ONLINE CLUSTERING WITH STOCHASTIC ARRIVALS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online k-clustering under strict consistency constraints: points are irrevocably assigned to a cluster and there is zero recourse (centers cannot be moved or closed). While achieving bounded costs under these constraints is impossible in the worst case without significantly relaxing the problem, we introduce the first framework for *online clustering with stochastic arrivals* to circumvent these adversarial lower bounds. By making mild stochastic modeling assumptions, we provide novel algorithms for the k-center and k-median objectives that achieve $O(1)$-approximations. Our algorithms strictly adhere to zero recourse and irrevocable assignments by utilizing a parsimonious amount of resource augmentation, opening at most $(1+\epsilon)k$ centers in expectation for any constant $\epsilon > 0$. Moreover, our algorithms have two additional properties that are extremely important in practice. First, they need only minimal information about the distribution: the expected cost of the optimal solution. Second, our algorithms are robust to model mis-specification — the performance of our algorithms smoothly degrades as our estimate of this expectation differs from the true expectation. Put together, these two properties mean that we need only to assume a weak estimate of one particular statistic about the distribution, rather than assuming full knowledge. We validate our theoretical results with experiments demonstrating the superior performance of our algorithms when distributional information is available.

1 INTRODUCTION

Clustering is fundamental in unsupervised learning, with k-center and k-median as standard objectives. In many modern large-scale applications—real-time network traffic monitoring, user clickstream analysis—data points arrive sequentially and must be processed *online*, requiring an algorithm to dynamically maintain a clustering under constraints. A common such set are *consistency* constraints Moseley et al. (2024): point assignments are *irrevocable* and centers cannot be moved or closed (*zero recourse*) Liberty et al. (2016).

Such stability is critical in downstream operational decisions, where frequent relabeling incurs high overhead and erodes interpretability Lattanzi & Vassilvitskii (2017); Liberty et al. (2016). In large-scale ML systems, for instance, cluster IDs augment feature vectors for downstream models, so altering historical assignments changes the entire feature set and necessitates costly retraining.

Achieving high-quality clustering under strict consistency constraints is impossible in the worst-case (adversarial) setting: no online algorithm achieves a competitive ratio bounded by any function of n and k for k-median Liberty et al. (2016), and these results extend easily to k-center. To circumvent these lower bounds, prior literature has primarily relied on three approaches Liberty et al. (2016); Lattanzi & Vassilvitskii (2017); Fichtenberger et al. (2021); Guo et al. (2021); Moseley et al. (2024):

- **Recourse:** Algorithms that maintain a low cost by frequently re-optimizing centers or relabeling points. State-of-the-art results require $O(k \cdot \text{poly}\log(n\Delta))$ center changes. These results of course violate the no-recourse constraint (Fichtenberger et al., 2021; Guo et al., 2021).
- **Large Resource Augmentation:** Approaches that achieve bounded costs open a number of clusters that far exceeds the target k, often scaling to $O(k \log n \log n\Delta)$ centers (Liberty et al., 2016).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **High Cost or Parameter Dependence:** Algorithms that strictly maintain k centers but suffer from competitive ratios that scale with the aspect ratio Δ or are exponential in k (Moseley et al., 2024).

In this paper, we move beyond worst-case analysis to investigate **stochastic online clustering**, modeling the data stream as independent draws from an unknown but fixed distribution $\mathcal{D}$ (our results extend beyond iid, as discussed in Section 1.1). This naturally captures domains where data is generated by underlying physical processes exhibiting statistical regularity.

Distributional inputs have been considered in the offline setting for k-clustering (Cormode & McGregor, 2008; Guha & Munagala, 2009), but little is known when we move to the online setting. By adopting this stochastic model, we aim to answer three fundamental theoretical questions:

1. **Circumventing Lower Bounds:** To what extent can the statistical regularity of a distribution be used to bypass worst-case lower bounds for online k-clustering?
2. **Distributional Knowledge:** What level of information about the distribution $\mathcal{D}$ is required? Can an algorithm operate effectively with minimal information (e.g., first and second moments), or is full knowledge of $\mathcal{D}$ necessary?
3. **Robustness to Incorrect Information:** What if the information we use about $\mathcal{D}$ is incorrect? Can we design algorithms with performance that degrades slowly and smoothly with the error of our information?

1.1 Our Contributions

We consider online stochastic k-clustering achieving an $O(1)$-approximation for both k-center and k-median in $O(nk)$ time. Crucially, our algorithms require only the expected optimal cost μ from $\mathcal{D}$, bypassing full distributional access, while strictly enforcing **zero recourse** and **irrevocable assignments**. To circumvent known lower bounds, we employ a parsimonious amount of resource augmentation, opening at most $(1+\epsilon)k$ centers in expectation for any $\epsilon > 0$. Moreover, given an estimate $\widehat{\mu}$ of μ, our algorithms' performance degrade slowly and smoothly as $\widehat{\mu}$ deviates from μ.

k-Center. We define k-center formally in Section 2, but informally it is the problem of choosing k centers so as to minimize the *maximum* distance of any point to its closest center. For this problem we show the following result:

Theorem 1 (Informal k-center)**.** *For every $0 < \epsilon \leq 1/2$, there exists an online algorithm which, when given a prediction $\widehat{\mu}$ of μ, opens at most $\left(1 + \frac{\mu}{\widehat{\mu}}\left(\frac{\epsilon}{1-\epsilon}\right)\right) k$ centers in expectation and has expected cost at most $\left(\frac{1}{\epsilon^2}(1 + \frac{\widehat{\mu}}{\mu})\right) \mu$ for stochastic online k-center.*

We prove this formally in Section 3 as Theorem 3. For k-center we can go significantly beyond iid: the point sequence need only be generated from *some* distribution, even with arbitrary correlations. Informally, if we knew the optimal value R^* for our particular input X, we could greedily cluster online with radius $2R^*$, yielding a 2-approximation with only k centers. The difficulty is that we do not know R^*: we only have an estimate of the *expected* optimal value. We show that by combining this idea with guess-and-double techniques and carefully exploiting the stochasticity of the input, we can nonetheless obtain Theorem 1.

k-Median. We define k-median formally in Section 2. Informally, it has the same setup as k-center, but the objective is to minimize the *sum* of every point's distance to its closest center (rather than the max). We obtain a result very similar to Theorem 1, but with a completely different algorithm:

Theorem 2 (Informal k-median)**.** *For every $0 < \epsilon \leq 1/2$, there exists an online algorithm for stochastic online k-median which, when given a prediction $\widehat{\mu}$ of μ, opens at most $\left(1 + O(\epsilon)\frac{\mu}{\widehat{\mu}} + O(\epsilon)\right) k$ centers in expectation and has expected cost at most $\left(\frac{1}{\epsilon^2}\left((1 + O(\epsilon))\frac{\widehat{\mu}}{\mu} + O(\epsilon)\right)\right) \mu$.*

We again can go beyond iid, but here we are slightly more limited: we require that the distribution is *exchangeable* (Kallenberg, 2002). Informally, in exchangeable distributions the distribution over the pointset can be arbitrary, but once a pointset is chosen its *ordering* is uniform. This is a particularly nice class for us because our algorithm for this problem is essentially a variant of Meyerson's algorithm for online facility location under random order arrivals Meyerson (2001). Note that online

facility location is a significantly easier problem than our settings, as it is essentially the Lagrangian relaxation of k-median: rather than a hard constraint on the number of clusters as in our problems, the number of clusters is essentially just added to the connection costs in the objective. This allows for a simple algorithm which, upon the arrival of point i, opens it as a center with probability d_i/f_i (where d_i is its current distance to the closest center and f_i is the facility opening cost). Our algorithm has to work in a setting with facility opening costs, but intuitively if we set f_i to μ/k for all points i then we might "expect" to open approximately k centers. This isn't quite true, though, and moreover we do not actually know μ. So our actual algorithm sets f_i to $\frac{\widehat{\mu}}{\epsilon^2 k}$.

Discussion. To get some intuition for the bounds of Theorems 1 and 2, consider what happens if there is significant distribution shift: $\mu/\widehat{\mu} = 2$ or $\widehat{\mu}/\mu = 2$. In other words, our prediction for the expected optimal cost is either *half* or *double* the true expected optimal cost. In this case we still only use $(1 + O(\epsilon))k$ centers, and get expected cost of at most $O\left(\mu/\epsilon^2\right)$. So for constant ϵ we get $O(1)$-competitive ratio with extremely mild resource augmentation. In other words, if our prediction is incorrect by a constant factor (in either direction), then we still use only $(1 + O(\epsilon))k$ centers and get competitive ratio $O(1/\epsilon^2)$.

The combination of this robustness with the minimal information needed by our algorithms is particularly powerful, since it means that our algorithms are quite robust to distribution shift. If we have an estimate $\widehat{\mu}$ due to some historical data, and then the distribution shifts to have actual expected optimal cost μ, our algorithm only pays based on the (multiplicative) gap between μ and $\widehat{\mu}$. This could be very small even if the distribution shifts quite dramatically, and is obviously upper bounded by natural notions of distribution distance like the Wasserstein-1 (Earth Mover's) distance.

