ROBUST ALGORITHMS FOR FACILITY LOCATION WITH NOISY PREDICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the fundamental problem of facility location in settings where algorithms receive noisy predictions about the optimal solution. Motivated by real-world scenarios in which side information from historical data or machine learning models is imperfect, we develop robust algorithms that effectively leverage these noisy predictors while maintaining provable guarantees. We study three distinct models for prediction noise. In the first model, each client is assigned a predicted cluster label, and the predictor may mislabel up to an α fraction of the clients within each true cluster and within each predicted cluster. In the second model, the algorithm can issue pairwise queries asking whether two clients are served by the same facility in an optimal solution, and each response is independently corrupted with probability at most $\alpha < 1/2$. In the third model, the predictor provides a candidate set of facilities that may include η^+ false positives and η^- false negatives relative to an optimal solution. For the Uncapacitated Facility Location (UFL) problem, we obtain several results under these noise models. With a noisy label predictor, we achieve a $(1 + O(\alpha))$-approximation provided the error rate $\alpha < 1/2$. Using noisy pairwise queries, we obtain a $(1 + \varepsilon)$-approximation algorithm with query complexity $\tilde{O}\left(\max\left\{\frac{|\mathcal{C}|k}{(1-2\alpha)^4 s_{\min}}, \frac{|\mathcal{C}|}{\delta s_{\min}}\right\}\right)$, where $\mathcal{C}$ is the set of clients, k is the number of clusters in the fixed optimum, and $s_{\min}$ is the minimum optimal cluster size. We complement the oracle results with an ETH-based query lower bound: for a constant $\beta_0 > 0$ any subexponential-time $(1 + \beta)$-approximation with $\beta < \beta_0$ must satisfy $q_1 \log k + q_2 = \Omega(k^{1-o(1)})$, where q_1 and q_2 are the numbers of label and pairwise queries. Finally, given a predicted facility set, we design an algorithm whose approximation guarantee degrades linearly with the opening-cost masses η^+ and η^-. We also show that such linear dependence on both error parameters is unavoidable in the high-accuracy regime. Together, these results establish a clear trade-off between prediction quality and approximation performance for the facility location problem across a variety of practical noise models.

1 INTRODUCTION

Facility location is a fundamental problem with broad applications (Xu & Wunsch, 2005). The NP-hard Uncapacitated Facility Location (UFL) problem admits constant-factor approximations (Hochbaum, 1982; Shmoys et al., 1997), with a best-known ratio of 1.488 (Li, 2013) and a theoretical lower bound of 1.463 (Guha & Khuller, 1999). Facility location and clustering share the common theme of assigning clients to centers (facilities or cluster centroids) to minimize connection costs, allowing insights and techniques to transfer between the two. For k-means and k-median, despite the dominance of practical heuristics such as Lloyd's algorithm (Lloyd, 1982; Arthur & Vassilvitskii, 2007; Sculley, 2010), theory has achieved improved approximations: 5.913 for Euclidean k-means (Cohen-Addad et al., 2022a) and $(2 + \varepsilon)$ for k-median (Cohen-Addad et al., 2025), along with strong hardness and coreset lower bounds (Cohen-Addad et al., 2021; 2022b;c). Classical algorithms assume full access to the input and seek to optimize the clustering costs.

However, in many practical settings, the algorithm's access to ground-truth information is limited, noisy, or indirect. The study of algorithms augmented with predictions has grown significantly, with early work primarily focused on online settings such as caching (Lykouris & Vassilvitskii, 2021;

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.

Rohatgi, 2020; Bansal et al., 2022), rent-or-buy problems (Purohit et al., 2018; Angelopoulos et al., 2024), scheduling (Azar et al., 2021; Lindermayr & Megow, 2022), and online graph problems (including the (capacitated) facility location problem) (Azar et al., 2022). The field continues to diversify, with advances in binary search using distributional predictions (Dinitz et al., 2024), warm-start competitive algorithms (Blum & Srinivas, 2025), and dynamic graph algorithms (Brand et al., 2024), demonstrating the broad applicability of prediction-based methods across many domains.

More recently, these techniques have been successfully applied to offline problems, including matching (Dinitz et al., 2021), clustering (Ergun et al., 2022), and sorting (Bai & Coester, 2023; Lu et al., 2021). In the setting with ε-accurate predictions, each input item is independently labeled as belonging to the optimal solution with probability $\frac{1}{2}+\varepsilon$. Recent studies have explored this framework in the contexts of Maximum Cut (Cohen-Addad et al., 2024), Independent Set (Braverman et al., 2024), and Constraint Satisfaction problems (Ghoshal et al., 2025). The authors in (Aamand et al., 2025) consider an edge-based variant of the model, where each edge contributes two bits of information; under this formulation, they obtain improved algorithms for NP-hard problems such as Vertex Cover and Set Cover.

In this paper, we study offline facility location in a setting where the algorithm receives only *noisy predictions* about the underlying structure. We focus on three natural and widely studied prediction models. In the next subsection, we explain these models and why we consider them in detail.

1.1 Problem Setting

The (metric) Uncapacitated Facility Location (UFL) problem is a classic optimization problem (Hochbaum, 1982) and consists of a metric space $(\mathcal{F}\cup\mathcal{C}, d)$, a set of candidate facilities $\mathcal{F}$ with opening costs $\{f_i\}_{i\in\mathcal{F}}$, and a set of clients $\mathcal{C}$. A solution opens a facility set $S\subseteq\mathcal{F}$ and connects each client to its nearest open facility. The objective is

$$\mathrm{cost}(S)=\sum_{i\in S} f_i+\sum_{j\in\mathcal{C}}\min_{i\in S} d(i,j).$$

A solution S is called a (λ_f,λ_c)-approximation if, for any solution $T\subseteq\mathcal{F}$, it holds that $\mathrm{cost}(S)\le\lambda_f F^*+\lambda_c C^*$, where F^* and C^* denote the opening and connection costs of T, respectively. When $\lambda_f=\lambda_c=\lambda$, this reduces to a standard λ-approximation. In terms of bifactor approximations, (Jain et al., 2002) show that achieving $\lambda_c<1+2e^{-\lambda_f}$ is impossible unless P = NP. For a large λ_f, (Byrka & Aardal, 2010) achieve an $(\lambda_f, 1+2e^{-\lambda_f})$-approximation for $\lambda_f\ge 1.6774$.

1.2 Prediction Models and Our Contributions

Let S^* denote an optimal solution and $\mathrm{OPT}=\mathrm{cost}(S^*)$. Since UFL may admit multiple optimal solutions, we fix an arbitrary optimal solution S^* and interpret all prediction/oracle answers with respect to this fixed S^*. We also adopt the clustering viewpoint of the UFL problem: the fixed optimal set S^* induces an optimal assignment (ties broken arbitrarily), and hence a partition of clients into $k=|S^*|$ *optimal clusters* $\mathcal{C}^*=\{C_1^*,\dots,C_k^*\}$, where each C_t^* is the set of clients served by a facility in S^*. We write $s_{\min}=\min_t|C_t^*|$ for the minimum optimal cluster size.

We consider three prediction models that provide different aspects of information about S^*. Our choice follows established lines of work in learning-augmented clustering, and we adopt the same predictors because UFL naturally induces a client partition via the optimal assignment.

(M1) Noisy label predictions. An α-noisy label predictor returns a cluster label $\Pi(j)\in[k]$ for each client $j\in\mathcal{C}$. We assume there exists a relabeling of $[k]$ such that, for every $t\in[k]$, the following two conditions hold: (i) at most an α-fraction of clients in the *true* cluster C_t^* are mislabeled, and (ii) among all clients predicted as label t, at most an α-fraction do *not* belong to C_t^*. Equivalently, each label achieves *precision* and *recall* at least $1-\alpha$. We assume $\alpha<1/2$, which is standard in this model (otherwise the predictor may be less accurate than random guessing).

With an α-noisy label predictor, we obtain a polynomial-time approximation with factor $1+\frac{4\alpha}{1-2\alpha}$. This guarantee assumes the algorithm has access to all predicted labels $\Pi(j)$. We also consider a query model in which the algorithm initially does not know these labels and a single query reveals the value of $\Pi(j)$ for a chosen client j. Under this model, we give a variant that achieves an additional

$(1+\varepsilon)$ multiplicative factor while using only $\tilde{O}(k\theta \cdot \mathrm{poly}(1/\varepsilon, 1/\delta))$ label queries, where θ captures cluster imbalance (definitions in Section 3). We refer to the number of such queries used by an algorithm as its *query complexity*.

We complement this upper bound with a query complexity lower bound. Assuming the Exponential Time Hypothesis (ETH), we construct an explicit hard metric UFL instance from Vertex Cover on 4-regular graphs and show that there is a constant $\beta_0 > 0$ such that any subexponential-time algorithm achieving a $(1+\beta)$-approximation for $\beta < \beta_0$ must make $\Omega(k^{1-\zeta}/\log k)$ label queries, for every $\zeta \in (0,1)$, even when the label oracle is exact. Thus, in the high-accuracy regime, one cannot hope to obtain a near-optimal solution while inspecting only a substantially smaller number of labels.

The α-noisy label prediction model has been studied extensively in the learning-augmented clustering literature. In particular, recent work develops learning-augmented algorithms for Euclidean k-clustering under noisy label predictions (Silwal, 2021; Gamlath et al., 2022; Ergun et al., 2022; Nguyen et al., 2023) and for general metric k-clustering under the same model (Fan & Shin, 2025). Moreover, the authors in (Silwal, 2021) implement such a predictor and empirically observe that it yields high-quality clusterings. At a high level, these works (and our result) show that given an α-noisy label predictor in a metric space, there exists a polynomial-time algorithm that achieves a $(1+f(\alpha))$-approximation for k-clustering, where $f(\alpha) \to 0$ as $\alpha \to 0$.

(M2) Noisy pairwise membership (same-cluster) queries. An oracle answers queries of the form

$$\Pi_{\mathrm{pair}}(u,v) \in \{0,1\} \quad \text{“are } u, v \text{ served by the same facility in } S^*\text{?”}.$$

Each answer is independently flipped with probability at most $\alpha < 1/2$. Once a pair is queried, the returned bit is fixed (no repeated-query amplification).

