LEARNING-AUGMENTED FACILITY LOCATION WITH SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Facility location is a fundamental problem in clustering and unsupervised learning. We consider the stochastic setting in which the clients arrive online from a sequence of unknown distributions. It is known (Gupta, Kehne, and Levin, SODA '24) that learning augmentation in the form of just a *single offline sample* from each distribution can significantly improve performance guarantees over the worst-case. However, this algorithm is non-robust: noisy samples invalidate the algorithmic guarantees. In this paper, we address this gap by designing the first *robust* learning-augmented algorithm for online facility location with samples. Our algorithm is robust to two natural forms of noise: adversarial corruption, in which individual samples may be manipulated, and distributional noise, in which samples are drawn from distributions that do not exactly match the corresponding input distributions. We complement our theoretical results with numerical simulations that empirically evaluate the performance and robustness of our algorithm.

1 INTRODUCTION

Clustering is one of the foundational pillars of unsupervised learning. In recent years, there has been increased focus on clustering for dynamically evolving datasets, leading to clustering problems in the online setting, where the set of points to be clustered arrives over time and clustering decisions have to be taken without knowing the future input (Meyerson, 2001; Alon et al., 2006; Fotakis, 2007; 2008; Liberty et al., 2016; Lattanzi & Vassilvitskii, 2017; Cohen-Addad et al., 2019; 2021; Almanza et al., 2021; Lattanzi et al., 2021; 2026). This reflects the growing interest in optimizing learning objectives in online and uncertain environments. A popular paradigm that seeks to combat this uncertainty is that of *learning-augmented algorithms*, where machine-learned predictions are used to augment online decision-making (Purohit et al., 2018; Wei & Zhang, 2020; Mitzenmacher & Vassilvitskii, 2020; Lykouris & Vassilvitskii, 2021; Antoniadis et al., 2020; Bamas et al., 2020; Anand et al., 2022). The basic desiderata of this framework are to obtain better than worst-case guarantees when the predictions are accurate, called the algorithm's *consistency*, while simultaneously ensuring worst-case guarantees when the predictions are arbitrary, called the algorithm's *robustness*, and to interpolate smoothly between these two extremes. In this paper, we present a new algorithm for the classical facility location problem when provided samples from the input distribution that achieves optimal consistency and robustness bounds, and degrades gracefully and optimally between the two.

The goal in clustering is to partition data points (called clients) into groups (or clusters) while optimizing a desired objective. A popular objective is the sum of distances of the data points from their cluster centers, known as the k-median problem. The fact that the number of clusters k is fixed rules out extending this problem to the online setting directly. Instead, the clustering problem of choice in the online setting has been the (Lagrangian) relaxation of k-median called the facility location problem. Here, clients have to be assigned to open facilities; while there is no limit on the number of facilities, every open facility adds an opening cost to the clustering objective. In online facility location, introduced by Meyerson (2001) and studied extensively since then (Fotakis, 2007; 2008; Anagnostopoulos et al., 2004; Lattanzi et al., 2026), clients arrive over time and have to be connected to an open facility, possibly after opening a new facility. The goal is to minimize the sum of all connection costs, measured as the distance between each client and the open facility they are connected to, and the opening costs of the open facilities. By now, it is known that the optimal competitive ratio for a sequence of n clients is $\Theta(\log n / \log \log n)$ (Meyerson, 2001; Fotakis, 2008).

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.

Recent work on learning-augmented facility location has considered several forms of side information to improve on the worst-case competitive ratio. These include predictions of the optimal facility serving each client (Jiang et al., 2022; Fotakis et al., 2025), predictions of the request set (Azar et al., 2022; Ameli et al., 2026), and multiple candidate solutions (Almanza et al., 2021; Anand et al., 2022; Lattanzi et al., 2026). Samples from the distributions generating the input are another natural and well-studied form of prediction. In this direction, Gupta et al. (2024) showed that for the case of uniform facility opening costs, an independent sample from each client distribution can be leveraged to obtain a competitive ratio of 6, thereby significantly improving over the $\Theta(\log n/\log\log n)$ worst-case bound.

While this result satisfies consistency, we show in Section C of the appendix that their algorithm is *not robust*: when samples can be noisy due to distributional shifts or adversarial corruptions, the algorithm fails to provide performance guarantees. In this paper, we address this gap by designing a *robust* algorithm for online facility location with samples. Our goal is to retain the consistency bounds for accurate samples while ensuring performance degrades gracefully with sample noise.