Finally, we discuss the resource augmentation used by our algorithm. As discussed above, algorithms that open exactly k centers must incur a high cost for worst-case arrivals. A simple construction shows that such a lower bound holds even in the iid setting; see Appendix B for details. This shows that our cost guarantee cannot be achieved without resource augmentation, even in the iid setting.

1.2 Related Work

Consistent Clustering. The closest line of prior work studies online clustering under *consistency* constraints: irrevocable point assignments and zero recourse on centers. A foundational lower bound of Liberty et al. (2016) shows that no online algorithm can achieve a competitive ratio bounded by any function of n and k under these constraints, for either k-median or k-means, motivating beyond-worst-case approaches, of which three main directions have been explored.

Large resource augmentation allows algorithms to open more than k centers. Liberty et al. (2016) give a randomized $O(1)$-competitive algorithm opening $O(k \log n \log(W^*/w^*))$ centers in expectation (where W^* and w^* are the optimal cost and a known lower bound), and $O(\log n)$-competitive with $O(k \log(n) \log(\gamma n))$ centers when n and w^* are unknown, where γ is the aspect ratio. Both are $\Omega(k \log n)$ centers, substantially larger than the $(1 + \epsilon)k$ our approach requires.

Recourse-based algorithms relax zero recourse by permitting a small number of center changes. Lattanzi & Vassilvitskii (2017) gave the first such result for k-median and k-means, maintaining $O(1)$-competitiveness with $O(k^2 \log^4(n\gamma))$ expected center changes, where γ is the aspect ratio of the metric. Fichtenberger et al. (2021) subsequently improved this to $O(k \operatorname{polylog}(n\gamma))$ center changes while maintaining $O(1)$-competitiveness for general metrics. Guo et al. (2021) extended these results to k-median with outliers. All of these algorithms violate our zero-recourse constraint

Moseley et al. (2024) show that given a budget B upper-bounding the optimal cost, they give a deterministic algorithm achieving a competitive ratio of $O(k^5 \cdot 3^k)$ relative to B, and show that any algorithm must pay $\Omega(k)$. This is the only result using no reassignments and no more than k centers.

No-Substitution Clustering. A related but distinct online model is *no-substitution* clustering, in which centers must be chosen from the stream at the moment a point is first seen, but point-to-center assignments may be made in hindsight, deferring assignment decisions and substantially simplifying the problem relative to our setting. Under adversarial orders, Bhattacharjee & Moshkovitz (2021) give an $O(k^3)$-competitive algorithm opening $O(dk \log(n) \log(k))$ centers, where $d = O(k^2 \log n)$ for mixtures of k Gaussians. In the random-order setting, Moshkovitz (2021) shows $O(\log n)$ (resp. $O(1)$) centers suffice for $O(1)$-competitiveness when n is unknown (resp. known), for k-means with

$k = O(1)$. For k-median, Hess et al. (2021) achieve $O(1)$-competitiveness with $O(k \log^2 k)$ centers when n is known, degrading to $O(k \log^2(k) \log(n))$ via a doubling trick when unknown.

Online Learning and Regret. A separate thread minimizes regret relative to the best fixed set of k centers in hindsight, maintaining k centers C_t at each round with cost $d(x_t, C_t)^2$ (for k-means) and no constraint on how centers evolve. Cohen-Addad et al. (2021) and Bhattacharjee et al. (2023) give algorithms with sublinear regret, but since centers may change arbitrarily each round, this setting is incomparable to our zero-recourse, irrevocable-assignment constraints.

Algorithms with Predictions. Our setup and results are related to the recent framework of algorithms with predictions (Mitzenmacher & Vassilvitskii, 2022). In this framework we are given some prediction, or advice, to help solve whatever problem we are facing; we normally think of this advice as being given by some machine learning system. Then the goal is to do well on the problem if the prediction is accurate while still obtaining good worst-case performance if the prediction is inaccurate. Clearly, we can think of the estimate $\widehat{\mu}$ that we are given of μ as a prediction in this framework. The most closely related results to ours in this setting are results on online facility location (e.g., (Almanza et al., 2021; Jiang et al., 2022; Fotakis et al., 2025)), (offline) clustering (e.g., (Ergun et al., 2022; Nguyen et al., 2023; Huang et al., 2025)), and distributional predictions (e.g., (Dinitz et al., 2024)). Theorems 1 and 2 give the following guarantees in the algorithms-with-predictions framework. Under perfect predictions ($\widehat{\mu} = \mu$), both algorithms achieve expected cost $O(\mu/\epsilon^2)$ using at most $(1 + O(\epsilon))k$ centers in expectation (*consistency*). Arbitrarily severe underestimation preserves the $O(1/\epsilon^2)$ cost ratio, while arbitrarily severe overestimation preserves the $(1 + O(\epsilon))k$ expected center bound (*one-sided robustness*).

2 Model and Preliminaries

We now formalize our model for online k-clustering with stochastic arrivals. Let $(\mathcal{X}, d)$ be a finite metric space on N points. An instance consists of a sequence $X = (X_1, \ldots, X_n)$ of n points from $\mathcal{X}$ which are revealed one at a time to the algorithm. The algorithm must maintain a set of centers $C \subseteq X$ and an assignment $\rho : [n] \to C$ online. Initially, $C = \emptyset$ and in each step $i \in [n]$ the algorithm faces the following irrevocable decision: either add the i'th arrival X_i to C (open it as a center) or assign X_i to some previously opened center $\rho(i) \in C$. Naturally, each center is assigned to itself in ρ (so that we have $d(X_i, \rho(i)) = 0$ for each $X_i \in C$). The algorithm is constrained to use zero recourse (cluster centers cannot be moved or closed) and assignments are irrevocable. We will refer to a pair (C, ρ) as a solution to the clustering problem; the resulting clusters are given by $S_c = \{X_i : \rho(i) = c\}$ for each $c \in C$.

Departing from prior work, we focus on inputs which are generated stochastically. We consider the sequence X to be sampled from some distribution $\mathcal{D}$ over length n sequences from $\mathcal{X}$ so that each $X_i, i \in [n]$ is a random variable. In the most general case, this distribution could involve arbitrary correlations between the various X_i. In the simplest case, we have that $\mathcal{D} = D^n$ for some distribution D over $\mathcal{X}$, i.e., X is an independent and identically distributed (iid) sequence. Another case of interest are exchangeable distributions. Given a sequence $X = (X_1, \ldots, X_n)$ and a permutation $\pi : [n] \to [n]$, we let $X^\pi = (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}$ be the permuted sequence. A distribution $\mathcal{D}$ is *exchangeable* if $\mathcal{D}(X) = \mathcal{D}(X^\pi)$ for all permutations π. In addition to stochastic generation, we will consider settings where the algorithm is given additional information about the distribution. In the weakest case, this could be a (potentially noisy) estimate of a parameter of the distribution $\mathcal{D}$, and in the strongest case this could be a predicted distribution $\widehat{\mathcal{D}}$.

Ideally, the algorithm opens at most k centers. As discussed in the introduction, achieving non-trivial guarantees under this constraint is impossible. Thus we allow the algorithm to open more than k centers, which is known as resource augmentation, and our goal is to keep this as close to k as possible. Since the input is generated stochastically and the algorithm itself may utilize randomness, we remark that the set of centers C (and hence $|C|$) is a random variable.

The final set of centers and assignment (C, ρ) are evaluated by a cost function $\mathrm{Cost}(X, C, \rho)$. In general we use $\mathrm{OPT}(X) := \min_{C:|C|\leq k,\rho} \mathrm{Cost}(X, C, \rho)$ to denote the hindsight optimal cost whenever the underlying cost function is clear. We consider the standard k-clustering objectives:

- k-center: $\mathrm{Cost}(X, C, \rho) = \max_{i\in[n]} d(X_i, \rho(i))$
- k-median: $\mathrm{Cost}(X, C, \rho) = \sum_{i=1}^{n} d(X_i, \rho(i))$

We say that an online algorithm which results in centers C and assignment ρ is (α, β)-competitive for a particular cost function if we have

$$\mathbb{E}[\text{Cost}(X, C, \rho)] \leq \alpha\mu \quad \text{and} \quad \mathbb{E}[|C|] \leq \beta k,$$

where $\mu = \mathbb{E}[\text{OPT}(X)]$ and all expectations are taken with respect to the random instance $X \sim \mathcal{D}$ and any internal randomness used by the algorithm. $\alpha \geq 1$ measures our competitiveness with the offline optimal cost in expectation, while $\beta \geq 1$ measures the amount of resource augmentation.

Finally, we will use $\mathbf{1}\{A\}$ to denote the indicator function for an event A.

3 Online k-Center Under Stochastic Arrivals

We now present our algorithm for online k-center with stochastic inputs. All omitted proofs from this section can be found in Appendix C. Our algorithm requires only minimal assumptions and information on the distribution $\mathcal{D}$. In fact, our analysis holds for *any* distribution $\mathcal{D}$ over $\mathcal{X}^n$, and our algorithm only requires an estimate $\widehat{\mu}$ of the expected optimal cost $\mu = \mathbb{E}[\text{OPT}(X)]$. The following theorem formally states the guarantees of our algorithm.

