min_x \sum_{p \in A} d(p, x)$. Let c be a point chosen from P such that $\mathbf{Pr}[c = p] = \frac{d(p,C)}{\mathsf{cost}(P;C)}$. Let $\mathcal{E}$ be the event $\{c \in A\}$. Then,

$$\mathbf{Exp}[\mathsf{cost}(A; C \cup c) \mid \mathcal{E}] \leq 16\mathsf{opt}(A)$$

Proof. The main observation in the proof is to show that conditioned on the event $\{c \in A\}$, the distribution of c in A is, although not uniform, not "too far" from it. In particular, we claim

Claim 1. For any point $r \in A$, $\mathbf{Pr}[c = r \mid \mathcal{E}] \leq \frac{2}{|A|} \cdot \left(\frac{\mathsf{cost}(A;r)}{\mathsf{cost}(A;C)} + 1\right)$

That is, if $\mathsf{cost}(A; r) \ll \mathsf{cost}(A; C)$, then the distribution is close to being (semi)-uniform. Before we prove the claim, let us see how it easily implies the lemma. Indeed,

$$\begin{aligned}
\mathbf{Exp}[\mathsf{cost}(A; C \cup c) \mid \mathcal{E}] &= \sum_{r \in A} \mathbf{Pr}[c = r \mid \mathcal{E}] \cdot \mathsf{cost}(A; C \cup r) \\
&\leq \frac{2}{|A|} \sum_{r \in A} \left(\frac{\mathsf{cost}(A; r)}{\mathsf{cost}(A; C)} + 1\right) \cdot \mathsf{cost}(A; C \cup r) \qquad \text{by Claim 1} \\
&\leq \frac{4}{|A|} \sum_{r \in A} \mathsf{cost}(A; r) \qquad \text{since } \mathsf{cost}(A; C \cup r) \leq \min\left(\mathsf{cost}(A; r), \mathsf{cost}(A; C)\right) \\
&\leq 16\mathsf{opt}(A) \qquad \text{by Lemma 1}
\end{aligned}$$

Proof of Claim 1. The probability that $\mathcal{E}$ occurs is $\mathbf{Pr}[\mathcal{E}] = \frac{\sum_{r \in A} d(r,C)}{\sum_{p \in P} d(p,C)} = \frac{\mathsf{cost}(A;C)}{\mathsf{cost}(P;C)}$. Therefore, conditioned on the event $\mathcal{E}$, the probability that c is a particular point $r \in A$ is simply the ratio of the distance of r to C divided by the sum of distances of points only in A to C. That is,

$$\mathbf{Pr}[c = r \mid \mathcal{E}] = \frac{d(r, C)/\mathsf{cost}(P; C)}{\mathsf{cost}(A; C)/\mathsf{cost}(P; C)} = \frac{d(r, C)}{\mathsf{cost}(A; C)} \tag{2}$$

Now, we notice the following. Fix any point $p \in A$. Then the relaxed triangle inequality gives us

$$d(r, C) \leq 2 \cdot (d(r, p) + d(p, C)) \tag{3}$$

To see this, let $z \in C$ be the one minimizing $d(p, z)$, that is $d(p, C) = d(p, z)$, and use $d(r, C) \leq d(r, z) \leq 2 \cdot (d(r, p) + d(p, z))$. Next, we *average* (3) over all $p \in A$ to give