With an α-noisy pairwise oracle, we give a $(1+\varepsilon)$-approximation with near-linear query complexity for any fixed $\varepsilon \in (0,1)$. With probability at least $1-\delta$, our algorithm uses at most

$$\tilde{O}\left(\max\left\{\frac{|\mathcal{C}|\ k}{(1-2\alpha)^4\, s_{\min}},\ \frac{|\mathcal{C}|}{\delta\, s_{\min}}\right\}\right)$$

queries. Moreover, using the same hard instance as for the noisy label prediction model (M1), we prove that there exists a constant $\beta_0 > 0$ such that, for every $\zeta \in (0,1)$, any subexponential-time algorithm achieving a $(1+\beta)$-approximation for $\beta < \beta_0$ must make $\Omega(k^{1-\zeta})$ pairwise queries, even when the oracle is exact.

Clustering with pairwise membership queries was initiated in (Ashtiani et al., 2016). Subsequent work established query-efficient $(1+\varepsilon)$-approximations for k-clustering with exact or noisy oracles, along with polynomial-time algorithms and matching lower bounds (Ailon et al., 2018; Mazumdar & Saha, 2017; Peng & Zhang, 2021; Xia & Huang, 2022). The model has since been generalized to richer query types, including non-adaptive subset queries yielding sub-quadratic algorithms (Black et al., 2024).

(M3) Noisy solution predictions. A solution predictor outputs a candidate facility set $S \subseteq \mathcal{F}$ intended to approximate S^*. We measure prediction error in terms of *opening-cost mass*:

$$\eta^+ := \sum_{i \in S \setminus S^*} f_i \qquad \text{and} \qquad \eta^- := \sum_{i \in S^* \setminus S} f_i,$$

capturing false positives/negatives relative to the fixed optimum S^*.

Given a predicted set S with error (η^+, η^-) and access to any bifactor (λ_f, λ_c)-approximation for UFL, we obtain an approximation ratio

$$\lambda_c \ + \ \frac{\eta^+ + (\lambda_f - \lambda_c)\eta^-}{\mathrm{OPT}}.$$

If $\lambda_f \le \lambda_c$, then taking the prediction to be $S = \emptyset$ is the best choice, since it minimizes η^+ (but maximizes η^-). This recovers the same behavior as the no-prediction baseline, since the natural bound is $\frac{\lambda_f F^* + \lambda_c C^*}{\mathrm{OPT}}$. To potentially beat the classical approximation hardness, we should use a

bifactor algorithm with $\lambda_f > \lambda_c$, assuming a good predictor exists. For example, using the $(\lambda, 1 + 2e^{-\lambda})$-approximation algorithm from (Byrka & Aardal, 2010) gives:

$$1 + \frac{2}{e^{\lambda}} + \frac{\eta^+ + \left(\lambda - 1 - \frac{2}{e^{\lambda}}\right)\eta^-}{\mathrm{OPT}}$$

for some $\lambda \geq 1.6774$.

Here, we complement our results with a matching lower bound in terms of the approximation quality: we prove that the additive dependence on prediction errors in this model is unavoidable. In other words, there exists a constant $\beta_0 > 0$ such that, for any fixed $\rho < 1 + \beta_0$, no algorithm can guarantee a $\rho\,\mathrm{OPT} + o(\eta^-)$ cost bound for all predictions, nor can it guarantee a $\rho\,\mathrm{OPT} + o(\eta^+)$ cost bound. Thus, in the high-accuracy regime, both false negatives and false positives must enter the approximation guarantee linearly up to constant factors.

This solution-prediction model has also been studied in the literature. In particular, the authors in (Antoniadis et al., 2025) consider a setting in which the algorithm is given a proposed solution set that may contain false positives (η^+) and false negatives (η^-), capturing (for example) historical solutions or heuristic outputs. They show that for a minimization problem admitting a black-box ρ-approximation, one can obtain a learning-augmented algorithm with approximation ratio

$$1 + \frac{\eta^+ + (\rho - 1)\eta^-}{\mathrm{OPT}}.$$

They instantiate this framework for several NP-hard problems, including Minimum Vertex Cover and Minimum Steiner Tree, and obtain analogous guarantees for maximization problems.

Overall, we show that imperfect predictions suffice for constant-factor approximations for UFL, with guarantees that degrade smoothly with noise, yielding the first offline learning-augmented algorithms for facility location.

2 Preliminaries

We consider metric UFL on input $(\mathcal{F} \cup \mathcal{C}, d)$ with facilities $\mathcal{F}$, clients $\mathcal{C}$, and opening costs $\{f_i\}_{i \in \mathcal{F}}$. For any opened facility set $S \subseteq \mathcal{F}$, define

$$\mathrm{open}(S) := \sum_{i \in S} f_i, \qquad \mathrm{conn}(S) := \sum_{j \in \mathcal{C}} \min_{i \in S} d(i, j),$$

so $\mathrm{cost}(S) = \mathrm{open}(S) + \mathrm{conn}(S)$. Let S^* be a fixed optimal solution and $\mathrm{OPT} = \mathrm{cost}(S^*)$.

For a client subset $X \subseteq \mathcal{C}$ and a facility $i \in \mathcal{F}$, we write

$$\mathrm{cost}(X, i) \;:=\; f_i + \sum_{j \in X} d(i, j), \qquad \mathrm{conn}(X, i) \;:=\; \sum_{j \in X} d(i, j).$$

Let $\mathrm{cost}(X) := \min_{i \in \mathcal{F}} \mathrm{cost}(X, i)$ and define a minimizer $C_X \in \arg\min_{i \in \mathcal{F}} \mathrm{cost}(X, i)$.

Assignments and induced clusters. For each $S \neq \emptyset$, fix a deterministic tie-breaking rule and define the assignment $\sigma_S(j) \in \arg\min_{i \in S} d(i, j)$. For each facility $i \in S$, define its served client set $\mathcal{C}_i(S) := \{j \in \mathcal{C} : \sigma_S(j) = i\}$. As recalled in Section 1, for the fixed optimum S^* this induces a partition $\mathcal{C}^* = \{C_1^*, \ldots, C_k^*\}$ where $k := |S^*|$. We write $s_{\min} := \min_{t \in [k]} |C_t^*|$ for the minimum optimal cluster size.

Prediction models. As discussed in Section 1, we consider three predictors/oracles. For simplicity, we denote all predictors by Π; when needed, we use subscripts (e.g., Π_{lab}, Π_{pair}) to disambiguate.

(M1) Noisy label predictions. Fix any optimal solution S^* and its induced clusters $\mathcal{C}^* = \{C_1^*, \ldots, C_k^*\}$. The predictor assigns each client a label in $[k]$; let $C_t \subseteq \mathcal{C}$ denote the set of clients predicted to have label t. After an appropriate relabeling of $[k]$, we say the label predictor is α-noisy if for every $t \in [k]$,

$$|C_t \cap C_t^*| \geq (1 - \alpha) \cdot \max\{|C_t|, |C_t^*|\}.$$

(M2) Noisy pairwise membership queries. Given a query pair $(u, v) \in \mathcal{C} \times \mathcal{C}$, the oracle returns a bit $\Pi_{\text{pair}}(u, v) \in \{0, 1\}$ intended to indicate whether u and v are served by the same facility in S^*. Each answer is independently corrupted with probability at most $\alpha < 1/2$.

(M3) Noisy solution predictions. The predictor outputs a facility set $S \subseteq \mathcal{F}$ intended to approximate S^*. We measure prediction error in terms of the opening-cost mass of false positives/negatives:

$$\eta^+ := \sum_{i \in S \setminus S^*} f_i, \qquad \eta^- := \sum_{i \in S^* \setminus S} f_i.$$

Query complexity. In (M1), the query complexity of an algorithm is the number of *label queries* it makes to Π, where one query reveals the label $\Pi(j)$ of a chosen client $j \in \mathcal{C}$. In (M2), the query complexity is the number of *pairwise membership queries* to Π_{pair}, each of the form $(u, v) \in \mathcal{C} \times \mathcal{C}$. When discussing lower bounds for algorithms that may use both oracle types, we write q_1 for the number of M1 label queries and q_2 for the number of M2 pairwise membership queries.

3 Label Predictions

In our first learning-augmented setting, we assume access to a predictor Π that assigns a label to each client. Using this predictor, we first establish the existence of a good approximation algorithm. In particular, we prove that the predictor's output is sufficiently accurate to be used directly.

Lemma 1. *Let X_i denote the set of clients labeled as i by the predictor, and X_i^* denote the set of clients in the i-th cluster of the optimal solution. Then for $\alpha < 1/2$,*

$$\text{cost}(X_i^*, C_{X_i}) \leq \left(1 + \frac{4\alpha}{1 - 2\alpha}\right) \cdot \text{cost}(X_i^*, C_{X_i^*}).$$

It implies that the following Algorithm 1 gives $(1 + \frac{4\alpha}{1-2\alpha})$-approximation algorithm. Algorithm 1 queries the label of every client from Π, requiring $|\mathcal{C}|$ queries. However, the query complexity can be reduced, particularly when k is not too large. To achieve this, Lemma 1 should be re-evaluated: our goal is to obtain an approximate solution consistent with the predictor's output, rather than with the (hidden) optimal solution of the facility location instance. The lemma guarantees that this relaxation incurs at most a $(1 + O(\alpha))$ multiplicative approximation error.