We consider robustness against two natural forms of sample noise, formalized in Section 2. In the *distributional noise* model, the distributions generating the offline samples may deviate from the distributions producing the online input. We quantify the sample error as the sum of total variation (i.e., ℓ_1) distance between the two sets of distributions. Next, we consider the *adversarial noise* model, where an adversary observes the realized offline samples and may corrupt them before they are presented to the algorithm. In this case, we define sample error as the number of samples that have been corrupted.

Our Contributions. Our main contributions are the following.

- **Robust Algorithm.** We give a single algorithm, ROBUST-ONLINE, that achieves a competitive ratio of $O(\log \ell/\log\log \ell)$ in both noise models (Theorem 3.1), where ℓ is the amount of noise in the samples. Thus, the algorithm's performance smoothly interpolates between $O(1)$-competitiveness with accurate samples and $O(\log n/\log\log n)$-competitiveness with arbitrary sample noise.
- **Matching Lower Bound.** We prove a matching $\Omega(\log \ell/\log\log \ell)$ lower bound on the competitive ratio in both noise models, showing that the dependence on the sample error is optimal (Theorem B.11 in Section B of the appendix). The lower bound holds even with uniform facility opening costs.
- **Experimental Evaluation.** We complement the theoretical analysis by evaluating our algorithm empirically on a variety of synthetically generated instances (Section 4). These numerical simulations demonstrate that (a) offline samples can improve performance over algorithms that do not use a sample, and (b) the performance degrades gracefully as the amount of noise in the samples increases.

1.1 RELATED WORK

Online facility location. Online clustering has been studied in various models, including irrevocable cluster assignments, regret against the best clustering in hindsight, and bounded-recourse or dynamic formulations (Liberty et al., 2016; Lattanzi & Vassilvitskii, 2017; Cohen-Addad et al., 2019; 2021; Lattanzi et al., 2021). We focus on online facility location, introduced by Meyerson (2001).We focus on online facility location, introduced by Meyerson (2001). The randomized algorithm of Meyerson (2001) is $O(\log n/\log\log n)$-competitive in the worst case (Fotakis, 2008) and $O(1)$-competitive under random arrival order. Fotakis (2008) also gives a deterministic algorithm with the same worst-case guarantee. For uniform facility costs, Kaplan et al. (2023) give a 3-competitive random-order algorithm. More broadly, there is also prior work on online facility location when the distributions are known and identical (Garg et al., 2008).

Learning-augmented algorithms. Our work fits in the learning-augmented framework, which seeks to improve worst-case online guarantees using machine-learned predictions when accurate, while retaining robustness if not (Purohit et al., 2018; Lykouris & Vassilvitskii, 2021; Mitzenmacher & Vassilvitskii, 2020). Of particular relevance to our work is the use of samples as predictions. The single-sample framework of Gupta et al. (2024) gives guarantees under exact sampling, while we study robustness to distributional and adversarial noise in the sample. A related model is that of

online algorithms with a sample, where the online input is adversarial but a sample of the input is provided offline to the algorithm (Kaplan et al., 2022; Argue et al., 2022; Kaplan et al., 2025).

The online facility location problem that we study in this paper has been previously studied in a variety of learning-augmented models. These works primarily differ in the nature of the predictions: the facility serving each arriving client in an optimal solution, the set of clients, several candidate solutions, and others (Azar et al., 2022; Gupta et al., 2022; Ameli et al., 2026; Almanza et al., 2021; Anand et al., 2022; Lattanzi et al., 2026; Jiang et al., 2022; Fotakis et al., 2025). These models differ from ours in that the advice predicts properties of the realized instance or of the optimal solution. In contrast, our side information is a random sample generated from distributions producing the online instance. Our work is also related to prior work on distributional predictions, where the algorithm receives information about the distribution generating its future input (Diakonikolas et al., 2021; Besbes et al., 2022; Dinitz et al., 2024; Angelopoulos et al., 2024; Cui & Dinitz, 2026). Our results provide consistency, robustness, and smoothness guarantees for sample-based advice.