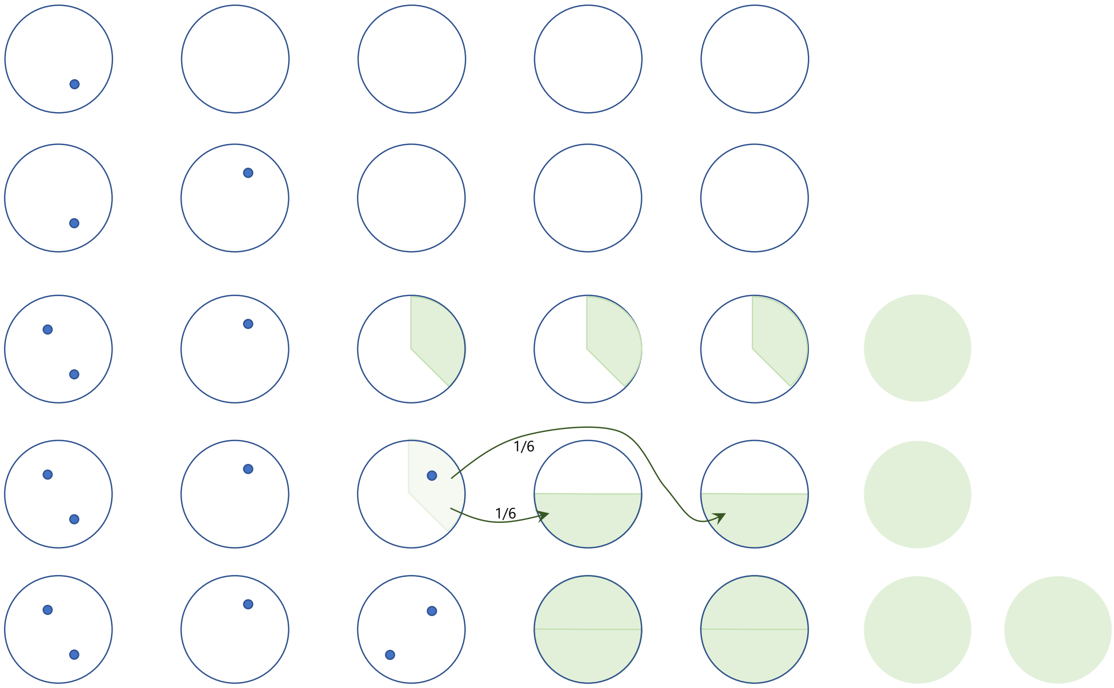
$$d(r, C) \leq \frac{2}{|A|} \cdot \left[\underbrace{\sum_{p \in A} d(p, r)}_{\mathsf{cost}(A;r)} + \underbrace{\sum_{p \in A} d(p, C)}_{\mathsf{cost}(A;C)}\right]$$

Substituting in (2) gives the claim. □

□

- **Main Workhorse : Managing Wastes and The Harmonic Sum.** Let's understand what Lemma 2 is telling us about the k-means++ algorithm. To this end, fix $\Pi^* = (P{*}_1, \ldots, P^*_k)$. Let $C := \{c_1, c_2, \ldots, c_k\}$ be the random centers picked by k-means++. From Lemma 2 we get that if $P^*_i \cap C \neq \emptyset$, then $\mathbf{Exp}[\mathsf{cost}(P^*_i; C)] = O(\mathsf{opt}(P^*_i))$. Therefore, if we were lucky and **every** part of Π^* got "hit" by the random C, then we would actually have an $O(1)$-approximation. This is what Lemma 2 is telling us. However, there is non-trivial chance that C "misses" quite a few of the P^*_i's. The remaining analysis (cleverly) bounds the expected cost of such misses. Indeed, the exposition below is heavily influenced by the analysis given by Sanjoy Dasgupta of UC San Diego in these lecture notes.

- *Intuition of the Charging Analysis.* Before I jump into the details, let me give a cartoon sketch of the argument that is to follow. For this, let me use the following figure.



We have $k = 5$ in this case, and the five blue-bordered circles in each row correspond to the 5 optimal clusters. The five rows correspond to the five iterations of the k-means++ algorithm. In the first iteration, we pick a center in the first ball, and we are happy. We are happy because the points in the first circle can be accounted for (in expectation, up to a constant factor). In the next iteration, we pick a center in the second ball, and we are happy again. In the third iteration, however, we pick a center in an already picked ball. And now, we know for ***sure*** that at least one of the three remaining balls will be "missed" by our final set C.

At this point, we pay a "penalty" which is represented by the green circle. Since we know that we will miss one ball, we pay for right at this point. However, we do not know *which* ball will be missed.

So, we smear this penalty over all three balls, and pay a third of each. A little more precisely, for each uncovered ball, we put one-third of the sum of the distances of its points to the current three centers into the penalty kitty. In the figure, all these are illustrated to be the same, and for now, this is not a bad image to keep in mind.

Here is the main analysis point: what is the probability that there was a miss in this third iteration? Let's bring in a little of notation. Suppose H was the sum of distances of points in the first two "hit" clusters to the first *two* centers. Similarly, suppose U was the sum of distances of points in the "uncovered" last three clusters. Then the probability that the third point is in one of the "hit" clusters is precisely $\frac{H}{U+H} \leq \frac{H}{U}$. And in that case, we put in $\frac{1}{3}$rd of the total sum of distances of points *to the three centers* in the uncovered last three clusters into the kitty. But this sum is at most U as well. Therefore, in expectation, the total amount of stuff we put in the kitty is $\leq \frac{H}{U} \cdot \frac{U}{3} \leq \frac{H}{3}$.