The challenge then becomes extracting sufficient information from the predictor using a limited number of queries (e.g., sublinear in $|\mathcal{C}|$). A key idea is that uniformly sampling a subset of clients and estimating their centers can already yield strong approximations. In related clustering problems, most notably k-means, the following lemma has played a central role in the development of sampling-based algorithms, including those in (Ailon et al., 2018; Chien et al., 2018) for k-means in Euclidean spaces:

Algorithm 1 Facility Location with Label Predictions

Require: A facility location instance $(X = (\mathcal{F}, \mathcal{C}), d)$, and a client set $\mathcal{C}$ with labels from predictor Π of label error rate α
Ensure: $(1 + \frac{4\alpha}{1-2\alpha})$-approximate opened facilities
1: **for** $i = 1$ to k **do**
2: Let X_i be the set of clients in X labeled as i by Π.
3: Compute C_{X_i}.
4: **end for**
5: **Return** $C_{X_1}, \ldots, C_{X_k}$

Lemma 2 ((Inaba et al., 1994)). *For a point set $X \subset \mathbb{R}^n$, let S be the set made by sampling m points with replacement from X. Let cost(X, p) is defined as the sum of squared distance between p and the points of X. C_X is defined still same. Then for any $\delta \in (0, 1)$, the following holds:*

$$\Pr[\textit{cost}(X, C_S) \leq \left(1 + \frac{1}{\delta m}\right) \textit{cost}(X, C_X)] \geq 1 - \delta.$$

However, this lemma does not directly carry over to the facility location problem due to the presence of facility opening costs. For instance, minimizing only the connection cost $\text{conn}(X, i)$ may yield a geometrically optimal facility with prohibitively high opening cost. If, however, the sizes of all k clusters are known, then for a single cluster X the clustering cost can be written as

$$\text{cost}(X, i) = \sum_{j \in X} d(i, j) + f_i = \sum_{j \in X} \left(d(i, j) + \tfrac{f_i}{|X|} \right).$$

This representation can be viewed as a shifted distance function, enabling an argument analogous to those used in other clustering settings. Specifically, if we define $d'(i, j) := d(i, j) + \frac{f_i}{|X|}$, and sample m clients (without replacement) from X, denoted by S, then the objective function to minimize is

$$\sum_{j \in S} d'(i, j) = \sum_{j \in S} d(i, j) + \tfrac{|S|}{|X|} f_i.$$

In what follows, we show that this method yields an accurate approximation for C_X in terms of cost.

Lemma 3. *Let $X \subseteq \mathcal{C}$ be a client set, and let S be a uniformly sampled subset of X of size m (without replacement). Let C'_S be a facility minimizing the objective $\sum_{j \in S} d(i, j) + \frac{|S|}{|X|} f_i$ over $i \in \mathcal{F}$. Then, for any $\varepsilon, \delta \in (0, 1)$ and $m > \frac{18}{\delta} \log\left(\frac{k|\mathcal{F}|}{\varepsilon^2}\right)$, the following holds:*

$$\Pr[\textit{cost}(X, C'_S) > (1 + \varepsilon) \cdot \textit{cost}(X, C_X)] < \frac{\delta}{k}.$$

Therefore, if the cluster sizes in the optimal solution are known, combining Lemma 3 with Lemma 1 yields a sampling-based algorithm that achieves a $(1+O(\alpha))\cdot(1+\varepsilon)$-approximation. We now argue that a similar guarantee can be obtained by uniformly sampling clients (without replacement) from $\mathcal{C}$. Suppose we randomly select m clients and reveal their labels from Π. Let m_i denote the number of sampled clients labeled as i. Then, the ratio m_i/m serves as an estimate of the relative size of the cluster associated with label i, since it approximates $|C_i| \,/\, |\mathcal{C}|$.

The main challenge arises when clusters are small. For instance, if there exists a cluster containing only a single client, the accuracy of the size estimation depends critically on whether that client is sampled. To capture this issue, we introduce an additional parameter θ that quantifies the imbalance between cluster sizes. Let $s_{\min}$ denote the size of the smallest cluster among the k clusters. We define $\theta := \frac{|\mathcal{C}|}{k \cdot s_{\min}}$, which is the ratio between the average cluster size and the minimum cluster size. A minimum cluster size condition of this form has been widely studied in clustering problems (Mazumdar & Saha, 2017; Chien et al., 2018; Xia & Huang, 2022; Gamlath et al., 2022). For our label-prediction algorithm, note that $s_{\min}$ refers not to the smallest cluster size in the optimal solution, but rather to the smallest cluster size in the predictor's output. It is stated as follows:

Lemma 4. *Suppose there are n balls that are numbered from 1 to k. The number of balls written as i is n_i, thus $n = \sum_{i=1}^{k} n_i$. Also, for all $1 \le i \le k$, $\theta \ge \frac{n}{k n_i}$ holds for some fixed parameter θ. Let m_i be a number of balls that comes from $m > 3k\theta \log(\frac{2k}{\delta})/\varepsilon^2$ sampling without replacement. Then for any $\varepsilon, \delta \in (0, 1)$ and $1 \le i \le k$, the following holds:*

$$\Pr\left[\left|\frac{m_i}{m} - \frac{n_i}{n}\right| \le \varepsilon \frac{n_i}{n}\right] \ge 1 - \frac{\delta}{k}.$$

To summarize, our goal is to estimate the number of clients labeled i by the predictor Π. This is achieved by uniformly sampling M_1 clients without replacement, as guaranteed by Lemma 4. Next, to obtain a good approximation of the corresponding center, we apply Lemma 3, which requires sampling at least $\frac{18}{\delta} \log\left(\frac{k \cdot |\mathcal{F}|}{\varepsilon^2}\right)$ clients from each cluster. To meet this requirement, the choice of M_2 is determined accordingly.ording to the guarantees provided by the following lemma:

Lemma 5 ((Chien et al., 2018), Theorem 3.3)**.** *Suppose there are n balls that are numbered from 1 to k. The number of balls written as i is n_i, thus $n = \sum_{i=1}^{k} n_i$. Without loss of generality, $n_1 \le n_2 \le \cdots \le n_k$. Let m_i be a number of balls that comes from m sampling without replacement. Then, the probability that for each label $i \in \{1, \ldots, k\}$, at least $m_o = \frac{m}{2}\frac{n_1}{n}$ balls are sampled is bounded as:*

$$\Pr\left[m_i \ge m_o\right] \ge 1 - k \exp\left(-\frac{m_o}{4}\right).$$

Algorithm 2 Facility Location with Label Predictions: Reduced Query Complexity

Require: A facility location instance $(X = (\mathcal{F}, \mathcal{C}), d)$, a fixed parameter θ, and a client set $\mathcal{C}$ with labels from predictor Π of label error rate $\alpha < \frac{1}{2}$

Ensure: Output $(1 + \frac{4\alpha}{1-2\alpha}) \cdot (1 + \varepsilon)$-approximate opened facilities with probability at least $1 - \delta$

1: Uniformly sample $M_1 = 27k\theta \log\left(\frac{6k}{\delta}\right)$ clients without replacement from $\mathcal{C}$.
2: **for** $i = 1$ to k **do**
3: $\quad m_i \leftarrow$ the number of clients labeled as i among the M_1 sampled clients.
4: $\quad n_i \leftarrow |\mathcal{C}| \cdot \frac{m_i}{M_1}$.
5: **end for**
6: Uniformly sample $M_2 = 2k\theta \cdot \max\left(\frac{216}{\delta} \log\left(\frac{k|\mathcal{F}|}{\varepsilon^2}\right), 4 \log \frac{3k}{\delta}\right)$ clients without replacement from $\mathcal{C}$.
7: **for** $i = 1$ to k **do**
8: $\quad S_i \leftarrow$ the set of clients labeled as i among the M_2 sampled clients.
9: $\quad$ Compute $C_i = \arg\min_{i' \in \mathcal{F}} \sum_{j \in S_i} d(i', j) + \frac{|S_i|}{n_i} f_{i'}$.
10: **end for**
11: **Return** $C_1, \ldots, C_k$

We now present our first result concerning Algorithm 2:

Theorem 1. *Algorithm 2 yields a* $\left(1 + \frac{4\alpha}{1-2\alpha}\right) \cdot (1 + \varepsilon)$*-approximate solution, given access to a label predictor* $\Pi : \mathcal{C} \to [k]$ *with a label error rate* $\alpha < 1/2$*. This holds with probability at least* $1 - \delta$ *for any* $\varepsilon, \delta \in (0, 1)$*. Then, for the parameter* $\theta = \frac{|\mathcal{C}|}{k \cdot s_{min}}$*, its query complexity is at most* $M_1 + M_2 = O\left(k\theta \cdot \max\left(\frac{1}{\delta} \log \frac{k|\mathcal{F}|}{\varepsilon^2}, \log \frac{k}{\delta}\right)\right)$.

4 Pairwise Membership Queries

We now introduce another predictor Π. Given any pair of clients $u, v \in \mathcal{C}$, this predictor outputs a boolean value indicating whether u and v are served by the same facility in the optimal solution. We say that Π has an error rate α if each of its outputs is incorrect with probability at most α, independently across pairs. Note that once a query is answered, the response is fixed, and thus we cannot *boost* the correctness probability through techniques such as majority voting. This type of pairwise prediction oracle has been extensively studied in the context of other clustering problems such as k-means (Mazumdar & Saha, 2017; Chien et al., 2018; Ailon et al., 2018), though its use in the facility location setting is less well explored. Moreover, many of the existing algorithms for pairwise queries are restricted to Euclidean spaces.

We now give a brief overview of our approach. Suppose we queried all possible $\binom{|\mathcal{C}|}{2}$ client pairs and collected the outputs of Π. This represents the maximum information available from the predictor. Ignoring the facility location context, the resulting task reduces to the classical problem of recovering an unknown clustering configuration from noisy pairwise similarity information. Such problems are well studied, and we recall one relevant result below:

Lemma 6 ((Mazumdar & Saha, 2017), Theorem 2)**.** *Let* k *be a number of clusters of hidden clustering configuration between* $|\mathcal{C}|$ *clients. Then, there exists an algorithm* A *with query complexity* $O(\frac{|\mathcal{C}|k \log|\mathcal{C}|}{(1-2\alpha)^4})$ *that returns all true clusters of size at least* $\frac{64k \log|\mathcal{C}|}{(1-2\alpha)^4}$ *with probability* $1 - \frac{2}{|\mathcal{C}|}$ *for predictor error rate* $\alpha < \frac{1}{2}$*. Moreover,* k *can be unknown.*

Note that the minimum cluster size constraint in recovery guarantees is essentially necessary. If there exist two clusters of constant size, the outcomes of both intra-cluster and inter-cluster queries can be arbitrary with some probability, rendering the problem inherently unsolvable. To account for this, we again introduce the parameter $\theta = \frac{|\mathcal{C}|}{k \cdot s_{\min}}$, which is the ratio between the average cluster size and the minimum cluster size.