Theorem 3. *For each $\epsilon \in (0, 1/2]$, there exists an online algorithm which, when given a prediction $\widehat{\mu}$ of μ, is (α, β)-competitive for $\alpha = \frac{1}{\epsilon^2}(1 + \frac{\widehat{\mu}}{\mu})$ and $\beta = 1 + \frac{\mu}{\widehat{\mu}}(\frac{\epsilon}{1-\epsilon})$.*

Algorithm 1: Online k-Center Algorithm

Data: Online sequence $X \sim \mathcal{D}$, predicted parameter $\widehat{\mu}$, number of centers k, $\epsilon \in (0, 1/2]$,
Result: Solution (C, ρ) with $\mathbb{E}[\text{Cost}(X, C, \rho)] \leq \frac{1}{\epsilon^2}(1 + \frac{\widehat{\mu}}{\mu}) \cdot \mu$ and $\mathbb{E}[|C|] \leq (1 + \frac{\mu}{\widehat{\mu}}(\frac{\epsilon}{1-\epsilon}))k$

Procedure `OnlineKCenter(`X, $\widehat{\mu}$, k, ϵ`)`**:**
- $C \leftarrow \{X_1\}, \rho(1) \leftarrow X_1$ `// Initialize the solution`
- $s \leftarrow 1$ `// Centers counter`
- $R \leftarrow (\frac{1}{\epsilon})^2\widehat{\mu}$ `// Initial cluster radius`
- **for** $i = 2, 3, \ldots, n$ **do**
  - $c \leftarrow \arg\min_{c' \in C} d(X_i, c')$
  - **if** $d(X_i, c) \leq R$ **then**
    - $\rho(i) \leftarrow c$ `// Assign` X_i `to` c `if it is within distance` R
  - **else**
    - $C \leftarrow C \cup \{X_i\}$ `// Open a center at` X_i
    - $\rho(i) \leftarrow X_i$
    - $s \leftarrow s + 1$
  - **if** $s = k$ **then**
    - $R \leftarrow \frac{1}{\epsilon}R$ `// Increase the radius`
    - $s \leftarrow 0$
- **return** C, ρ

Our algorithm is formally described in Algorithm 1. The idea is as follows. If we knew $\text{OPT}(X)$, we could simulate standard offline approximation algorithms (Gonzalez, 1985; Hochbaum & Shmoys, 1985) by cutting out clusters of radius $R = O(\text{OPT}(X))$ online. That is, we assign arrival X_i to any previously open center that is within distance R of it. If no such center exists then we open X_i as a center. When $R \geq 2\text{OPT}(X)$, it is impossible for this algorithm to open more than k centers (as otherwise we would have $k+1$ points all at a distance $> 2\text{OPT}(X)$ from each other, a contradiction).

Since we do not know $\text{OPT}(X)$ a natural approach is to apply a guess-and-double scheme: each time the algorithm opens k centers we double our guess of R and continue. The problem with doing this naively is that we could open much more than $(1 + \epsilon)k$ centers. Here we leverage the stochastic nature of the input to give a stronger guarantee. First, we use the predicted parameter $\widehat{\mu}$ to give a better starting point for the radius R. Second, we increase R by a factor of $1/\epsilon$ to control the number of centers; this ensures that the probability of opening more than k centers is small and decreases quickly as we increase R.

We will say that Algorithm 1 has reached phase ℓ when it sets $R = (\frac{1}{\epsilon})^{2+\ell}\widehat{\mu}$. To start the analysis, we analyze the quality of the solution produced by our algorithm.

Lemma 4. *Let C, ρ be the solution produced by Algorithm 1 when run on input sequence X. We have* $\mathrm{Cost}(X, C, \rho) \leq \frac{1}{\epsilon^2} \max\left(\widehat{\mu}, \mathrm{OPT}(X)\right)$

Lemma 5. *Let C be the set of centers produced by Algorithm 1 when run on input sequence $X \sim \mathcal{D}$. We have $\mathbb{E}[|C|] \leq (1 + \frac{\mu}{\widehat{\mu}}(\frac{\epsilon}{1-\epsilon}))k$.*

Proof of Theorem 3. The bound on β follows directly from Lemma 5. For the bound on α, we compute the expected cost of Algorithm 1 on an input sequence $X \sim \mathcal{D}$ as follows:

$$\mathbb{E}[\mathrm{Cost}(X, C, \rho)] \leq \frac{1}{\epsilon^2}\mathbb{E}[\max(\widehat{\mu}, \mathrm{OPT}(X))] \leq \frac{1}{\epsilon^2}\mathbb{E}[\widehat{\mu} + \mathrm{OPT}(X)] = \frac{1}{\epsilon^2}\left(\widehat{\mu} + \mu\right).$$

The first inequality follows from Lemma 4 and the second inequality follows since $\max(a, b) \leq a + b$ for all $a, b \in \mathbb{R}$. Rearranging the right hand side yields the expression for α, completing the proof. □

4 Online k-Median Under Stochastic Arrivals

This section gives our results for the k-median objective. All omitted proofs from this section can be found in Appendix D. Again, our algorithm only requires an estimate $\widehat{\mu}$ of the expected optimal cost $\mu = \mathbb{E}[\mathrm{OPT}(X)]$. Our analysis works not just for i.i.d. sequences, but for the more general class of exchangeable distributions. The following theorem states the formal result we wish to prove.

Theorem 6. *For each $\epsilon \in (0, 1/2]$ there is an online algorithm which, when given a prediction $\widehat{\mu}$ of μ, is (α, β)-competitive for $\alpha = \frac{1}{\epsilon^2}\left((1 + O(\epsilon))\frac{\widehat{\mu}}{\mu} + O(\epsilon)\right)$ and $\beta = 1 + O(\epsilon)\frac{\mu}{\widehat{\mu}} + O(\epsilon)$ whenever $\mathcal{D}$ is an exchangeable distribution.*

Algorithm 2: Online k-Median Algorithm

Data: Online sequence $X \sim \mathcal{D}$, predicted parameter $\widehat{\mu}$, number of centers k, $\epsilon \in (0, 1/2]$
Result: Solution (C, ρ) which is (α, β)-competitive for (α, β) as in Theorem 6

Procedure `OnlineKMedian(`X, $\widehat{\mu}$ k, ϵ`)`**:**
- $C \leftarrow \{X_1\}, \rho(1) \leftarrow X_1$ `//Initialize the solution`
- **for** $i = 2, 3, \ldots, n$ **do**
  - $c \leftarrow \arg\min_{c' \in C} d(X_i, c')$ `//Find the closest center to` X_i
  - $p_i \leftarrow \min(1, \epsilon^2 k \cdot d(X_i, c)/\widehat{\mu})$ `//Probability of opening a new center`
  - $B_i \sim \mathrm{Bernoulli}(p_i)$
  - **if** $B_i = 1$ **then**
    - $C \leftarrow C \cup \{X_i\}$ `//If` $B_i = 1$`, open a center at` X_i
    - $\rho(i) = X_i$
  - **else**
    - $\rho(i) = c$ `//Otherwise assign` X_i `to the closest center` c
- **return** (C, ρ)

Our algorithm is a modification of Meyerson's algorithm for the online facility location problem (Meyerson, 2001). The key idea is that points which are expensive to assign, far from all existing centers, are likely to belong to a new cluster. Thus, we open a center there with probability proportional to their assignment cost. Doing this probabilistically helps ensure we do not open too many centers. Formally, we set the opening probability $p_i = \min\left(1, \epsilon^2 k D_i/\widehat{\mu}\right)$, treating $\widehat{\mu}/k$ as a per-cluster budget. The stochastic arrival order then controls the total center count: within each optimal cluster, far points tend to arrive after a nearby close point has already opened a center, limiting their contribution.

Our analysis follows a similar structure as Meyerson's original paper (Meyerson, 2001) and lecture notes by Roughgarden (Roughgarden, 2017), but adapted to our setting (note that the k-median objective is related to, but distinct from, the facility location objective). We start by conditioning on the (multi-)set of points P which show up in X, but not the exact ordering that they arrive in X. Note that conditioning on P fixes $\mathrm{OPT}(X)$ since $\mathrm{OPT}(X)$ is independent of the arrival order. Further, since $\mathcal{D}$ is exchangeable, we have that X is a random permutation of the points in S. For this reason, we focus on the case where X is a random permutation of the points in a fixed collection S. For simplicity of notation we will often write $\mathrm{OPT} := \mathrm{OPT}(P) = \mathrm{OPT}(X)$.

Let $\mathcal{F} = (\mathcal{F}_i)_{i=0}^n$ be the filtration capturing all information observed up to step i of the algorithm ($i = 0$ corresponds to no observed information). Note that p_i, the probability of opening a center

in each step, is an $\mathcal{F}_{i-1}$-measurable random variable. Proving Theorem 6 reduces to establishing a bound on $\mathbb{E}[\sum_{i=1}^n p_i]$. We state this bound as follows.

Lemma 7. *Fix a collection of points P and suppose that Algorithm 2 is run on the sequence X, where X is a random permutation of P. Then we have $\mathbb{E}[\sum_{i=1}^n p_i] \leq (1 + O(\epsilon) + O(\epsilon \mathrm{OPT}/\widehat{\mu}))k$.*

The majority of the analysis is focused on establishing Lemma 7. Before getting into this, let's see how this lemma implies Theorem 6.