Robust statistics. Our two noise models, distributional and adversarial noise, are closely related to classical noise models in robust statistics. Huber (1964) introduced a contamination model in which observations are drawn from a target distribution contaminated by arbitrary noise; see also Huber & Ronchetti (2009). Modern algorithmic robust statistics also studies strong contamination, in which a clean sample is drawn first and an adversary may inspect it before replacing some of its components (Diakonikolas & Kane, 2023). These models have motivated extensive work on robust estimation of statistical parameters under contamination and arbitrary outliers (Diakonikolas et al., 2016; Lai et al., 2016; Charikar et al., 2017; Steinhardt et al., 2018; Cheng et al., 2019; Diakonikolas et al., 2020; Dalalyan & Minasyan, 2022; Diakonikolas et al., 2023; Kane et al., 2024).

Roadmap. The rest of the paper is organized as follows. In Section 2, we formalize the facility location problem that we study in this paper. In Section 3, we present and analyze our robust algorithm for both the distributional and adversarial noise models. In Section 4, we evaluate the algorithm empirically via numerical simulations. Finally, we conclude with some closing remarks in Section 5. In the appendix, Section A contains additional experimental details, Section B contains proofs omitted from Section 3, and Section C shows that the algorithm of Gupta et al. (2024) is not robust in either noise model.

2 PRELIMINARIES

A facility location instance consists of a metric space (V, d) containing a set of facilities $\mathcal{F} \subseteq V$ and a set of n clients $X \subseteq V$. Each facility $f \in \mathcal{F}$ has an opening cost $c_f \geq 1$ (by scaling). The goal is to assign each client to an open facility such that the sum of distances between the clients and their assigned facilities plus the opening cost of the open facilities is minimized.

In the online problem, the clients are not known upfront to the algorithm and arrive online, and must be assigned to an open facility on arrival, possibly after opening a new facility. We consider an input sequence $X = (v^1, \ldots, v^n)$ where v^i is drawn independently from some distribution D_i. Note that $D_1, \ldots, D_n$ need not be identical, and these distributions are not known to the algorithm. Further, since the distributions need not be identical, this model can generate any input sequence by choosing appropriate point distributions.

For a facility set $F \subseteq \mathcal{F}$ and a client set $Y \subseteq V$, write

$$\mathrm{Obj}(F, Y) := \sum_{f \in F} c_f + \sum_{v \in Y} d(v, F).$$

We denote $\mathrm{OPT}(Y) := \min_F \mathrm{Obj}(F, Y)$. We treat client collections as multisets when different indices realize the same client location in V, but use set notation for simplicity.

Distributional Noise. The offline sample $\hat{X} = (\hat{v}^1, \ldots, \hat{v}^n)$ is generated independently from predicted distributions $\hat{D}_1, \ldots, \hat{D}_n$. The prediction error ℓ is the sum of the ℓ_1 distances between the predicted and actual distributions, i.e.,

$$\ell := \sum_{i=1}^{n} \|\hat{D}_i - D_i\|_1 = \sum_{i=1}^{n} \sum_{u \in V} |\hat{D}_i(u) - D_i(u)|.$$

The algorithm is given only the sample $\hat{X}$; it knows neither the distributions nor ℓ. It will be convenient to decompose each pair of distributions into common and distinct parts:

$$D_i = P_i + Q_i, \qquad \hat{D}_i = P_i + \hat{Q}_i, \qquad P_i(u) = \min\{D_i(u), \hat{D}_i(u)\}. \tag{1}$$

Then, define

$$q_i := \|Q_i\|_1 = \|\hat{Q}_i\|_1 = \frac{1}{2}\|D_i - \hat{D}_i\|_1.$$

We view $\hat{v}^i$ as being drawn from (normalized) P_i with probability $1 - q_i$, and from (normalized) $\hat{Q}_i$ with probability q_i. Let G_i indicate that $\hat{v}^i$ is drawn from $\hat{Q}_i$, and define

$$\hat{Y} := \{\hat{v}^i : G_i = 0\}, \qquad \hat{Z} := \{\hat{v}^i : G_i = 1\}.$$

We refer to $\hat{Y}$ and $\hat{Z}$ as the uncorrupted and corrupted parts of the sample, respectively.

Adversarial Noise. An independent clean sample $\tilde{X} = (\tilde{v}^1, \ldots, \tilde{v}^n)$ is first generated with $\tilde{v}^i \sim D_i$. *After observing* $\tilde{X}$, an adversary replaces exactly ℓ entries and gives the resulting sample $\hat{X}$ to the algorithm. We again write

$$\hat{X} = \hat{Y} \cup \hat{Z},$$

where $\hat{Y}$ comprises the unmodified entries and $\hat{Z}$ comprises the adversarial replacements. Note that $|\hat{Z}| = \ell$. As in the distributional noise model, the sample error ℓ is not known to the algorithm.