We now recall Lemma 8, which implies that a small number of randomly sampled clients suffices to obtain a good approximate solution. Building on this idea, our approach is as follows: we randomly

sample a subset of clients, recover the underlying clustering configuration using Lemma 6, and then, for each recovered cluster, compute the center that minimizes the cost for the sampled clients using Lemma 8. These centers are then returned as the output of the algorithm. The procedure is summarized in Algorithm 3, and the corresponding results are obtained as follows:

Algorithm 3 Facility Location with Pairwise Membership Predictions

Require: A facility location instance $(X = (\mathcal{F}, \mathcal{C}), d)$, a fixed parameter θ, and a pairwise membership predictor Π of error rate $\alpha < \frac{1}{2}$

Ensure: Output $(1+\varepsilon)$-approximate opened facilities with probability at least $1-\delta$

1: Uniformly sample $M_1 = 27k\theta \log\left(\frac{8k}{\delta}\right)$ clients without replacement from $\mathcal{C}$.
2: **for** $i = 1$ to k **do**
3: $\quad m_i \leftarrow$ the number of clients labeled as i among the M_1 sampled clients.
4: $\quad n_i \leftarrow |\mathcal{C}| \cdot \frac{m_i}{M_1}$.
5: **end for**
6: Uniformly sample $M_2 = \max\left(M_2', \frac{576k\theta}{\delta} \log\left(\frac{k|\mathcal{F}|}{\varepsilon^2}\right), 8k\theta \log\frac{4k}{\delta}\right)$ clients without replacement from $\mathcal{C}$, where M_2' is defined as the minimum integer that satisfies $\frac{M_2'}{\log M_2'} \geq \frac{128k^2\theta}{(1-2\alpha)^4}$.
7: Run the algorithm of Lemma 6 on the above M_2 clients.
8: **for** $i = 1$ to k **do**
9: $\quad S_i \leftarrow$ the i-th cluster recovered from the above step.
10: $\quad$ Compute $C_i = \arg\min_{i' \in \mathcal{F}} \sum_{j \in S_i} d(i', j) + \frac{|S_i|}{n_i} f_{i'}$.
11: **end for**
12: **Return** $C_1, \ldots, C_k$

Theorem 2. *Algorithm 3 yields a $(1+\varepsilon)$-approximate solution for the (metric) uncapacitated facility location problem, given access to a pairwise membership predictor $\Pi : \mathcal{C} \times \mathcal{C} \to \{0,1\}$ with an error rate $\alpha < 1/2$. This holds with probability at least $1-\delta$ for any $\varepsilon, \delta \in (0,1)$. Then, for the parameter $\theta = \frac{|\mathcal{C}|}{k \cdot s_{\min}}$, its query complexity is at most*

$$M_1 + M_2 = \tilde{O}\left(\max\left(\frac{k^2\theta}{(1-2\alpha)^4}, \frac{k\theta}{\delta} \log\frac{k\,|\mathcal{F}|}{\varepsilon^2}, k\theta \log\frac{k}{\delta}\right)\right).$$

5 Solution Predictions

We also consider a different type of predictor Π, which outputs a subset of facilities forming a feasible solution to the facility location problem. Given an optimal solution S^* for a facility location instance and a predicted solution S, we define false positive prediction weight $\eta^+ = \sum_{i \in S \setminus S^*} f_i$, and false negative prediction weight: $\eta^- = \sum_{i \in S^* \setminus S} f_i$. This prediction error model, denoted (η^+, η^-), was introduced in (Antoniadis et al., 2025).

To utilize this predictor, we apply a classical bifactor approximation algorithm as a black box. Suppose there exists a (λ_f, λ_c)-approximation algorithm A. Given a prediction S with errors (η^+, η^-), our approach constructs a modified instance in which the facility opening costs are adjusted as:

$$f_i' = \begin{cases} 0 & \text{if } i \in S, \\ f_i & \text{otherwise.} \end{cases}$$

Algorithm A is then executed on this modified instance. If the algorithm outputs S', we define the final solution as $S \cup S'$. Notice that in the modified instance, A incurs zero cost for opening facilities in S, so it is reasonable to assume $S \subseteq S'$. Consequently, for the purpose of analysis, we treat the final output as S' rather than $S \cup S'$. The following theorem quantifies the performance of this approach, showing how the prediction errors (η^+, η^-) impact the approximation guarantee.

Theorem 3. *Algorithm 4 yields a $(\lambda_c + \frac{\eta^+ + (\lambda_f - \lambda_c)\eta^-}{\mathrm{OPT}})$-approximation solution, given access to a solution predictor Π with error (η^+, η^-), where OPT is the optimal cost of the given instance.*

Several bifactor approximation algorithms for the (metric) UFL problem exist. As we discussed in Section 1.2, our algorithm performs more effectively when λ_c is small. Using the $(\lambda, 1 + 2e^{-\lambda})$-approximation algorithm from (Byrka & Aardal, 2010) for any $\lambda \geq 1.6774$, we obtain $(1 + \frac{2}{e^\lambda} +$

Algorithm 4 Facility Location with Solution Predictions

Require: A facility location instance $(X = (\mathcal{F}, \mathcal{C}), d)$, a (λ_f, λ_c) bifactor approximation algorithm A, and a solution predictor Π with error (η^+, η^-)

Ensure: Output $(\lambda_c + \frac{\eta^+ + (\lambda_f - \lambda_c)\eta^-}{\mathrm{OPT}})$-approximate opened facilities

$S \leftarrow$ Obtain the solution S from predictor Π.
for $i \in \mathcal{F}$ **do**
  $f'_i \leftarrow 0$ **if** $i \in S$ **else** f_i.
end for
$S' \leftarrow$ the solution generated by running A on instance X using the modified costs $\{f'_i\}_{i \in \mathcal{F}}$.
Return $S \cup S'$

$\frac{\eta^+ + (\lambda - 1 - \frac{2}{e^\lambda})\eta^-}{\mathrm{OPT}})$-approximation algorithm. Therefore, in light of the 1.463 approximation hardness of the UFL problem (Guha & Khuller, 1999), and for sufficiently large λ, our learning-augmented algorithm has the potential to significantly outperform classical approaches.

6 Lower Bounds

This section presents lower bound results that complement the algorithms for the three prediction models studied above. The lower bounds are based on the bounded-degree Vertex Cover problem.

Lemma 7 ((Chlebík & Chlebíková, 2006), ETH-based 4-regular Vertex Cover gap)**.** *Assuming ETH, there exist constants $0 < a < b < 1$ such that, for every $\zeta \in (0, 1)$, no algorithm running in time $2^{O(n^{1-\zeta})}$ can distinguish the following two cases for an n-vertex 4-regular graph G:* ***(YES)*** *G has a vertex cover of size at most an,* ***(NO)*** *every vertex cover of G has size at least bn.*

We next describe the hard UFL instance obtained from this vertex-cover gap. The UFL instance will be called a YES instance or a NO instance according to whether the underlying graph G is in the YES case or the NO case of Lemma 7. Given a 4-regular graph $G = (V, E)$ with $|V| = n$ and $|E| = 2n$, fix a constant $\lambda \in (0, 2)$. We construct a metric UFL instance as follows. For every vertex $v \in V$, create a facility with opening cost λ, and for every edge $e \in E$, create a client. A client corresponding to edge e is at distance 1 from the two endpoint facilities of e, and at distance 3 from every other facility. Distances between two distinct facilities, and between two distinct clients, are set to 2. The triangle inequality follows immediately.

We identify facilities with their corresponding vertices. For any opened facility set $S \subseteq \mathcal{F}$, let $u(S)$ denote the number of edges not covered by the corresponding vertex set. Then the cost of S in the constructed instance is $\mathrm{cost}(S) = \lambda|S| + 2n + 2u(S)$. Since $\lambda < 2$, an optimal UFL solution cannot leave an edge uncovered: opening one endpoint facility decreases the connection cost of the uncovered edge-client by 2 and costs only λ. Hence every optimum corresponds to a minimum vertex cover, and $\mathrm{OPT}(I_G) = 2n + \lambda\tau(G)$, where $\tau(G)$ is the minimum vertex cover size. Consequently, if $B = (2 + \lambda a)n$ and $1 + \beta_0 := (2 + \lambda b)/(2 + \lambda a)$, then $\beta_0 > 0$, and distinguishing $\mathrm{OPT}(I_G) \le B$ from $\mathrm{OPT}(I_G) \ge (1 + \beta_0)B$ in time $2^{O(n^{1-\zeta})}$ would contradict Lemma 7. This construction yields the query lower bound in Theorem 4 for (M1) and (M2), even with exact oracles, and the prediction-error lower bound in Theorem 5 for (M3).

Theorem 4. *For the hard instances constructed from n-vertex graphs above, for every $\zeta \in (0, 1)$ and every $\beta \in (0, \beta_0)$, no algorithm can output a $(1+\beta)$-approximate UFL solution in time $2^{O(n^{1-\zeta})}$ while making q_1 M1 label queries and q_2 M2 pairwise queries satisfying*

$$q_1 \log k + q_2 = O(k^{1-\zeta}),$$

where k is the number of facilities in the fixed optimum. In particular, an M1-only algorithm requires $q_1 = \Omega(k^{1-\zeta}/\log k)$ queries, and an M2-only algorithm requires $q_2 = \Omega(k^{1-\zeta})$ queries.

Theorem 5. *For every $\zeta \in (0, 1)$ and every $\rho < 1 + \beta_0$, under ETH, no algorithm running in time $2^{O(n^{1-\zeta})}$ on the hard instances constructed from n-vertex graphs can provide the following guarantee for every metric UFL instance I and every predicted facility set $S \subseteq \mathcal{F}$:*

$$\mathrm{cost}(\mathcal{A}(I, S)) \le \rho\ \mathrm{OPT}(I) + r_-(\eta^-), \mathrm{cost}(\mathcal{A}(I, S)) \le \rho\ \mathrm{OPT}(I) + r_+(\eta^+)$$

with $r_-(x) = o(x)$ and $r_+(x) = o(x)$.