Proof of Theorem 6. Let $X \sim \mathcal{D}$. For exchangeable distribution $\mathcal{D}$ and a collection of points P, we have that $X \mid P$ is random permutation of the points in P. Thus we first focus on the case where the set of points P is fixed and X is a random permutation of these points. Too keep notation concise, expectations will be taken over the random permutation of P and any internal randomness utilized by the algorithm unless otherwise stated.

The number of centers opened by Algorithm 2 is clearly given by $|C| = \sum_{i=1}^n B_i$. Thus we have

$$\mathbb{E}[|C|] = \sum_{i=1}^n \mathbb{E}[B_i] = \sum_{i=1}^n \mathbb{E}[\mathbb{E}[B_i \mid \mathcal{F}_{i-1}]] = \sum_{i=1}^n \mathbb{E}[p_i] \leq \left(1 + O(\epsilon) + O(\epsilon)\frac{\mathrm{OPT}}{\widehat{\mu}}\right) k,$$

where we used Lemma 7 to get the inequality. For the cost, let D_i be the random distance from the i'th arrival X_i to its closest center at the time of X_i's arrival. We either pay 0 cost if we open X_i as a center, or cost D_i if not. Thus, conditioned on the history prior to step i, we have

$$\mathbb{E}[d(X_i, \rho(i)) \mid \mathcal{F}_{i-1}] = (1 - p_i)D_i \leq p_i \frac{\widehat{\mu}}{\epsilon^2 k}.$$

This follows since either $p_i = 1$ (in which case the left hand side of the inequality is 0 and the right hand side is non-negative) or $p_i = \epsilon^2 k D_i/\widehat{\mu} < 1$ (in which case the right hand side equals $D_i \geq (1 - p_i)D_i$ since $p_i \in (0, 1)$). Summing over all steps yields

$$\mathbb{E}[\mathrm{Cost}(X, C, \rho)] = \sum_{i=1}^n \mathbb{E}[d(X_i, \rho(i))] = \sum_{i=1}^n \mathbb{E}[\mathbb{E}[d(X_i, \rho(i)) \mid \mathcal{F}_{i-1}]] \leq \frac{\widehat{\mu}}{\epsilon^2 k} \sum_{i=1}^n \mathbb{E}[p_i].$$

So substituting the bound on $\mathbb{E}[\sum_i p_i]$ from Lemma 7 yields

$$\mathbb{E}[\mathrm{Cost}(X, C, \rho)] \leq \frac{\widehat{\mu}}{\epsilon^2 k}\left(1 + O(\epsilon) + O(\epsilon)\frac{\mathrm{OPT}}{\widehat{\mu}}\right) k \leq \frac{\widehat{\mu}}{\epsilon^2}\left(1 + O(\epsilon) + O(\epsilon)\frac{\mathrm{OPT}}{\widehat{\mu}}\right).$$

Finally, taking the expectation over the collection of points P and rearranging yields

$$\mathbb{E}_{X\sim\mathcal{D}}[\mathrm{Cost}(X, C, \rho)] \leq \frac{\mu}{\epsilon^2}\left((1 + O(\epsilon))\frac{\widehat{\mu}}{\mu} + O(\epsilon)\right) \text{ and } \mathbb{E}_{X\sim\mathcal{D}}[|C|] \leq k\left(1 + O(\epsilon)\frac{\mu}{\widehat{\mu}} + O(\epsilon)\right),$$

completing the proof. □

4.1 Establishing Lemma 7

At this point, we may focus on bounding $\mathbb{E}[\sum_{i=1}^n p_i]$ for a fixed collection of arriving points P which are presented to Algorithm 2 in a random order. Since we have fixed P, we may also fix a particular optimal solution, (C^*, ρ^*). We focus on a particular optimal center $c^* \in C^*$ and let $S(c^*) := \{X_i : \rho^*(i) = c^*, i \in [n]\}$ be c^*'s cluster in the optimal solution. Note that the points in $S(c^*)$ still arrive to the algorithm in a uniformly random order. We may abuse notation and write $i \in S(c^*)$ as a shorthand for $X_i \in S(c^*)$ Thus our goal will be to bound $\mathbb{E}[\sum_{i\in S(c^*)} p_i]$. Summing over all optimal clusters will then yield Lemma 7.

The rough intuition for the analysis is as follows. Once a point that is close enough to c^* is chosen as a center, we may use the triangle inequality to bound subsequent p_i's in terms of distances incurred by OPT. Closeness will depend on the average cost paid by c^*'s cluster, which is given by:

$$\bar{d}(c^*) := \frac{1}{|S(c^*)|} \sum_{i\in S(c^*)} d(X_i, c^*)$$

To formalize this, we introduce the following definition.

Definition 8. *For a fixed optimal cluster $S(c^*)$, we say that a point $X_i \in S(c^*)$ is* close *if $d(X_i, c^*) \leq \bar{d}(c^*)/\epsilon$ and otherwise it is* far. *Let $S_{\mathrm{c}}(c^*), S_{\mathrm{f}}(c^*)$ denote the sets of close and far points in c^*'s cluster, respectively.*

A key observation entailed by this definition is that there are many close points (and therefore relatively few far points).

Proposition 9. *We have $|S_{\mathrm{c}}(c^*)| \geq (1-\epsilon)|S(c^*)|$ and $|S_{\mathrm{f}}(c^*)| \leq \epsilon|S(c^*)|$*

The next two lemmas bound the contributions of the close and the far points to the sum of the p_i's.

Lemma 10. *We have $\mathbb{E}[\sum_{i \in S_{\mathrm{c}}(c^*)} p_i] \leq 1 + \frac{2\epsilon k}{\widehat{\mu}} \sum_{i \in S(c^*)} d(X_i, c^*)$.*

Lemma 11. *We have $\mathbb{E}[\sum_{i \in S_{\mathrm{f}}(c^*)} p_i] \leq O(\epsilon) + \frac{O(\epsilon)k}{\widehat{\mu}} \sum_{i \in S(c^*)} d(X_i, c^*)$*

For the close points, we use the fact that once a close point has been opened as a center, the triangle inequality bounds the p_i's for subsequent close arrivals, giving the latter term in Lemma 10. A careful inductive argument shows that the expected contribution of the earlier p_i's is at most 1.

The idea for the far points is that they are unlikely to contribute much since they arrive in a random order and there are at most $\epsilon|S(c^*)|$ such far points. Far points that arrive before every close point could have a large p_i, but since there are few far points, this only contributes $O(\epsilon)$ in expectation. If a far point X_i arrives after some close point X_j, then we can bound p_i in terms of p_j and the distances paid by OPT for X_i and X_j by using the triangle inequality. This is because every center that was available for X_j is also available for the later arrival X_i. This allows us to use Lemma 10 to help charge the p_i's for far points to that of close points.

Lemma 7 easily follows from Lemmas 10 and 11, completing our analysis for the k-median objective.

5 Experiments

We describe our experimental methodology and empirical results, designed to answer four questions: (1) *Theory vs. practice*: does the algorithm outperform its theoretical guarantees? (2) *Cost-ϵ tradeoff*: how does cost increase as ϵ decreases? (3) *Calibration*: when ϵ is tuned so that our algorithm opens approximately k centers, how does cost compare to the offline optimum? (4) *ϵ-crossover consistency*: is the value of ϵ at which our algorithm opens closest to k centers consistent across datasets?

Evaluation protocol. For each dataset, a random 5% training sample is used to estimate OPT via `pyclustering` Novikov (2019). This estimate, scaled to the test-set size, is supplied to our algorithm. We vary ϵ and record clustering cost and center counts, comparing against baselines with $k = 10$. To evaluate sensitivity to arrival order, we generate 3 independent random permutations of the test set; for each, we run our randomized algorithm 3 times. We report the mean and one standard deviation across these 9 runs. Our implementation, together with the datasets used in the experiments, is included in the supplementary material.

Baselines. We compare against two baselines.

The first baseline is the fully *offline* k-median solution computed by the `pyclustering` with $k = 10$. This baseline has access to the entire test set in advance and computes a clustering before any online arrival process begins. The fully offline baseline has three substantial advantages over our algorithm. First, it has access to all points in advance. Second, it does not need to make center decisions or point assignments in an online manner. Third, it is not restricted to choosing centers from among the input points, and may instead place centers anywhere in the underlying space.

The second baseline, which we call the **Oracle-Assisted Baseline**, first computes the k-median clustering of the full test set using `pyclustering`, then receives the test points in a random permutation and sets as each cluster's center the first point from that cluster appearing in the permutation. Though it processes points online, it leverages information unavailable in truly online settings: implicit knowledge of the ground-truth clusters. By selecting one center per offline cluster, this *unachievable* baseline maintains exactly $k = 10$ centers. We include it to isolate the performance gap attributable strictly to online decision-making when underlying clusters are already known.