3 Robust Algorithm for Facility Location

We give a single algorithm that we call Robust-Online for both noise models.

Theorem 3.1 (Sample robustness)**.** Robust-Online *has expected cost at most* $O\left(\frac{\log \ell}{\log\log \ell}\right)$ *times* $\mathbb{E}_X[\mathrm{OPT}(X)]$*, where* ℓ *is the appropriate notion of sample error in each noise model.*

The dependence on ℓ is optimal up to constant factors; a reduction from the classical $\Omega(\log n/\log\log n)$ online lower bound (Fotakis, 2008) appears in Section B of the appendix.

We first describe the algorithm, then prove a refined guarantee for Meyerson's algorithm (Meyerson, 2001), and use it to analyze the robustness of our algorithm and establish Theorem 3.1.

3.1 The Robust-Online Algorithm

The Robust-Online algorithm uses two runs of the algorithm from Meyerson (2001), which we will describe shortly. The first run, denoted RO, is simulated offline on a uniformly random ordering $\hat{X}_1, \ldots, \hat{X}_n$ of the sample. The second run, denoted ON, serves the actual online clients.

Meyerson's algorithm (Meyerson, 2001). A facility f has type j if $2^j \le c_f < 2^{j+1}$. Let s be the largest facility type and let $\mathcal{F}_{\le j}$ denote facilities of type $\le j$, with $\mathcal{F}_{\le -1} = \varnothing$.

When a client v arrives, let $F_{<v}$ be the set of facilities currently open. Define

$$\delta_j(v) := d(\mathcal{F}_{\le j} \cup F_{<v}, v) \qquad \text{for } -1 \le j \le s.$$

Independently for every $0 \le j \le s$, open the closest type-j facility to v with probability

$$\min\left\{\frac{\delta_{j-1}(v) - \delta_j(v)}{2^j}, 1\right\}.$$

The client is then connected to the nearest open facility. Note that if a type-j facility f_j is opened when v arrives, then $d(v, f_j) = \delta_j(v)$.

Offline simulation. We first generate a uniform random ordering $\hat{X}_1, \ldots, \hat{X}_n$ of the offline sample $\hat{X}$ and run Meyerson's algorithm on this sequence (RO). Let F_i^{RO} denote the facilities opened after processing the first i samples, and let C_i^{RO} denote the cumulative opening and connection cost incurred on this prefix, with $F_0^{\mathrm{RO}} = \varnothing$ and $C_0^{\mathrm{RO}} = 0$. This run is only a simulation; its facilities will be opened when processing the real clients only if budget permits.

Online phase (ON)**.** Let B_t denote the cost incurred by the online run of Meyerson's algorithm (ON) through time t, excluding purchases from the offline simulation. Start with no open facilities, $B_0 = 0$, and $\eta_0 = 0$. When client v^t arrives:

1. Serve v^t using the online run of Meyerson's algorithm (ON). Add this step's opening and connection cost to B_{t-1} to obtain B_t.
2. Set $\eta_t = \max\{i \in \{0, \dots, n\} : C_i^{\mathrm{RO}} \le B_t\}$, and open all unopened facilities in $F_{\eta_t}^{\mathrm{RO}}$.

3.2 Refined Meyerson Guarantee

We will need to analyze the performance of Meyerson's algorithm after some facilities from the offline sample have already been opened. The standard worst-case guarantee does not capture the benefit of these initial facilities, and depends only on the number of remaining clients. We instead bound the expected cost in terms of how well the initial facility set serves the clients.

For any facility set $F \subseteq \mathcal{F}$ and client $v \in V$, define

$$\kappa_F(v) := \min_{f \in \mathcal{F}} \{\mathbf{1}\{f \notin F\} \cdot c_f + d(v, f)\}.$$

This is the minimum additional cost needed to serve v if all facilities in F are already open. We also have the monotonicity property that $\kappa_F(v) \ge \kappa_{F'}(v)$ if $F \subseteq F'$.

We will need two local cost bounds. First, the expected cost incurred by ON on a client v, conditioned on the facilities $F_{<v}$ open just before v arrives, is at most $O(\kappa_{F_{<v}}(v))$ (Lemma B.1). Second, the bound given in Lemma 3.2 applies to any subset of clients contained in a ball. Both proofs appear in Section B of the appendix.