AI USE STATEMENT

We used generative AI tools for proofreading, editing, and formatting the manuscript. They were not used for solving the research problems or developing the proofs.

REFERENCES

Anders Aamand, Justin Y. Chen, Siddharth Gollapudi, Sandeep Silwal, and Hao Wu. Improved approximations for hard graph problems using predictions. *CoRR*, abs/2505.23967, 2025. doi: 10.48550/ARXIV.2505.23967. URL https://doi.org/10.48550/arXiv.2505.23967.

Nir Ailon, Anup Bhattacharya, Ragesh Jaiswal, and Amit Kumar. Approximate Clustering with Same-Cluster Queries. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 40:1–40:21, 2018.

Spyros Angelopoulos, Christoph Dürr, Shendan Jin, Shahin Kamali, and Marc Renault. Online computation with untrusted advice. *Journal of Computer and System Sciences*, 144:103545, 2024.

Antonios Antoniadis, Marek Elias, Adam Polak, and Moritz Venzin. Approximation algorithms for combinatorial optimization with predictions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AEFVa6VMu1.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1027–1035, 2007.

Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3216–3224, 2016.

Yossi Azar, Stefano Leonardi, and Noam Touitou. Flow time scheduling with uncertain processing time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1070–1080, 2021.

Yossi Azar, Debmalya Panigrahi, and Noam Touitou. Online graph algorithms with predictions. In Joseph (Seffi) Naor and Niv Buchbinder (eds.), *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pp. 35–66. SIAM, 2022.

Xingjian Bai and Christian Coester. Sorting with predictions. *Advances in Neural Information Processing Systems*, 36:26563–26584, 2023.

Nikhil Bansal, Christian Coester, Ravi Kumar, Manish Purohit, and Erik Vee. Learning-augmented weighted paging. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 67–89. SIAM, 2022.

Hadley Black, Euiwoong Lee, Arya Mazumdar, and Barna Saha. Clustering with non-adaptive subset queries. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Avrim Blum and Vaidehi Srinivas. Competitive strategies to use "warm start" algorithms with predictions. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 3775–3801. SIAM, 2025.

Jan van den Brand, Sebastian Forster, Yasamin Nazari, and Adam Polak. On dynamic graph algorithms with predictions. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 3534–3557. SIAM, 2024.

Vladimir Braverman, Prathamesh Dharangutte, Vihan Shah, and Chen Wang. Learning-augmented maximum independent set. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2024, August 28-30, 2024, London School of Economics, London, UK*, volume 317 of *LIPIcs*, pp. 24:1–24:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.

Jaroslaw Byrka and Karen Aardal. An optimal bifactor approximation algorithm for the metric uncapacitated facility location problem. *SIAM Journal on Computing*, 39(6):2212–2231, 2010.

I. Eli Chien, Chao Pan, and Olgica Milenkovic. Query k-means clustering and the double dixie cup problem. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6650–6659, 2018.

Miroslav Chlebík and Janka Chlebíková. Complexity of approximating bounded variants of optimization problems. *Theoretical Computer Science*, 354(3):320–338, 2006.

Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. On approximability of clustering problems without candidate centers. In Dániel Marx (ed.), *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pp. 2635–2648. SIAM, 2021.

Vincent Cohen-Addad, Hossein Esfandiari, Vahab S. Mirrokni, and Shyam Narayanan. Improved approximations for euclidean k-means and k-median, via nested quasi-independent sets. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pp. 1621–1628. ACM, 2022a.

Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. Johnson coverage hypothesis: Inapproximability of k-means and k-median in l_p-metrics. In *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pp. 1493–1530. SIAM, 2022b.

Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pp. 1038–1051. ACM, 2022c.

Vincent Cohen-Addad, Tommaso d'Orsi, Anupam Gupta, Euiwoong Lee, and Debmalya Panigrahi. Learning-augmented approximation algorithms for maximum cut and related problems. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Vincent Cohen-Addad, Fabrizio Grandoni, Euiwoong Lee, Chris Schwiegelshohn, and Ola Svensson. A (2+ϵ)-approximation algorithm for metric k-median. In Michal Koucký and Nikhil Bansal (eds.), *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23-27, 2025*, pp. 615–624. ACM, 2025.

Michael Dinitz, Sungjin Im, Thomas Lavastida, Benjamin Moseley, and Sergei Vassilvitskii. Faster matchings via learned duals. *Advances in Neural information processing systems*, 34:10393–10406, 2021.

Michael Dinitz, Sungjin Im, Thomas Lavastida, Ben Moseley, Aidin Niaparast, and Sergei Vassilvitskii. Binary search with distributional predictions. *Advances in Neural Information Processing Systems*, 37:90456–90472, 2024.

Jon C. Ergun, Zhili Feng, Sandeep Silwal, David P. Woodruff, and Samson Zhou. Learning-augmented k-means clustering. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Chenglin Fan and Kijun Shin. Learning augmented graph k-clustering. In *The Thirty Eighth Annual Conference on Learning Theory, 30-4 July 2025, Lyon, France*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1919–1934. PMLR, 2025.

Basilios Gamlath, Konstantin Makarychev, Yury Makarychev, and Arya Mazumdar. Approximate cluster recovery from noisy labels. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT)*, pp. 2456–2494, 2022.

Suprovat Ghoshal, Konstantin Makarychev, and Yury Makarychev. Constraint satisfaction problems with advice. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2025, New Orleans, LA, USA, January 12-15, 2025*, pp. 1202–1221. SIAM, 2025.

Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *Journal of Algorithms*, 31(1):228–248, 1999.

Dorit S. Hochbaum. Heuristics for the fixed cost median problem. *Mathematical Programming*, 22 (1):148–162, 1982.

Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering (extended abstract). In *Proceedings of the Tenth Annual Symposium on Computational Geometry*, pp. 332–339, 1994.

Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 731–740, 2002.

Shi Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. *Information and Computation*, 222:45–58, 2013.

Alexander Lindermayr and Nicole Megow. Permutation predictions for non-clairvoyant scheduling. In *Proceedings of the 34th ACM Symposium on Parallelism in Algorithms and Architectures*, pp. 357–368, 2022.

Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Pinyan Lu, Xuandi Ren, Enze Sun, and Yubo Zhang. Generalized sorting with predictions. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 111–117. SIAM, 2021.

Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *Journal of the ACM (JACM)*, 68(4):1–25, 2021.

Arya Mazumdar and Barna Saha. Clustering with noisy queries. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5790–5801, 2017.

Thy Dinh Nguyen, Anamay Chaturvedi, and Huy L. Nguyen. Improved learning-augmented algorithms for k-means and k-medians clustering. *International Conference on Learning Representations (ICLR)*, 2023.

Pan Peng and Jiawei Zhang. Towards a query-optimal and time-efficient algorithm for clustering with a faulty oracle. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3662–3680. PMLR, 2021.

Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving online algorithms via ml predictions. *Advances in Neural Information Processing Systems*, 31, 2018.

Dhruv Rohatgi. Near-optimal bounds for online caching with machine learned advice. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1834–1845. SIAM, 2020.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 1177–1178. ACM, 2010. doi: 10.1145/1772690.1772862.

David B. Shmoys, Eva Tardos, and Karen Aardal. Approximation algorithms for facility location problems. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC '97)*, pp. 265–274, New York, NY, USA, 1997. ACM.

Sandeep Silwal. Learning-augmented algorithms. *Massachusetts Institute of Technology*, 2021.

Jinghui Xia and Zengfeng Huang. Optimal clustering with noisy queries via multi-armed bandit. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24315–24331. PMLR, 2022.

Rui Xu and Donald Wunsch. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2005.

A Proof of Lemma 1

Lemma 1 Let X_i denote the set of clients labeled as i by the predictor, and X_i^* denote the set of clients in the i-th cluster of the optimal solution. Then for $\alpha < 1/2$,

$$\mathrm{cost}(X_i^*, C_{X_i}) \leq \left(1 + \frac{4\alpha}{1-2\alpha}\right) \cdot \mathrm{cost}(X_i^*, C_{X_i^*}).$$

Our proof builds upon and extends the result of (Fan & Shin, 2025), which was originally developed for k-clustering problems such as k-means and k-median.

Proof. Since label names are arbitrary, we abuse notation by dropping subscripts. Let $D := d(C_X, C_{X^*})$. Then, by the triangle inequality, the following holds:

$$\begin{aligned} |X \cap X^*| \cdot D = \sum_{j \in X \cap X^*} d(C_X, C_{X^*}) &\leq \sum_{j \in X \cap X^*} d(C_X, j) + \sum_{j \in X \cap X^*} d(C_{X^*}, j) \\ &= \mathrm{conn}(X \cap X^*, C_X) + \mathrm{conn}(X \cap X^*, C_{X^*}). \end{aligned} \tag{1}$$

Similarly, applying the triangle inequality $d(C_{X^*}, j) \leq d(C_X, j) + D$ for any $j \in X \setminus X^*$, we obtain:

$$\mathrm{conn}(X \cap X^*, C_X) \leq \mathrm{conn}(X \cap X^*, C_{X^*}) + |X \setminus X^*| \cdot D. \tag{2}$$

Summing (1) and (2) yields the following upper bound for D:

$$(|X \cap X^*| - |X \setminus X^*|) \cdot D \leq 2 \cdot \mathrm{conn}(X \cap X^*, C_{X^*}).$$

From the definition of error rate α, $|X \cap X^*| \geq (1-\alpha) \cdot \max(|X|, |X^*|)$ holds. Therefore, from the above inequality, we get:

$$D \leq \frac{2 \cdot \mathrm{conn}(X \cap X^*, C_{X^*})}{(1-2\alpha)\max(|X|, |X^*|)}. \tag{3}$$

On the other hand, by the definition of the center C_X, $\mathrm{cost}(X, C_X) \leq \mathrm{cost}(X, C_{X^*})$ holds. Given that $X = (X \cap X^*) \cup (X \setminus X^*)$, we obtain:

$$\begin{aligned} &\mathrm{cost}(X \cap X^*, C_X) + \mathrm{conn}(X \setminus X^*, C_X) \\ &= \mathrm{cost}(X, C_X) \\ &\leq \mathrm{cost}(X, C_{X^*}) \\ &= \mathrm{cost}(X \cap X^*, C_{X^*}) + \mathrm{conn}(X \setminus X^*, C_{X^*}). \end{aligned}$$