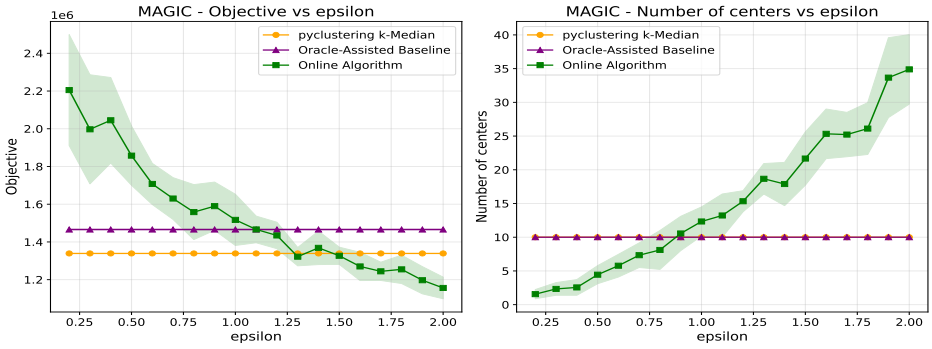


Figure 1: MAGIC dataset results: objective (left) and centers (right) vs. ϵ.

Datasets. We use four UCI benchmark datasets: Adult, MAGIC, Skin, and Power. We retain only numerical features and remove rows with missing values. Dataset sizes and features are shown in Table 1; additional details are in Appendix E.

Dataset	Points	Features	Crossover ϵ	Online/Offline	Online/Oracle
MAGIC	19,020	10	0.9	1.19	1.08
Adult	32,561	6	0.9	1.47	1.18
Skin	245,057	3	0.9	1.84	1.23
Power	2,049,280	7	1.1	1.35	1.05

Table 1: Crossover behavior across datasets. The crossover ϵ is the value in $\{0.2, 0.3, \ldots, 2\}$ for which the average number of centers opened by our algorithm is closest to $k = 10$. Ratios are reported at that value of ϵ.

Discussion. Figure 1 shows our results on the MAGIC dataset, and Table 1 summarizes the crossover behavior across all datasets. The plots for the remaining datasets are included in Appendix E.

Theory is Predictive of Empirical Performance. Our theoretical guarantee gives an expected cost of $O(1/\epsilon^2)$ times the optimal. The data suggests our theoretical bound, while correct, is highly conservative, and the algorithm performs substantially better in practice.

Cost-ϵ tradeoff. As shown in Figure 1 (left), cost decreases as ϵ increases. The cost reaches the offline benchmark for larger values of ϵ, and is already close to the offline benchmark near the crossover point ($\epsilon = 0.9$). This is notable given that our algorithm must make irrevocable assignment decisions without knowledge of future arrivals, while the offline baseline sees all points in advance.

Calibration. An informative operating point is the crossover value of ϵ, where our algorithm opens approximately the same number of centers as the baselines. For the MAGIC dataset, this occurs at $\epsilon = 0.9$. At this point, our cost is $1.19\times$ the offline optimum and $1.08\times$ the Oracle-Assisted Baseline. Given that both baselines operate in substantially easier settings — with full advance access to the data and, in the oracle case, ground-truth cluster structure — this gap is quite moderate. It demonstrates that with only an estimate of OPT as input, our algorithm remains competitive with baselines granted significantly stronger advantages.

ϵ-crossover consistency. As shown in Table 1, the crossover value is consistent across all datasets: it is 0.9 for MAGIC, Adult, and Skin, and 1.1 for Power. This consistency suggests that choosing the hyperparameter ϵ is not highly dataset-dependent, which makes setting it much easier in practice.

We conclude and discuss directions for future work in Appendix A.

AI USE STATEMENT

Generative AI tools were used only to polish the writing to improve readability.

REPRODUCIBILITY STATEMENT

In terms of reproducing our theoretical results, we have carefully stated our model and assumptions in Section 2. Further, any missing proofs can be found in Appendix C and Appendix D.

The supplementary material contains all code necessary to reproduce our experimental results. We have provided links to the datasets we utilized and additional experimental details in Appendix E.

REFERENCES

Matteo Almanza, Flavio Chierichetti, Silvio Lattanzi, Alessandro Panconesi, and Giuseppe Re. Online facility location with multiple advice. In *NeurIPS*, pp. 4661–4673, 2021.

Robi Bhattacharjee and Michal Moshkovitz. No-substitution k-means clustering with adversarial order. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, Proceedings of Machine Learning Research, pp. 345–366. PMLR, 2021. URL http://proceedings.mlr.press/v132/bhattacharjee21a.html.

Robi Bhattacharjee, Jacob Imola, Michal Moshkovitz, and Sanjoy Dasgupta. Online k-means clustering on arbitrary data streams. In Shipra Agrawal and Francesco Orabona (eds.), *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, Proceedings of Machine Learning Research, pp. 204–236. PMLR, 2023. URL https://proceedings.mlr.press/v201/bhattacharjee23b.html.

Vincent Cohen-Addad, Benjamin Guedj, Varun Kanade, and Guy Rom. Online k-means clustering. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 1126–1134. PMLR, 2021. URL http://proceedings.mlr.press/v130/cohen-addad21a.html.

Vincent Cohen-Addad, Silvio Lattanzi, Andreas Maggiori, and Nikos Parotsidis. Online and consistent correlation clustering. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pp. 4157–4179. PMLR, 2022. URL https://proceedings.mlr.press/v162/cohen-addad22a.html.

Graham Cormode and Andrew McGregor. Approximation algorithms for clustering uncertain data. In Maurizio Lenzerini and Domenico Lembo (eds.), *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2008, June 9-11, 2008, Vancouver, BC, Canada*, pp. 191–200. ACM, 2008. doi: 10.1145/1376916.1376944. URL https://doi.org/10.1145/1376916.1376944.

Michael Dinitz, Sungjin Im, Thomas Lavastida, Benjamin Moseley, Aidin Niaparast, and Sergei Vassilvitskii. Binary search with distributional predictions. In *NeurIPS*, 2024.

Jon C. Ergun, Zhili Feng, Sandeep Silwal, David P. Woodruff, and Samson Zhou. Learning-augmented k-means clustering. In *ICLR*. OpenReview.net, 2022.

Hendrik Fichtenberger, Silvio Lattanzi, Ashkan Norouzi-Fard, and Ola Svensson. Consistent k-clustering for general metrics. In *ACM-SIAM Symposium on Discrete Algorithms*, 2021.

Dimitris Fotakis, Evangelia Gergatsouli, Themistoklis Gouleakis, Nikolas Patris, and Thanos Tolias. Improved bounds for online facility location with predictions. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 26973–26981. AAAI Press, 2025. doi: 10.1609/AAAI.V39I25.34903. URL https://doi.org/10.1609/aaai.v39i25.34903.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5. URL https://doi.org/10.1016/0304-3975(85)90224-5.

Sudipto Guha and Kamesh Munagala. Exceeding expectations and clustering uncertain data. In Jan Paredaens and Jianwen Su (eds.), *Proceedings of the Twenty-Eigth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009, June 19 - July 1, 2009, Providence, Rhode Island, USA*, pp. 269–278. ACM, 2009. doi: 10.1145/1559795.1559836. URL https://doi.org/10.1145/1559795.1559836.

Xiangyu Guo, Janardhan Kulkarni, Shi Li, and Jiayi Xian. Consistent k-median: Simpler, better and robust. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1135–1143, 2021.

Tom Hess, Michal Moshkovitz, and Sivan Sabato. A constant approximation algorithm for sequential random-order no-substitution k-median clustering. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 3298–3308, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1aa057313c28fa4a40c5bc084b11d276-Abstract.html.

Dorit S. Hochbaum and David B. Shmoys. A best possible heuristic for the *k*-center problem. *Math. Oper. Res.*, 10(2):180–184, 1985. doi: 10.1287/MOOR.10.2.180. URL https://doi.org/10.1287/moor.10.2.180.

Junyu Huang, Qilong Feng, Ziyun Huang, Zhen Zhang, Jinhui Xu, and Jianxin Wang. New algorithms for the learning-augmented k-means problem. In *ICLR*. OpenReview.net, 2025.

Shaofeng H.-C. Jiang, Erzhi Liu, You Lyu, Zhihao Gavin Tang, and Yubo Zhang. Online facility location with predictions. In *ICLR*. OpenReview.net, 2022.

Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002. ISBN 0-387-95313-2. doi: 10.1007/978-1-4757-4015-8. URL http://dx.doi.org/10.1007/978-1-4757-4015-8.

Silvio Lattanzi and Sergei Vassilvitskii. Consistent k-clustering. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pp. 1975–1984. PMLR, 2017. URL http://proceedings.mlr.press/v70/lattanzi17a.html.

Silvio Lattanzi, Benjamin Moseley, Sergei Vassilvitskii, Yuyan Wang, and Rudy Zhou. Robust online correlation clustering. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 4688–4698, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/250dd56814ad7c50971ee4020519c6f5-Abstract.html.

Edo Liberty, Ram Sriharsha, and Maxim Sviridenko. An algorithm for online k-means clustering. In Michael T. Goodrich and Michael Mitzenmacher (eds.), *Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments, ALENEX 2016, Arlington, Virginia, USA, January 10, 2016*, pp. 81–89. SIAM, 2016. doi: 10.1137/1.9781611974317.7. URL https://doi.org/10.1137/1.9781611974317.7.

Claire Mathieu, Ocan Sankur, and Warren Schudy. Online correlation clustering. In Jean-Yves Marion and Thomas Schwentick (eds.), *27th International Symposium on Theoretical Aspects of Computer Science, STACS 2010, Nancy, France, March 4-6, 2010*, LIPIcs, pp. 573–584. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2010. doi: 10.4230/LIPICS.STACS.2010.2486. URL https://doi.org/10.4230/LIPIcs.STACS.2010.2486.