Now, in the fourth iteration, we pick a point in the third cluster. Therefore, the third cluster is now hit, and is now happy. We don't need to pay any penalty for it. At this point, we *move* the penalty that was charged to it (denoted by the green zone), and equally divide it among the remaining uncovered clusters. In the case all the cluster "sizes" are the same, this doesn't change the total penalty. However, in the case of unequal "sizes" as well, because of our sampling process, this can only "help". This will be made formal soon, but for intuition, keep the equal sized balls picture in find. And in that case, the two equal balls now pay $1/2$ each.

Finally, in the last iteration there is another "miss". This time around, we smear $1/2$-the mass to the two remaining balls. And now, we see that we pay for them both, which indeed we should since the game is over. Again, the probability of the miss occurring is $\leq \frac{H}{U}$, where of course the values of H and U have changed. And the expected cost we pay is $\frac{U}{2}$ this time. Thus, in expectation in this round, we pay $\frac{U}{2}$.

Perhaps the reader can see where the "$\ln k$" is coming from. If there are k iterations, in each iteration there is a hit or a miss. The hits are good events, and the final hits will be bounded by Lemma 2 to $O(\mathsf{opt})$. The misses are bad events and generate penalty. But the expected penalty it generates in any round t is $\leq \frac{H}{k-t}$, where H is the expected cost of the "hit" clusters, which is $O(\mathsf{opt})$ by Lemma 2. Summing over all iterations gives us that the total penalties is $O(H_k) \cdot \mathsf{opt}$ where H_k is the harmonic sum.

- **Detailed Analysis.** The rigorous analysis goes through via a "potential" argument which formalizes the notion of penalty. Let us start with some definitions. For iteration $1 \leq t \leq k$, we use C_t to denote the centers picked by the k-means++ algorithm at the end of the iteration. So, C_k is the final output of the algorithm. We fix the optimum clustering $\Pi^* = (P*_1, \ldots, P^*_k)$. We let $\mathsf{opt} := \mathsf{opt}(\Pi^*) := \sum_{i=1}^{k} \mathsf{opt}(P^*_i)$ denote the cost of the optimum clustering. We wish to prove that $\mathbf{Exp}[\mathsf{cost}(P; C_k)] \leq O(\ln k)\mathsf{opt}$.

  At the end of iteration t, we let $\mathcal{H}_t := \{P^*_i \; : \; P^*_i \cap C_t \neq \emptyset\}$ denote the optimal clusters which have been "hit" so far by the algorithm. We let $\mathcal{U}_t$ denote the remaining clusters which are "uncovered". So, $|\mathcal{H}_t| + |\mathcal{U}_t| = k$. We also have, from Lemma 2, that

$$\mathsf{cost}(\mathcal{H}_t; C_t) := \sum_{P_i \in \mathcal{H}_t} \mathbf{Exp}[\mathsf{cost}(P_i; C_t)] \leq 16 \sum_{P_i \in \mathcal{H}_t} \mathsf{opt}(P_i) \leq 16\mathsf{opt} \tag{4}$$

  The main non-triviality in the analysis is hand $\mathsf{cost}(\mathcal{U}_t; C_t) := \sum_{P_i \in \mathcal{U}_t} \mathbf{Exp}[\mathsf{cost}(P_i; C_t)]$. To this end, we define the following ***potential function***. Let miss_t denote the number of "misses" we have had so far. That is, the number of times C_t picks more than one element from the same P_i^*. Note that $\mathsf{miss}_t = t - |\mathcal{H}_t|$. We define the potential function as

$$\Phi(t) := \mathsf{miss}_t \cdot \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|} \tag{5}$$

It is instructive to connect with the intuitive picture from the previous bullet point. Every time we made a mistake, we paid a penalty which was the total cost of the so-far uncovered clusters smeared equally over. The above potential is capturing that idea. Observe two things

$$\Phi(0) = 0, \text{ because } \mathsf{miss}_0 = 0 \quad \text{and} \quad \Phi(k) = \mathsf{cost}(\mathcal{U}_k; C_k), \text{ because } \mathsf{miss}_k = |\mathcal{U}_k|$$

We wish to prove that $\mathbf{Exp}[\Phi(k)]$ is small. To show this, we will show that the *increase* in every iteration is bounded in expectation.