Then applying the triangle inequality $d(C_X, j) - d(C_{X^*}, j) \leq D$ for any $j \in X \setminus X^*$, we obtain:

$$\mathrm{cost}(X \cap X^*, C_X) \leq \mathrm{cost}(X \cap X^*, C_{X^*}) + |X \setminus X^*| \cdot D. \tag{4}$$

Therefore, the final results can be derived as follows:

$$\begin{aligned} \mathrm{cost}(X^*, C_X) &= \mathrm{cost}(X \cap X^*, C_X) + \mathrm{conn}(X^* \setminus X, C_X) \\ &\leq \mathrm{cost}(X \cap X^*, C_X) + \mathrm{conn}(X^* \setminus X, C_{X^*}) \\ &\quad + |X^* \setminus X| \cdot D \quad (\because \text{triangle inequality}) \\ &\leq \mathrm{cost}(X \cap X^*, C_{X^*}) + |X \setminus X^*| \cdot D \\ &\quad + \mathrm{conn}(X^* \setminus X, C_{X^*}) + |X^* \setminus X| \cdot D \quad (\because (4)) \\ &\leq \mathrm{cost}(X^*, C_{X^*}) + 2\alpha \cdot \max(|X|, |X^*|) \cdot D \quad (\because \text{definition of } \alpha) \\ &\leq \mathrm{cost}(X^*, C_{X^*}) + \frac{4\alpha}{1-2\alpha} \cdot \mathrm{conn}(X \cap X^*, C_{X^*}) \quad (\because (3)) \\ &\leq \left(1 + \frac{4\alpha}{1-2\alpha}\right) \cdot \mathrm{cost}(X^*, C_{X^*}). \end{aligned}$$

□

B Proof of Lemma 3

Lemma 3. Let $X \subseteq \mathcal{C}$ be a client set, and let S be a uniformly sampled subset of X of size m (without replacement). Let C'_S be a facility minimizing the objective $\sum_{j\in S} d(i,j) + \frac{|S|}{|X|} f_i$ over $i \in \mathcal{F}$. Then, for any $\varepsilon, \delta \in (0,1)$ and $m > \frac{18}{\delta} \log\left(\frac{k|\mathcal{F}|}{\varepsilon^2}\right)$, the following holds:

$$\Pr[\text{cost}(X, C'_S) > (1+\varepsilon) \cdot \text{cost}(X, C_X)] < \frac{\delta}{k}.$$

Proof. Let $\mathcal{F}_{\text{bad}}$ be a set of "bad" facilities, defined as

$$\mathcal{F}_{\text{bad}} = \{i \in \mathcal{F} \mid \text{cost}(X,i) > (1+\varepsilon) \cdot \text{cost}(X, C_X)\}.$$

Therefore, $C'_S \in \mathcal{F}_{\text{bad}}$ is equivalent to $\text{cost}(X, C'_S) > (1+\varepsilon) \cdot \text{cost}(X, C_X)$. For any fixed $i \in \mathcal{F}_{\text{bad}}$, if i minimizes the objective $\sum_{j\in S} d(i,j) + \frac{|S|}{|X|} f_i$, then it implies:

$$\sum_{j\in S} d(i,j) + \frac{|S|}{|X|} f_i < \sum_{j\in S} d(C_X, j) + \frac{|S|}{|X|} f_{C_X}.$$

This is equivalent to $\sum_{j\in S} W_j \leq 0$, where we define a random variable W_j for each client $j \in X$ corresponding to the pair of facilities (i, C_X):

$$W_j = \left(d(i,j) + \frac{f_i}{|X|}\right) - \left(d(C_X, j) + \frac{f_{C_X}}{|X|}\right).$$

We will use Hoeffding's inequality to show that the probability of this event is small. Let μ_W denote the average of the W_j values over the entire set X:

$$\begin{aligned}
\mu_W &= \frac{1}{|X|} \sum_{j\in X} W_j \\
&= \frac{1}{|X|} \left(\sum_{j\in X} \big(d(i,j) - d(C_X, j)\big) + \big(f_i - f_{C_X}\big) \right) \\
&= \frac{\text{cost}(X,i) - \text{cost}(X, C_X)}{|X|}.
\end{aligned}$$

For brevity, let $\text{OPT} := \text{cost}(X, C_X)$. Since $i \in \mathcal{F}_{\text{bad}}$, we have $\text{cost}(X,i) > (1+\varepsilon)\,\text{OPT}$, which implies

$$\mu_W > \frac{\varepsilon \cdot \text{OPT}}{|X|}.$$

Hence, the population mean is strictly positive. We are interested in the probability that the sample mean

$$\overline{W}_S := \frac{1}{|S|} \sum_{j\in S} W_j$$

is non-positive. To apply Hoeffding's inequality, it remains to bound the range of the random variables W_j.

It can be done by the property of metric space. By triangle inequality, $|d(i,j) - d(C_X, j)| \leq d(i, C_X)$. The term $(f_i - f_{C_X})/|X|$ is a constant shift. Thus, the total range of W_j over all $j \in X$ is $R_i := \max_j W_j - \min_j W_j \leq 2 \cdot d(i, C_X)$.

Moreover, we can bound $d(i, C_X)$ using the triangle inequality as follows:

$$|X| \cdot d(i, C_X) = \sum_{j\in X} d(i, C_X) \leq \sum_{j\in X} (d(i,j) + d(C_X, j)) = \text{conn}(X,i) + \text{conn}(X, C_X).$$

Since facility opening costs are non-negative, we have $\text{conn}(X,i) \leq \text{cost}(X,i)$ and $\text{conn}(X, C_X) \leq \text{OPT}$. Thus,

$$|X| \cdot d(i, C_X) \leq \text{cost}(X,i) + \text{OPT}\,.$$

This gives a bound on the range: $R_i \leq \frac{2(\text{cost}(X,i)+\text{OPT})}{|X|}$.

Now we can apply Hoeffding's inequality to the sample mean $\overline{W}_S$:

$$\Pr[\overline{W}_S \leq 0] = \Pr[\overline{W}_S - \mu_W \leq -\mu_W] \leq \exp\left(\frac{-2\,|S|\,\mu_W^2}{R_i^2}\right).$$

We need to find a uniform lower bound for the term μ_W^2/R_i^2 that is independent of the specific bad facility i.

$$\frac{\mu_W^2}{R_i^2} \geq \frac{\left(\frac{\text{cost}(X,i)-\text{OPT}}{|X|}\right)^2}{\left(\frac{2(\text{cost}(X,i)+\text{OPT})}{|X|}\right)^2} = \frac{(\text{cost}(X,i)-\text{OPT})^2}{4(\text{cost}(X,i)+\text{OPT})^2}.$$

Let $\text{cost}(X,i) = (1+\alpha)\cdot\text{OPT}$. Since i is a bad facility, $\alpha > \varepsilon$. Then the expression becomes:

$$\frac{((1+\alpha)\,\text{OPT}-\text{OPT})^2}{4((1+\alpha)\,\text{OPT}+\text{OPT})^2} = \frac{\alpha^2}{4(2+\alpha)^2}.$$

The function $g(\alpha) = \frac{\alpha}{2+\alpha}$ is monotonically increasing for $\alpha > 0$. Therefore, its value is minimized as α approaches its lower bound ε. This gives us a uniform lower bound:

$$\frac{\mu_W^2}{R_i^2} > \frac{\varepsilon^2}{4(2+\varepsilon)^2}.$$

This bound is independent of i. Therefore, for any $i \in \mathcal{F}_{\text{bad}}$, the probability that it is preferred over C_X in the sample is bounded by:

$$\Pr\left[\sum_{j\in S} W_j \leq 0\right] < \exp\left(\frac{-m\varepsilon^2}{2(2+\varepsilon)^2}\right).$$

Therefore, the probability $C_S \in \mathcal{F}_{\text{bad}}$ happens is bounded by union bound for all facilities $\mathcal{F}$, that is:

$$|\mathcal{F}|\cdot\exp\left(\frac{-m\varepsilon^2}{2(2+\varepsilon)^2}\right) < \frac{\delta}{k}.$$

For any $\varepsilon \in (0,1)$, solving for m yields:

$$m > \frac{18\log(\frac{k|\mathcal{F}|}{\delta})}{\varepsilon^2}.$$

Choosing m of at least this size guarantees that with probability at least $1-\delta/k$, the facility C_S selected by the sample is not a bad facility. This implies $\text{cost}(X,C_S) \leq (1+\varepsilon)\cdot\text{cost}(X,C_X)$, which completes the proof. □

C Proof of Lemma 4

Lemma 4 Suppose there are n balls that are numbered from 1 to k. The number of balls written as i is n_i, thus $n = \sum_{i=1}^k n_i$. Also, for all $1 \leq i \leq k$, $\theta \geq \frac{n}{kn_i}$ holds for some fixed parameter θ. Let m_i be a number of balls that comes from $m > 3k\theta\log(\frac{2k}{\delta})/\varepsilon^2$ sampling without replacement. Then for any $\varepsilon, \delta \in (0,1)$, the following holds:

$$\Pr\left[\left|\frac{m_i}{m} - \frac{n_i}{n}\right| \leq \varepsilon\frac{n_i}{n}\right] \geq 1 - \frac{\delta}{k}$$

for all $1 \leq i \leq k$.

Proof. Let $p_i = n_i/n$ be the true proportion of balls of type i in the population. The random variable m_i represents the number of balls of type i in a sample of size m drawn with replacement.

The expected value of m_i is $\mathbb{E}[m_i] = m \cdot p_i$. Substituting p_i and $\mathbb{E}[m_i]$ into the inequality, we get $\left|\frac{m_i}{m} - p_i\right| \leq \varepsilon p_i$. It is equivalent to $|m_i - \mathbb{E}[m_i]| \leq \varepsilon \mathbb{E}[m_i]$.