Adam Meyerson. Online facility location. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, Las Vegas, Nevada, USA, October 14-17, 2001*, pp. 426–431. IEEE Computer Society, 2001. doi: 10.1109/SFCS.2001.959917. URL https://doi.org/10.1109/SFCS.2001.959917.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. *Commun. ACM*, 65(7): 33–35, 2022. doi: 10.1145/3528087. URL https://doi.org/10.1145/3528087.

Benjamin Moseley, Heather Newman, and Kirk Pruhs. Online k-median with consistent clusters. In Amit Kumar and Noga Ron-Zewi (eds.), *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2024, London School of Economics, London, UK, August 28-30, 2024*, LIPIcs, pp. 20:1–20:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024. doi: 10.4230/LIPICS.APPROX/RANDOM.2024.20. URL https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2024.20.

Michal Moshkovitz. Unexpected effects of online no-substitution k-means clustering. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, Proceedings of Machine Learning Research, pp. 892–930. PMLR, 2021. URL http://proceedings.mlr.press/v132/moshkovitz21a.html.

Thy Dinh Nguyen, Anamay Chaturvedi, and Huy L. Nguyen. Improved learning-augmented algorithms for k-means and k-medians clustering. In *ICLR*. OpenReview.net, 2023.

Andrei V. Novikov. Pyclustering: Data mining library. *Journal of Open Source Software*, 4(36):1230, 2019. doi: 10.21105/joss.01230.

Tim Roughgarden. Lecture #16: The random-order model for online algorithms. Lecture Notes, CS264: Beyond Worst-Case Analysis, Stanford University, March 2017. Available at https://timroughgarden.org/w17/l/l16.pdf.

A Conclusion and Future Work

We gave the first framework for online k-clustering under strict consistency constraints (zero recourse and irrevocable assignments) that achieves an $O(1)$-approximation. The key insight is that a natural stochastic modeling assumption is sufficient to circumvent the strong adversarial lower bounds that have dominated this line of work. Crucially, our algorithms do not know the distribution and require only a single distributional parameter (the expected optimal cost), are robust to errors in this estimate, and perform well empirically, often far exceeding what the theoretical guarantees suggest.

We believe the stochastic modeling perspective adopted here is broadly applicable beyond k-center and k-median. A natural direction for future work is other online clustering problems. For example, online correlation clustering under consistency constraints (Mathieu et al., 2010; Lattanzi et al., 2021; Cohen-Addad et al., 2022) poses similar adversarial lower bound barriers, and may be another natural setting where stochastic arrivals unlock non-trivial guarantees.

B Lower Bound for Algorithms with No Resource Augmentation

We show that resource augmentation is necessary for constant competitiveness even under i.i.d. arrivals and complete knowledge of the input distribution.

Proposition 12. *For every $n \geq 4$, there exists a three-point metric and an i.i.d. distribution over n arrivals such that every online algorithm with zero recourse, irrevocable assignments, and at most $k = 2$ centers has competitive ratio $\Omega(n^2)$ for k-median and $\Omega(n^3)$ for k-center. These bounds hold even for randomized algorithms given the distribution and the horizon n in advance.*

Proof. Consider the metric space $\{A, B, O\}$ with

$$d(A, B) = n^2, \qquad d(A, O) = d(B, O) = n^6.$$

Each arrival is drawn independently according to

$$\Pr[X_i = A] = \Pr[X_i = B] = \frac{1 - n^{-4}}{2}, \qquad \Pr[X_i = O] = n^{-4}.$$

Let F denote the event that O appears at least once. By a union bound, $\Pr[F] \leq n^{-3}$.

Offline cost. If F does not occur, opening centers at A and B gives zero cost for either objective. If F occurs, opening a center at O and another at either A or B gives k-median cost at most $n \cdot n^2 = n^3$ and k-center cost at most n^2. Thus, writing μ_{med} and μ_{ctr} for the respective expected optimal costs,

$$\mu_{\mathrm{med}} \le \Pr[F]\, n^3 \le 1, \qquad \mu_{\mathrm{ctr}} \le \Pr[F]\, n^2 \le \frac{1}{n}.$$

Online cost. Let E be the event that the first two arrivals are A and B in either order. Then

$$\Pr[E] = \frac{(1-n^{-4})^2}{2} \ge \frac{1}{4}.$$

Condition on E. The algorithm must open the first arrival as a center. If it does not open the second arrival, its irrevocable assignment immediately incurs distance n^2, giving cost at least n^2 for either objective.

Otherwise, both center slots are occupied by A and B. Since future arrivals are independent of the observed prefix and the algorithm's random choices, the probability that O appears among the remaining $n-2$ arrivals is

$$q = 1 - (1-n^{-4})^{n-2}.$$

Using $1 - x \le e^{-x}$ and $1 - e^{-t} \ge t/2$ for $t \in [0,1]$, we obtain

$$q \ge 1 - e^{-(n-2)n^{-4}} \ge \frac{(n-2)n^{-4}}{2} \ge \frac{1}{4n^3},$$

where the second inequality uses $(n-2)n^{-4} \le 1$, and the last uses $n - 2 \ge n/2$ for $n \ge 4$. If O appears, it must be assigned to an existing center at distance n^6. Opening the second center therefore incurs conditional expected cost at least

$$qn^6 \ge \frac{n^3}{4} \ge n^2$$

for either objective.

Both decisions consequently incur conditional expected cost at least n^2, and the same holds for any randomized mixture of them. Hence

$$\mathbb{E}[\mathrm{ALG}] \ge \Pr[E]\, \mathbb{E}[\mathrm{ALG} \mid E] \ge \frac{n^2}{4}.$$

Dividing by the respective expected optimal costs gives

$$\frac{\mathbb{E}[\mathrm{ALG}_{\mathrm{med}}]}{\mu_{\mathrm{med}}} \ge \frac{n^2}{4}, \qquad \frac{\mathbb{E}[\mathrm{ALG}_{\mathrm{ctr}}]}{\mu_{\mathrm{ctr}}} \ge \frac{n^3}{4},$$

as claimed. □

Thus, stochastic arrivals and full distributional knowledge alone do not suffice for constant competitiveness under a strict budget of k centers, even when $k = 2$.

C Missing Proofs from Section 3

Proof of Lemma 4. Suppose that $\mathrm{OPT}(X) \le \frac{1}{\epsilon}\widehat{\mu}$. In this case, we claim that the algorithm opens at most k centers (i.e., the algorithm never increases R). If not, then when the algorithm opens the $(k+1)$'st center, this constitutes $k+1$ points whose pairwise distances are all at least $\frac{1}{\epsilon^2}\widehat{\mu} \ge \frac{2}{\epsilon}\widehat{\mu} \ge 2\mathrm{OPT}(X)$, contradicting the definition of $\mathrm{OPT}(X)$. Since the algorithm opens at most k centers in this case, we have that $\mathrm{Cost}(X, C, \rho) \le \frac{1}{\epsilon^2}\widehat{\mu}$.

Now suppose that $(\frac{1}{\epsilon})^{\ell}\widehat{\mu} < \mathrm{OPT}(X) \le (\frac{1}{\epsilon})^{\ell+1}\widehat{\mu}$ for some $\ell \ge 1$. If the algorithm opens its last center in phase ℓ or earlier, then we clearly have $\mathrm{Cost}(X, C, \rho) \le (\frac{1}{\epsilon})^{2+\ell}\widehat{\mu} \le \frac{1}{\epsilon^2}\mathrm{OPT}(X)$. Suppose for contradiction that Algorithm 1 opens a center in phase $\ell + 1$. This means that the algorithm opened k centers in phase ℓ and at least 1 center in phase $\ell + 1$. This constitutes $k+1$ points whose pairwise distances are all at least $(\frac{1}{\epsilon})^{2+\ell}\widehat{\mu} \ge \frac{1}{\epsilon}\mathrm{OPT}(X) \ge 2\mathrm{OPT}(X)$, contradicting the definition of $\mathrm{OPT}(X)$. □

Proof of Lemma 5. Let $\mathcal{E}_\ell$ be the event that $(\frac{1}{\epsilon})^\ell \widehat{\mu} < \mathrm{OPT}(X) \le (\frac{1}{\epsilon})^{\ell+1}\widehat{\mu}$ for $\ell \ge 1$ and $\mathcal{E}_0$ be the event that $\mathrm{OPT}(X) \le \frac{1}{\epsilon}\widehat{\mu}$. As discussed in the proof of Lemma 4, when event $\mathcal{E}_\ell$ has occurred (for $\ell \ge 0$), Algorithm 1 opens at most $k \cdot (1+\ell)$ centers. Since the events $\{\mathcal{E}_\ell\}_{\ell \ge 0}$ are mutually exclusive and exhaustive we have

$$\begin{aligned}
\mathbb{E}[|C|] &= \sum_{\ell=0}^{\infty} \mathbb{E}[|C| \mid \mathcal{E}_\ell] \Pr[\mathcal{E}_\ell] \le \sum_{\ell=0}^{\infty} k \cdot (1+\ell) \Pr[\mathcal{E}_\ell] = k\left(1 + \sum_{\ell=0}^{\infty} \ell \Pr[\mathcal{E}_\ell]\right) \\
&= k\left(1 + \sum_{\ell=1}^{\infty}\sum_{j=1}^{\infty} \mathbf{1}\{j \le \ell\} \Pr[\mathcal{E}_\ell]\right) = k\left(1 + \sum_{j=1}^{\infty}\sum_{\ell=j}^{\infty} \Pr[\mathcal{E}_\ell]\right) \\
&= k\left(1 + \sum_{j=1}^{\infty} \Pr\left[\mathrm{OPT}(X) > \left(\frac{1}{\epsilon}\right)^j \widehat{\mu}\right]\right) \le k\left(1 + \frac{\mathbb{E}[\mathrm{OPT}(X)]}{\widehat{\mu}} \sum_{j=1}^{\infty} \epsilon^j\right) \\
&= k\left(1 + \frac{\mu}{\widehat{\mu}}\left(\frac{\epsilon}{1-\epsilon}\right)\right),
\end{aligned}$$

where the first inequality follows from the observation above, and the second inequality follows from Markov's inequality. This completes the proof of the lemma. □