- Let us consider the situation at the end of iteration t. In what follows, we will be implicitly conditioning all the random events that have occurred so far, and the expectation will be over the randomness in the $(t+1)$th iteration. We will upper bound $\mathbf{Exp}[\Phi(t+1) - \Phi(t)]$. To do so, we condition on two disjoint events: $\mathcal{E}_{\mathsf{hit}}$, which occurs if c_{t+1} lies in a cluster in $\mathcal{U}_t$, that is, c_{t+1} lies in an hitherto uncovered cluster, and $\mathcal{E}_{\mathsf{miss}}$, which occurs if c_{t+1} lies in a cluster in $\mathcal{H}_t$.

  First, we observe that

$$\mathbf{Pr}[\mathcal{E}_{\mathsf{miss}}] = \frac{\mathsf{cost}(\mathcal{H}_t; C_t)}{\mathsf{cost}(\mathcal{H}_t; C_t) + \mathsf{cost}(\mathcal{U}_t; C_t)} \le \frac{\mathsf{cost}(\mathcal{H}_t; C_t)}{\mathsf{cost}(\mathcal{U}_t; C_t)} \tag{6}$$

  The following two lemmas establish upper bounds on the increase in the potential conditioned on each event. We first complete the proof of Theorem 1 using these lemmas, and then we go on to prove them.

  Lemma 3. $\mathbf{Exp}[\Phi(t+1) - \Phi(t) \mid \mathcal{E}_{\mathsf{hit}}] \le 0$.

  Going back to the intuitive picture, the above lemma corresponds to the case when we pick a center which already had some "penalty" on it. This happened in the fourth row in the illustration above. At that time, we redistributed the penalty on the remaining uncovered clusters. The above lemma says that this redistribution can be done without increasing the total penalty.

  Lemma 4. $\mathbf{Exp}[\Phi(t+1) - \Phi(t) \mid \mathcal{E}_{\mathsf{miss}}] \le \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|}$.

  Going back to the intuitive picture, the above lemma corresponds to the case that when we miss, the expected increase in the potential can be "charged" to the cost of connecting the "hit" clusters divided by the total number of uncovered clusters left.

- **Proof of Theorem 1.** The final cost $\mathbf{Exp}[\mathsf{cost}(P; C_k)]$ can be partitioned as

$$\mathbf{Exp}[\mathsf{cost}(P; C_k)] = \mathbf{Exp}[\mathsf{cost}(\mathcal{H}_k; C_k)] + \mathbf{Exp}[\mathsf{cost}(\mathcal{U}_k; C_k)]$$

From (4), we know that $\mathbf{Exp}[\mathsf{cost}(\mathcal{H}_k; C_k)] \leq 16\mathsf{opt}$. We also know that $\mathsf{cost}(\mathcal{U}_k; C_k) = \Phi(k)$. From Lemma 3 and Lemma 4, we get that for any t,

$$\begin{aligned}
\mathbf{Exp}[\Phi(t+1) - \Phi(t)] &= \mathbf{Pr}[\mathcal{E}_{\mathsf{hit}}] \cdot \mathbf{Exp}[\Phi(t+1) - \Phi(t) \mid \mathcal{E}_{\mathsf{hit}}] + \mathbf{Pr}[\mathcal{E}_{\mathsf{miss}}] \cdot \mathbf{Exp}[\Phi(t+1) - \Phi(t) \mid \mathcal{E}_{\mathsf{miss}}] \\
&\leq \frac{\mathsf{cost}(\mathcal{H}_t; C_t)}{\mathsf{cost}(\mathcal{U}_t; C_t)} \cdot \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|} \\
&= \frac{\mathsf{cost}(\mathcal{H}_t; C_t)}{|\mathcal{U}_t|} \quad \underbrace{\leq}_{(4),\ |\mathcal{U}_t| = k - |\mathcal{H}_t| \geq k - t} \quad \frac{16\mathsf{opt}}{k-t}
\end{aligned} \tag{7}$$