We want to prove that this holds for all $i \in \{1, \ldots, k\}$ with probability at least $1 - \delta$. If we can show that for all $1 \leq i \leq k$, the event $|m_i - \mathbb{E}[m_i]| > \varepsilon \mathbb{E}[m_i]$ happens not more than $\frac{\delta}{k}$, then by union bound the total failure probability will be bounded by δ.

The Chernoff bound for a sum of m independent Bernoulli trials X with expectation $\mu = \mathbb{E}[X]$ and for $0 < \varepsilon < 1$ is:

$$\Pr\left[|X - \mu| > \varepsilon\mu\right] \leq 2\exp\left(-\frac{\mu\varepsilon^2}{3}\right).$$

Applying this bound to our random variable m_i with expectation $\mu_i = \mathbb{E}[m_i] = mp_i$, we get:

$$2\exp\left(-\frac{mp_i\varepsilon^2}{3}\right) \leq \frac{\delta}{k}$$

which implies $m \geq \frac{3\log(\frac{2k}{\delta})}{p_i\varepsilon^2}$. From the condition $\theta n_i \geq n/k$, we get $p_i \geq \frac{1}{k\theta}$. This completes the proof. □

D Proof of Lemma 8

A subtle issue arises regarding whether it is permissible to use an *estimated* cluster size, rather than the true cluster size, when applying Lemma 3. We show that this substitution is indeed valid, at the expense of only a constant-factor increase in the required sample size. The following lemma formalizes this insight, representing a slight refinement of Lemma 3:

Lemma 8. *Let S be a uniformly sampled subset of $X \subseteq \mathcal{C}$ of size m (without replacement). Let C'_S be the facility that minimizes the objective $\sum_{j\in S} d(i,j) + \frac{|S|}{x} f_i$ with respect to i, where x is an estimation of $|X|$ such that $\left|\frac{x}{|X|} - 1\right| < \varepsilon'$. Then, for any $\varepsilon, \varepsilon', \delta \in (0, \frac{1}{2})$, and $m > \frac{72}{\delta}\log\left(\frac{k|\mathcal{F}|}{\varepsilon^2}\right)$, the following holds:*

$$\Pr[\textit{cost}(X, C'_S) > (1+\varepsilon)\cdot \textit{cost}(X, C_X)] < \frac{\delta}{k}.$$

Proof. Let $\mathcal{F}_{\text{bad}}$ be a set of "bad" facilities, defined as

$$\mathcal{F}_{\text{bad}} = \{i \in \mathcal{F} \mid \text{cost}(X, i) > (1+\varepsilon)\cdot \text{cost}(X, C_X)\}.$$

Therefore, $C'_S \in \mathcal{F}_{\text{bad}}$ is the equivalent to $\text{cost}(X, C'_S) > (1+\varepsilon)\cdot\text{cost}(X, C_X)$. For any fixed $i \in \mathcal{F}_{\text{bad}}$, if i minimizes an objective $\sum_{j\in S} d(i,j) + \frac{|S|}{|X|} f_i$, then it implies:

$$\sum_{j\in S} d(i,j) + \frac{|S|}{|X|} f_i < \sum_{j\in S} d(C_X, j) + \frac{|S|}{|X|} f_{C_X}.$$

This is equivalent to $\sum_{j\in S} W_j \leq 0$, where we define a random variable W_j for each client $j \in X$ corresponding to the pair of facilities (i, C_X):

$$W_j = \left(d(i,j) + \frac{f_i}{|X|}\right) - \left(d(C_X, j) + \frac{f_{C_X}}{|X|}\right).$$

We will use Hoeffding's inequality to show that the probability of this event is small. Let μ_W be the average of the W_j values over the entire set X.

$$\begin{aligned}
\mu_W &= \frac{1}{|X|}\sum_{j\in X} W_j \\
&= \frac{1}{|X|}\left(\sum_{j\in X}(d(i,j) - d(C_X, j)) + (f_i - f_{C_X})\right) \\
&= \frac{\text{cost}(X, i) - \text{cost}(X, C_X)}{|X|}.
\end{aligned}$$

Let $\mathrm{OPT} := \mathrm{cost}(X, C_X)$ for brevity. Since $i \in \mathcal{F}_{\text{bad}}$, we have $\mathrm{cost}(X, i) > (1+\varepsilon)\,\mathrm{OPT}$, which implies $\mu_W > \frac{\varepsilon \cdot \mathrm{OPT}}{|X|}$. The population mean is strictly positive. We are interested in the probability that the sample mean $\overline{W}_S := \frac{1}{|S|}\sum_{j\in S} W_j$ is non-positive. To apply Hoeffding's inequality, we must bound the range of the variables W_j.

It can be done by the property of metric space. By triangle inequality, $|d(i,j) - d(C_X, j)| \le d(i, C_X)$. The term $(f_i - f_{C_X})/|X|$ is a constant shift. Thus, the total range of W_j over all $j \in X$ is $R_i := \max_j W_j - \min_j W_j \le 2 \cdot d(i, C_X)$.

Moreover, we can bound $d(i, C_X)$ using the triangle inequality as follows:

$$|X| \cdot d(i, C_X) = \sum_{j\in X} d(i, C_X) \le \sum_{j\in X} (d(i,j) + d(C_X, j)) = \mathrm{conn}(X, i) + \mathrm{conn}(X, C_X).$$

Since facility opening costs are non-negative, we have $\mathrm{conn}(X, i) \le \mathrm{cost}(X, i)$ and $\mathrm{conn}(X, C_X) \le \mathrm{OPT}$. Thus,

$$|X| \cdot d(i, C_X) \le \mathrm{cost}(X, i) + \mathrm{OPT}\,.$$

This gives a bound on the range: $R_i \le \frac{2(\mathrm{cost}(X,i)+\mathrm{OPT})}{|X|}$.

Now we can apply Hoeffding's inequality to the sample mean $\overline{W}_S$:

$$\Pr[\overline{W}_S \le 0] = \Pr[\overline{W}_S - \mu_W \le -\mu_W] \le \exp\left(\frac{-2\,|S|\,\mu_W^2}{R_i^2}\right).$$

We need to find a uniform lower bound for the term μ_W^2/R_i^2 that is independent of the specific bad facility i.

$$\frac{\mu_W^2}{R_i^2} \ge \frac{\left(\frac{\mathrm{cost}(X,i)-\mathrm{OPT}}{|X|}\right)^2}{\left(\frac{2(\mathrm{cost}(X,i)+\mathrm{OPT})}{|X|}\right)^2} = \frac{(\mathrm{cost}(X,i)-\mathrm{OPT})^2}{4(\mathrm{cost}(X,i)+\mathrm{OPT})^2}.$$

Let $\mathrm{cost}(X, i) = (1+\alpha)\cdot \mathrm{OPT}$. Since i is a bad facility, $\alpha > \varepsilon$. Then the expression becomes:

$$\frac{((1+\alpha)\,\mathrm{OPT} - \mathrm{OPT})^2}{4((1+\alpha)\,\mathrm{OPT} + \mathrm{OPT})^2} = \frac{\alpha^2}{4(2+\alpha)^2}.$$

The function $g(\alpha) = \frac{\alpha}{2+\alpha}$ is monotonically increasing for $\alpha > 0$. Therefore, its value is minimized as α approaches its lower bound ε. This gives us a uniform lower bound:

$$\frac{\mu_W^2}{R_i^2} > \frac{\varepsilon^2}{4(2+\varepsilon)^2}.$$

This bound is independent of i. Therefore, for any $i \in \mathcal{F}_{\text{bad}}$, the probability that it is preferred over C_X in the sample is bounded by:

$$\Pr\left[\sum_{j\in S} W_j \le 0\right] < \exp\left(\frac{-m\varepsilon^2}{2(2+\varepsilon)^2}\right).$$

Therefore, the probability $C_S \in \mathcal{F}_{\text{bad}}$ happens is bounded by union bound for all facilities $\mathcal{F}$, that is:

$$|\mathcal{F}| \cdot \exp\left(\frac{-m\varepsilon^2}{2(2+\varepsilon)^2}\right) < \frac{\delta}{k}.$$

For any $\varepsilon \in (0, 1)$, solving for m yields:

$$m > \frac{18\log(\frac{k|\mathcal{F}|}{\delta})}{\varepsilon^2}.$$

Choosing m of at least this size guarantees that with probability at least $1 - \delta/k$, the facility C_S selected by the sample is not a bad facility. This implies $\mathrm{cost}(X, C_S) \le (1+\varepsilon)\cdot \mathrm{cost}(X, C_X)$, which completes the proof. □

E Proof of Theorem 1

Theorem 1 Algorithm 2 yields a $\left(1+\frac{4\alpha}{1-2\alpha}\right)\cdot(1+\varepsilon)$-approximate solution, given access to a label predictor $\Pi:\mathcal{C}\to[k]$ with a label error rate $\alpha<1/2$. This holds with probability at least $1-\delta$ for any $\varepsilon,\delta\in(0,1)$. Then, for the parameter $\theta=\frac{|\mathcal{C}|}{k\cdot s_{min}}$, its query complexity is at most $M_1+M_2=O\left(k\theta\cdot\max\left(\frac{1}{\delta}\log\frac{k|\mathcal{F}|}{\varepsilon^2},\log\frac{k}{\delta}\right)\right)$.

Proof. In the initial step, we uniformly sample M_1 clients to estimate the cluster sizes labeled by Π. Lemma 4 guarantees that, with M_1 client samples, we can estimate all cluster sizes with a multiplicative error of at most $1/3$. This holds with probability at least $1-\frac{\delta}{3}$, as ensured by a union bound over all k clusters. Subsequently, the algorithm samples M_2 clients. By Lemma 5, with probability at least $1-k\exp\left(-\frac{m_o}{4}\right)\geq 1-\frac{\delta}{3}$, the number of clients sampled from each cluster is at least:

$$m_o=\frac{M_2}{2}\frac{s_{min}}{|\mathcal{C}|}\geq\frac{M_2}{2k\theta}.$$

Finally, given that the lower bound for m_o satisfies:

$$m_o\geq 2k\theta\cdot\frac{216}{\delta}\log\left(\frac{k\,|\mathcal{F}|}{\varepsilon^2}\right),$$

it follows from Lemma 8 that, with probability at least $1-\frac{\delta}{3}$, we obtain a $(1+\varepsilon)$-approximate center for all k clusters. This concludes the proof. □

F Proof of Theorem 2

Theorem 2 Algorithm 3 yields a $(1+\varepsilon)$-approximate solution for (metric) uncapacitated facility location problem, given access to a pairwise membership predictor $\Pi:\mathcal{C}\times\mathcal{C}\to\{0,1\}$ with an error rate $alpha<1/2$. This holds with probability at least $1-\delta$ for any $\varepsilon,\delta\in(0,1)$. Then, for the parameter $\theta=\frac{|\mathcal{C}|}{k\cdot s_{min}}$, its query complexity is at most

$$M_1+M_2=\tilde{O}\left(\max\left(\frac{k^2\theta}{(1-2\alpha)^4},\frac{k\theta}{\delta}\log\frac{k\,|\mathcal{F}|}{\varepsilon^2},k\theta\log\frac{k}{\delta}\right)\right)$$

.