D Missing Proofs from Section 4

Proof of Proposition 9. Since $|S(c^*)| = |S_{\mathrm{f}}(c^*)| + |S_{\mathrm{c}}(c^*)|$, we only need to prove one of the statements above. Suppose for the sake of contradiction that $|S_{\mathrm{f}}(c^*)| > \epsilon |S(c^*)|$. Then we have

$$\bar{d}(c^*) = \frac{1}{|S(c^*)|} \sum_{i \in S(c^*)} d(X_i, c^*) \ge \frac{1}{|S(c^*)|} \sum_{i \in S_{\mathrm{f}}(c^*)} d(X_i, c^*) \ge \frac{|S_{\mathrm{f}}(c^*)|}{|S(c^*)|} \cdot \frac{\bar{d}(c^*)}{\epsilon} > \bar{d}(c^*),$$

which is clearly a contradiction. □

Proof of Lemma 10. Let j^* be the period where the first close point is opened as a center. Note that j^* is a random variable so some care is needed to make the argument rigorous. Clearly, by linearity of expectations and since $\mathbf{1}\{i \le j^*\} + \mathbf{1}\{i > j^*\} = 1$, we have

$$\mathbb{E}\left[\sum_{i \in S_{\mathrm{c}}(c^*)} p_i\right] = \sum_{i \in S_{\mathrm{c}}(c^*)} \mathbb{E}\left[p_i \mathbf{1}\{i \le j^*\}\right] + \sum_{i \in S_{\mathrm{c}}(c^*)} \mathbb{E}[p_i \mathbf{1}\{i > j^*\}]. \tag{1}$$

We claim that the first sum on the right hand side of (1) is at most 1. To prove this, we construct a sequence $a_1, \ldots, a_m$, where $m = |S_{\mathrm{c}}(c^*)|$ and a_1 equals the target sum, then inductively show that $a_j \le 1$ for all $j \in [m]$. Without loss of generality, re-index the arrivals in $S_{\mathrm{c}}(c^*)$ so that they are $1, 2, \ldots, m$. Now for each $j \in [m]$, we define $a_j = \mathbb{E}[\sum_{i \ge j} p_j \mathbf{1}\{j^* \ge i\} \mid j^* \ge j]$. Since we always have $j^* \ge 1$, we see that $a_1 = \sum_{i \in S_{\mathrm{c}}(c^*)} \mathbb{E}\left[p_i \mathbf{1}\{i \le j^*\}\right]$. We now prove by induction that $a_j \le 1$ for all $j \in [m]$. The base case is when $j = m$. Clearly, we have $a_m = \mathbb{E}[p_m \mathbf{1}\{j^* \ge m\} \mid j^* \ge m] \le 1$ since $p_m \le 1$ almost surely. Now for the inductive step, consider $1 \le j < m$ and suppose that $a_{j+1} \le 1$. We have

$$a_j = \mathbb{E}\left[\sum_{i \ge j} p_j \mathbf{1}\{j^* \ge i\} \middle| j^* \ge j\right] = \mathbb{E}\left[p_j \mid j^* \ge j\right] + \mathbb{E}\left[\sum_{i \ge j+1} p_i \mathbf{1}\{j^* \ge i\} \middle| j^* \ge j\right].$$

Focusing on the latter term, since $j^* \ge j+1 \implies j^* \ge j$, we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \ge j+1} p_i \mathbf{1}\{j^* \ge i\} \middle| j^* \ge j\right] &= \mathbb{E}\left[\sum_{i \ge j+1} p_i \mathbf{1}\{j^* \ge i\} \middle| j^* \ge j+1\right] \Pr[j^* \ge j+1 \mid j^* \ge j] \\
&= a_{j+1} \Pr[j^* \ge j+1 \mid j^* \ge j] = a_{j+1} \Pr[B_j = 0 \mid j^* \ge j] \\
&= a_{j+1}(1 - \mathbb{E}[p_j \mid j^* \ge j]).
\end{aligned}$$

Substituting this into our equation for a_j, we have

$$a_j = \mathbb{E}[p_j \mid j^* \geq j] + (1 - \mathbb{E}[p_j \mid j^* \geq j])a_{j+1} \leq \mathbb{E}[p_j \mid j^* \geq j] + 1 - \mathbb{E}[p_j \mid j^* \geq j] = 1,$$

where in the inequality we use that $a_{j+1} \leq 1$ by the induction hypothesis. Thus we have $a_1 \leq 1$, and so the first term in the right hand side of (1) is at most 1.

Moving to the latter term, recall that D_i is the distance from X_i to its closest center when X_i arrives. Note that by definition of j^*, X_{j^*} is a center. So for $i > j^*$, we have (by the triangle inequality) that $D_i \leq d(X_i, X_{j^*}) \leq d(X_i, c^*) + d(X_{j^*}, c^*) \leq 2\bar{d}(c^*)/\epsilon$, since both X_i and X_{j^*} are close. We have $p_i \leq \epsilon^2 k D_i/\widehat{\mu}$, so

$$\sum_{i \in S_{\mathrm{c}}(c^*)} \mathbb{E}[p_i \mathbf{1}\{i > j^*\}] \leq \frac{\epsilon^2 k}{\widehat{\mu}} \sum_{i \in S_{\mathrm{c}}(c^*)} \mathbb{E}[D_i \mathbf{1}\{i > j^*\}] \leq \frac{2\epsilon k}{\widehat{\mu}} m\bar{d}(c^*) \leq \frac{2\epsilon k}{\widehat{\mu}} \sum_{i \in S(c^*)} d(X_i, c^*),$$

where we use that $m\bar{d}(c^*) = \frac{|S_{\mathrm{c}}(c^*)|}{|S(c^*)|} \sum_{i \in S(c^*)} d(X_i, c^*) \leq \sum_{i \in S(c^*)} d(X_i, c^*)$. Substituting the bounds we derived for each term on the right hand side of (1) yields the lemma. □

Proof of Lemma 11. Consider some $X_i \in S_{\mathrm{f}}(c^*)$ and suppose that it arrives after some close point $X_j \in S_{\mathrm{c}}(c^*)$. Since every center that was open when X_j arrived is available when X_i arrives, so by the triangle inequality we have

$$D_i \leq D_j + d(X_i, X_j) \leq D_j + d(X_i, c^*) + d(X_j, c^*).$$

This implies that

$$p_i \leq p_j + \frac{\epsilon^2 k}{\widehat{\mu}}(d(X_i, c^*) + d(X_j, c^*))$$

whenever X_i arrives after X_j. To see this, first note that the inequality is trivial if $p_j = 1$ since $p_i \leq 1$. If $p_j < 1$, then $p_j = \epsilon^2 k D_j/\widehat{\mu}$, and so the right hand side equals $\frac{\epsilon^2 k}{\widehat{\mu}}(D_j + d(X_i, c^*) + d(X_j, c^*)) \geq \epsilon^2 k D_i/\widehat{\mu} \geq p_i$. From here, we use the fact that points arrive in a random order. Recall that $m = |S_{\mathrm{c}}(c^*)|$ is the number of close points. There are $m + 1$ ways that X_i can arrive relative to the close points, each of which is equally likely: it can arrive before all of the close points or between two of the close points. Let $\mathcal{E}_0$ be the event that X_i arrives before all of the close points and let $\mathcal{E}_j$ be the event that X_i arrives after X_j but before the next close arrival (if one exists) for each $X_j \in S_{\mathrm{c}}(c^*)$. In the case that X_i arrives before all of the close points, we can use the trivial bound $p_i \leq 1$. Thus we have:

$$\begin{aligned}
\mathbb{E}[p_i] &= \sum_{j \in \{0\} \cup S_{\mathrm{c}}(c^*)} \mathbb{E}[p_i \mid \mathcal{E}_j] \Pr[\mathcal{E}_j] \\
&\leq \frac{1}{|S_{\mathrm{c}}(c^*)| + 1} \left(1 + \sum_{j \in S_{\mathrm{c}}(c^*)} \left(\mathbb{E}[p_j] + \frac{\epsilon^2 k}{\widehat{\mu}} (d(X_i, c^*) + d(X_j, c^*))\right)\right) \\
&\leq \frac{1}{|S_{\mathrm{c}}(c^*)|} \left(1 + \sum_{j \in S_{\mathrm{c}}(c^*)} \mathbb{E}[p_j] + \frac{\epsilon^2 k}{\widehat{\mu}} \sum_{j \in S_{\mathrm{c}}(c^*)} d(X_j, c^*)\right) + \frac{\epsilon^2 k}{\widehat{\mu}} d(X_i, c^*)
\end{aligned}$$