Therefore,

$$\mathbf{Exp}[\Phi(k)] = \underbrace{\mathbf{Exp}[\Phi(0)]}_{=0} + \sum_{t=0}^{k-1} \mathbf{Exp}[\Phi(t+1) - \Phi(t)] \leq 16\mathsf{opt} \cdot H_k$$

Therefore, we get $\mathbf{Exp}[\mathsf{cost}(P; C_k)] \leq 16\mathsf{opt}(1 + H_k)$. □

- **Proof of Lemma 3.** When a hit occurs, the parameters change as follows: $\mathsf{miss}_{t+1} = \mathsf{miss}_t$ and $|\mathcal{U}_{t+1}| = |\mathcal{U}_t| - 1$. Therefore,

$$\begin{aligned}
\Phi(t+1) - \Phi(t) &= \mathsf{miss}_{t+1} \cdot \frac{\mathsf{cost}(\mathcal{U}_{t+1}; C_{t+1})}{|\mathcal{U}_{t+1}|} - \mathsf{miss}_t \cdot \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|} \\
&= \mathsf{miss}_t \cdot \left(\frac{\mathsf{cost}(\mathcal{U}_{t+1}; C_{t+1})}{|\mathcal{U}_t| - 1} - \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|} \right)
\end{aligned}$$

The lemma now follows from the following claim

$$\mathbf{Exp}[\mathsf{cost}(\mathcal{U}_{t+1}; C_{t+1}) \mid \mathcal{U}_t, C_t] \leq \left(1 - \frac{1}{|\mathcal{U}_t|}\right) \cdot \mathsf{cost}(\mathcal{U}_t; C_t) \tag{8}$$

Let us now upper bound the expected "drop" in the cost when we open an uncovered cluster. Let $\mathcal{U}_t := \{Q_1, Q_2, \ldots, Q_r\}$. Conditioned on $\mathcal{E}_{\mathsf{hit}}$, the probability that Q_j is "hit" is precisely $\frac{\mathsf{cost}(Q_j; C_t)}{\mathsf{cost}(\mathcal{U}_t; C_t)}$. Therefore,

$$\mathbf{Exp}[\mathsf{cost}(\mathcal{U}_t; C_t) - \mathsf{cost}(\mathcal{U}_{t+1}; C_t) \mid \mathcal{U}_t, C_t] = \sum_{j=1}^{r} \frac{\mathsf{cost}(Q_j; C_t)}{\mathsf{cost}(\mathcal{U}_t; C_t)} \cdot \mathsf{cost}(Q_j; C_t) \geq \frac{1}{|\mathcal{U}_t|} \cdot \mathsf{cost}(\mathcal{U}_t; C_t)$$

where the inequality follows by the Cauchy-Schwarz/Jensen/however-you-think-of-it inequality: for any non-negative x_j's, $\sum_{j=1}^{r} x_j^2 \geq \frac{1}{r} \left(\sum_{j=1}^{r} x_j \right)^2$. Rearranging, we get

$$\mathbf{Exp}[\mathsf{cost}(\mathcal{U}_{t+1}; C_{t+1}) \mid \mathcal{U}_t, C_t] \leq \mathbf{Exp}[\mathsf{cost}(\mathcal{U}_{t+1}; C_t)] \leq \left(1 - \frac{1}{|\mathcal{U}_t|}\right) \cdot \mathsf{cost}(\mathcal{U}_t; C_t)$$

which establishes (8), and proves the lemma.

- **Proof of Lemma 4.** When a miss occurs, the parameters change as follows: $\mathsf{miss}_{t+1} = \mathsf{miss}_t + 1$ and $|\mathcal{U}_{t+1}| = |\mathcal{U}_t|$. Therefore,

  $$\begin{aligned}
  \Phi(t+1) - \Phi(t) &= \mathsf{miss}_{t+1} \cdot \frac{\mathsf{cost}(\mathcal{U}_{t+1}; C_{t+1})}{|\mathcal{U}_{t+1}|} - \mathsf{miss}_t \cdot \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|} \\
  &\leq (\mathsf{miss}_t + 1) \cdot \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|} - \mathsf{miss}_t \cdot \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|} \\
  &= \frac{\mathsf{cost}(\mathcal{U}_t; C_t)}{|\mathcal{U}_t|}
  \end{aligned}$$

  where the inequality follows since opening more centers can only decrease cost.