Proof. In the initial step, we uniformly sample M_1 clients to estimate the cluster sizes labeled by Π. Lemma 4 guarantees that, with M_1 client samples, we can estimate all cluster sizes with a multiplicative error of at most $1/3$. This holds with probability at least $1-\frac{\delta}{4}$, as ensured by a union bound over all k clusters.

Subsequently, the algorithm samples M_2 clients. By Lemma 5, with probability at least $1-k\exp\left(-\frac{m_o}{4}\right)\geq 1-\frac{\delta}{4}$, the number of clients sampled from each cluster is at least:

$$m_o=\frac{M_2}{2}\frac{s_{min}}{|\mathcal{C}|}\geq\frac{M_2}{2k\theta}.$$

Given that the lower bound for m_o satisfies:

$$m_o\geq\frac{64k\log M_2}{(1-2\alpha)^4},$$

it follows from Lemma 6 that, with probability at least $1-\frac{2}{M_2}\geq 1-\frac{\delta}{4}$, we successfully recover all k clusters among M_2 clients.

Lastly, the lower bound for m_o also satisfies:

$$m_o\geq 2k\theta\cdot\frac{288}{\delta}\log\left(\frac{k\,|\mathcal{F}|}{\varepsilon^2}\right),$$

it follows from Lemma 8 that, with probability at least $1-\frac{\delta}{4}$, a $(1+\varepsilon)$-approximate center is obtained for all k clusters.

Since each of these events occurs with probability at least $1-\frac{\delta}{4}$, a union bound implies that all events occur simultaneously with probability at least $1-\delta$. This concludes the proof. □

G Proof of Theorem 3

Theorem 3 Algorithm 4 yields a $(\lambda_c + \frac{\eta^+ + (\lambda_f - \lambda_c)\eta^-}{\mathrm{OPT}})$-approximation solution, given access to a solution predictor Π with error (η^+, η^-), where OPT is the optimal cost of the given instance.

Proof. When the instance with $\{f'_i\}_{i\in\mathcal{F}}$ is given, by the definition of the bifactor approximation, the following holds:

$$\sum_{i\in S'} f'_i + \sum_{j\in\mathcal{C}} \min_{i\in S'} d(i,j) \le \lambda_f \cdot \sum_{i\in S^*} f'_i + \lambda_c \cdot \sum_{j\in\mathcal{C}} \min_{i\in S^*} d(i,j).$$

Thus, the original cost correspond to S', $\mathrm{cost}(S')$ is:

$$\mathrm{cost}(S') = \sum_{i\in S'} f_i + \sum_{j\in\mathcal{C}} \min_{i\in S'} d(i,j) = \sum_{i\in S'\setminus S} f_i + \sum_{i\in S} f_i + \sum_{j\in\mathcal{C}} \min_{i\in S'} d(i,j).$$

From the definition of $\{f'_i\}_{i\in\mathcal{F}}$, $\sum_{i\in S'\setminus S} f_i = \sum_{i\in S'} f'_i$. Thus,

$$\begin{aligned}
\sum_{i\in S'\setminus S} f_i + \sum_{i\in S} f_i + \sum_{j\in\mathcal{C}} \min_{i\in S'} d(i,j) &= \sum_{i\in S'} f'_i + \sum_{i\in S} f_i + \sum_{j\in\mathcal{C}} \min_{i\in S'} d(i,j) \\
&\le \lambda_f \cdot \sum_{i\in S^*} f'_i + \lambda_c \cdot \sum_{j\in\mathcal{C}} \min_{i\in S^*} d(i,j) + \sum_{i\in S} f_i \\
&= \lambda_f \cdot \sum_{i\in S^*\setminus S} f_i + \lambda_c \cdot (\mathrm{OPT} - \sum_{i\in S^*} f_i) + \sum_{i\in S} f_i \\
&\le \lambda_f \cdot \eta^- + \lambda_c \cdot \mathrm{OPT} - \lambda_c \cdot \eta^- + \eta^+ \\
&\le \left(\lambda_c + \frac{\eta^+ + (\lambda_f - \lambda_c)\eta^-}{\mathrm{OPT}}\right) \mathrm{OPT}.
\end{aligned}$$

□

H Proof of Theorem 4

Theorem 4 For the hard instances constructed from n-vertex graphs above, for every $\zeta \in (0,1)$ and every $\beta \in (0, \beta_0)$, no algorithm can output a $(1+\beta)$-approximate UFL solution in time $2^{O(n^{1-\zeta})}$ while making q_1 M1 label queries and q_2 M2 pairwise queries satisfying

$$q_1 \log k + q_2 = O(k^{1-\zeta}),$$

where k is the number of facilities in the fixed optimum. In particular, an M1-only algorithm requires $q_1 = \Omega(k^{1-\zeta}/\log k)$ label queries, and an M2-only algorithm requires $q_2 = \Omega(k^{1-\zeta})$ pairwise queries.

Proof. Assume that such an algorithm A exists. On a YES instance, fix a minimum vertex cover T^* and the corresponding optimal UFL solution S^*. Since G is 4-regular and $|E| = 2n$, every vertex cover has size at least $n/2$, while trivially $|T^*| \le n$; hence $k = |S^*| = |T^*| = \Theta(n)$. The optimum induces a partition of the edge-clients into k clusters by assigning each edge-client to an open endpoint facility, with deterministic tie-breaking. This defines exact answers to both the M1 label oracle and the M2 pairwise oracle.

If the number of queries is too small, the entire set of possible answers can be enumerated. We remove the oracles by transcript enumeration: a label query has k possible answers, while a pairwise query has two possible answers. Thus the number of adaptive joint transcripts is at most

$$k^{q_1} 2^{q_2} = 2^{q_1 \log k + q_2} = 2^{O(k^{1-\zeta})} = 2^{O(n^{1-\zeta})}.$$

An oracle-free algorithm can enumerate all such transcripts, simulate A on each transcript, compute the true UFL cost of each output solution, and keep the cheapest one.

If G is a YES instance, then the transcript consistent with the exact oracles is among the enumerated transcripts. Along that transcript, A returns a solution of cost at most

$$(1+\beta)\,\mathrm{OPT}(I_G) \le (1+\beta)B < (1+\beta_0)B$$

with the success probability guaranteed by A. If G is a NO instance, then every UFL solution has cost at least $(1+\beta_0)B$, regardless of the transcript. Therefore the enumeration algorithm distinguishes the two UFL gap cases in time $2^{O(n^{1-\zeta})}$, contradicting Lemma 7. Hence $q_1 \log k + q_2 = \Omega(k^{1-\zeta})$. □

I Proof of Theorem 5

Theorem 5 For every $\zeta \in (0,1)$ and every $\rho < 1+\beta_0$, under ETH, no algorithm running in time $2^{O(n^{1-\zeta})}$ on the hard instances constructed from n-vertex graphs can provide the following guarantee for every metric UFL instance I and every predicted facility set $S \subseteq \mathcal{F}$:

$$\mathrm{cost}(\mathcal{A}(I,S)) \le \rho\ \mathrm{OPT}(I) + r_-(\eta^-)$$

with $r_-(x) = o(x)$. Likewise, no such algorithm can provide the guarantee

$$\mathrm{cost}(\mathcal{A}(I,S)) \le \rho\ \mathrm{OPT}(I) + r_+(\eta^+)$$

with $r_+(x) = o(x)$.

Proof. We use the same hard family. For the false-negative lower bound, run the assumed algorithm with the prediction $S = \emptyset$. Note that this is a vacuous prediction: it carries no information about the optimum and requires no oracle access. If S^* is an optimal solution, then all optimal facilities are missing from the prediction, so

$$\eta^- = \sum_{i \in S^*} f_i = \lambda |S^*| = \lambda \tau(G).$$

On YES instances, $\tau(G) = \Theta(n)$ and $\mathrm{OPT}(I_G) \le B = \Theta(n)$, so $\eta^- = \Theta(n)$ and $r_-(\eta^-) = o(n)$. The assumed guarantee would then give

$$\mathrm{cost}(\mathcal{A}(I_G, \emptyset)) \le \rho B + o(n).$$

Since $\rho < 1+\beta_0$, the gap $(1+\beta_0)B - \rho B$ is $\Theta(n)$, and therefore $\rho B + o(n) < (1+\beta_0)B$ for all sufficiently large n. In NO instances every solution has cost at least $(1+\beta_0)B$. Thus checking whether the output cost is below $(1+\beta_0)B$ distinguishes the hard gap, a contradiction.

For the false-positive lower bound, run the algorithm with the prediction $S = \mathcal{F}$, the set of all facilities. Then $\eta^- = 0$, and

$$\eta^+ = \sum_{i \in \mathcal{F} \setminus S^*} f_i = \lambda(n - \tau(G)).$$

On YES instances, $\tau(G) \le an$, so $\eta^+ = \Theta(n)$ and $r_+(\eta^+) = o(n)$. The same gap calculation gives

$$\mathrm{cost}(\mathcal{A}(I_G, \mathcal{F})) \le \rho B + o(n) < (1+\beta_0)B$$

for all sufficiently large n, while on every NO instance all solutions have cost at least $(1+\beta_0)B$. This again distinguishes the ETH-hard UFL gap. Hence neither error term can be replaced by a sublinear dependence. □