Summing over all far points yields

$$\begin{aligned}
\sum_{i \in S_{\mathrm{f}}(c^*)} \mathbb{E}[p_i] &\leq \frac{|S_{\mathrm{f}}(c^*)|}{|S_{\mathrm{c}}(c^*)|} \left(1 + \sum_{j \in S_{\mathrm{c}}(c^*)} \mathbb{E}[p_j] + \frac{\epsilon^2 k}{\widehat{\mu}} \sum_{j \in S_{\mathrm{c}}(c^*)} d(X_j, c^*)\right) + \frac{\epsilon^2 k}{\widehat{\mu}} \sum_{i \in S_{\mathrm{f}}(c^*)} d(X_i, c^*) \\
&\leq \frac{\epsilon}{1 - \epsilon} \left(2 + \frac{2\epsilon k}{\widehat{\mu}} \sum_{j \in S(c^*)} d(X_j, c^*)\right) + \frac{\epsilon^2 k}{\widehat{\mu}} \sum_{i \in S(c^*)} d(X_i, c^*) \\
&\leq 4\epsilon + \frac{5\epsilon^2 k}{\widehat{\mu}} \sum_{i \in S(c^*)} d(X_i, c^*),
\end{aligned}$$

where in the first step we use Proposition 9. This completes the proof of the lemma. □

Proof of Lemma 7. Using Lemmas 10 and 11, we have

$$\sum_{i=1}^{n}\mathbb{E}[p_i] = \sum_{c^*\in C^*}\sum_{i\in S(c^*)}\mathbb{E}[p_i] = \sum_{c^*\in C^*}\left(\sum_{i\in S_c(c^*)}\mathbb{E}[p_i] + \sum_{i\in S_f(c^*)}\mathbb{E}[p_i]\right)$$
$$\leq \sum_{c^*\in C^*}\left(1 + \frac{2\epsilon k}{\widehat{\mu}}\sum_{i\in S(c^*)} d(X_i, c^*) + 4\epsilon + \frac{5\epsilon^2 k}{\widehat{\mu}}\sum_{i\in S(c^*)} d(X_i, c^*)\right)$$
$$= k + \frac{2\epsilon k}{\widehat{\mu}}\mathrm{OPT} + 4\epsilon k + \frac{5\epsilon^2 k}{\widehat{\mu}}\mathrm{OPT} = k(1 + O(\epsilon) + O(\epsilon)\mathrm{OPT}/\widehat{\mu}),$$

which is the bound we wanted to prove. □

E Additional Experimental Details and Results

E.1 Dataset Details

We use four datasets from the UC Irvine Machine Learning Repository: Adult, MAGIC, Skin, and Power (all of which have been released under the CC-BY 4.0 license). In each case, we retain only numerical features and remove rows with missing values before running the experiments. We give a more detailed discussion of the datasets, their original features, and any preprocessing steps here.

- **Magic** Gamma Telescope: Includes 10 features and 19,020 data points. Source: UCI Machine Learning Repository.
- **Skin** Segmentation: Consists of 3 features and 245,057 data points. Source: UCI Machine Learning Repository.
- **Adult** Census Income: Includes 32,561 points. For this study, 6 numeric features were retained from the original 14. Source: UCI Machine Learning Repository.
- Individual Household Electric **Power** Consumption: Includes 2,049,280 points. Date and time attributes were excluded, leaving 7 features. Source: UCI Machine Learning Repository.

E.2 Experimental Setup

We use Python 3.13 for conducting our experiments on a system equipped with an 11th Generation Intel Core i7 CPU running at 2.80GHz, 32GB of RAM, a 128GB NVMe KIOXIA disk drive, and a 64-bit Windows 11 Enterprise operating system. It's worth noting that the costs of the algorithms, i.e., k-median cost function and the number of centers, are hardware-independent.

E.3 Additional Plots

This appendix contains the experimental plots for the datasets not shown in the main text, see Figures 2, 3, and 4. The left plot in each figure shows objective value versus ϵ, and the right plot shows the number of selected centers versus ϵ. In each plot, the solid line shows the mean over runs and the surrounding shaded region shows one standard deviation.

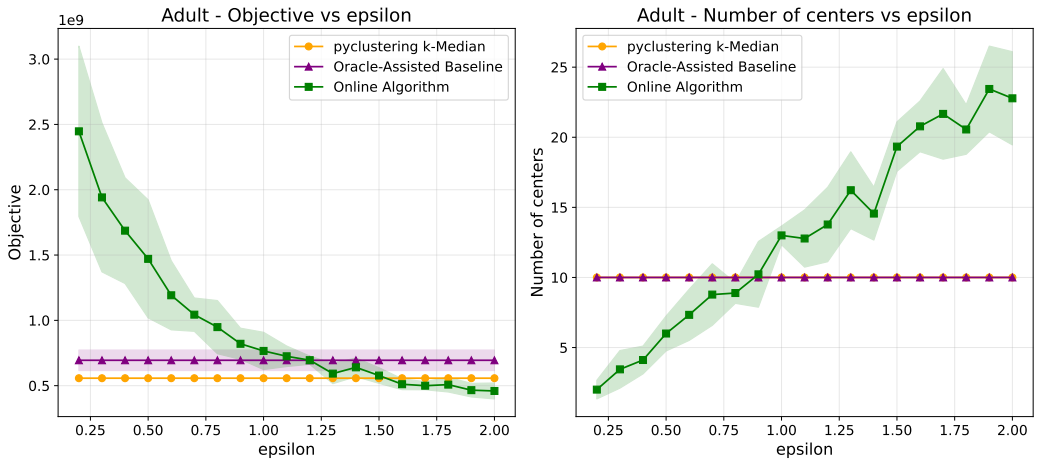


Figure 2: Results on the Adult dataset.

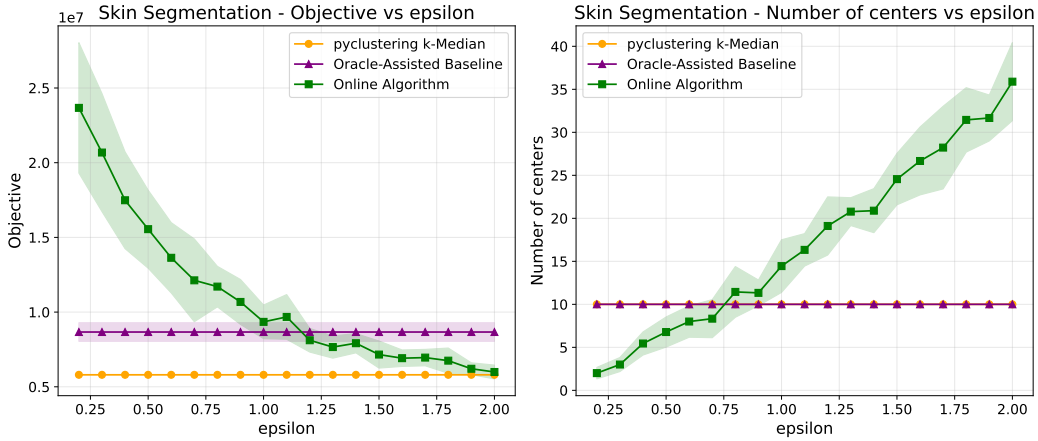


Figure 3: Results on the Skin dataset.

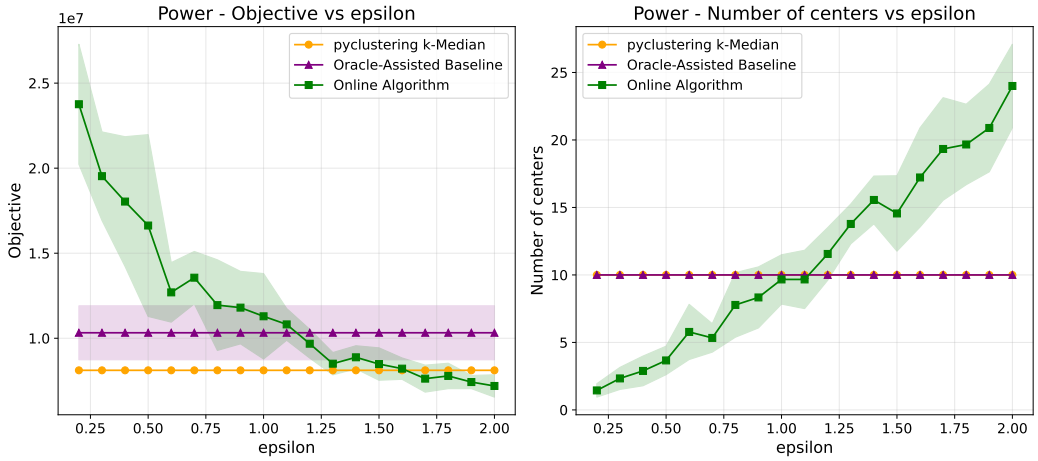


Figure 4: Results on the Power dataset.