Lemma 3.2 (Ball cost)**.** *Let F_0 be the initial open set of facilities. Fix a facility $g \in \mathcal{F}$ and a subsequence $S \subseteq X$ of clients satisfying $d(v, g) \le R$ for every $v \in S$. Then*

$$\mathbb{E}[\mathsf{cost}_{\mathrm{ON}}(S)] \le O\left(\mathbf{1}\{g \notin F_0\} \cdot c_g + \sum_{v \in S} d(v, g) + |S|R\right).$$

The clients in S need not arrive consecutively. Here $\mathsf{cost}_{\mathrm{ON}}(S)$ is the cost of facility openings and connections made by ON *in response to clients in S.*

Now, for an initial facility set F_0, let

$$\Gamma := \sum_{v \in X} \kappa_{F_0}(v).$$

Thus, Γ is the total cost of serving each client greedily given F_0. We now show a refined guarantee for Meyerson's algorithm using Lemma 3.2 at geometrically increasing radii determined by F_0.

Lemma 3.3 (Refined Meyerson bound)**.** *For any client sequence X and initial facility set F_0 and for every $a \ge 2$, the expected cost of* ON *excluding the cost of F_0 is at most*

$$O\left(a + \log_a\left(2 + \frac{\Gamma}{\mathrm{OPT}(X)}\right)\right) \cdot \mathrm{OPT}(X).$$

The bound also holds when additional facilities are made available between arrivals. Since $\Gamma \le n \cdot \mathrm{OPT}(X)$, choosing $a = \Theta(\log n / \log\log n)$ and $F_0 = \varnothing$ recovers the usual worst-case guarantee.

Proof. Fix an optimal facility location solution for X. For a facility f^* and its assigned clients C^*, write

$$d_v^* = d(v, f^*), \qquad B^* = c_{f^*} + \sum_{v \in C^*} d_v^*, \qquad \Gamma^* = \sum_{v \in C^*} \kappa_{F_0}(v).$$

Set $\lambda = \kappa_{F_0}(f^*)$. The triangle inequality gives

$$\lambda - d_v^* \le \kappa_{F_0}(v) \le \lambda + d_v^*. \tag{2}$$

Far clients. If $d_v^* \ge \lambda/2$, then (2) gives $\kappa_{F_{<v}}(v) \le \kappa_{F_0}(v) \le 3d_v^*$. Using Lemma B.1, the expected cost incurred by ON for such clients is at most $O(B^*)$.

Near clients. Let $C^*_{\text{in}} = \{v \in C^* : d^*_v < \lambda/2\}$ and $m = |C^*_{\text{in}}|$. Every near client has $\kappa_{F_0}(v) > \lambda/2$, and hence

$$m\lambda/2 \le \Gamma^*. \tag{3}$$

Partition these clients into disjoint sets as follows:

$$S_0 = \{v \in C^*_{\text{in}} : d^*_v \le B^*/m\},$$
$$S_j = \{v \in C^*_{\text{in}} : a^{j-1}B^*/m < d^*_v \le a^j B^*/m\} \quad (j \ge 1).$$

Let b be the index of the last nonempty annulus, with $b = 0$ if all near clients lie in S_0. If $b \ge 1$, its inner radius is less than $\lambda/2$, so (3) gives

$$a^{b-1} < \frac{m\lambda}{2B^*} \le \frac{\Gamma^*}{B^*} \implies b \le 1 + \log_a(2 + \Gamma^*/B^*).$$

Lemma 3.2 applied to S_0 with center f^* and radius B^*/m shows that the expected cost of ON on S_0 is at most $O(B^*)$ as $|S_0| \le m$. We then apply Lemma 3.2 on S_j for $1 \le j \le b$ with center f^* and radius $a^j B^*/m$. This radius is less than ad^*_v for every $v \in S_j$, yielding

$$\mathbb{E}[\mathsf{cost}_{\text{ON}}(S_j)] \le O(c_{f^*} + a \sum_{v \in S_j} d^*_v).$$

Thus, summing the expected cost of ON over the near and far clients gives

$$\mathbb{E}[\mathsf{cost}_{\text{ON}}(C^*)] \le O\left(a + \log_a(2 + \Gamma^*/B^*)\right) B^*. \tag{4}$$

Summing over clusters. We have $\sum_{C^*} B^* = \text{OPT}(X)$ and $\sum_{C^*} \Gamma^* = \Gamma$. Since $\log_a(2 + x)$ is concave, weighted Jensen's inequality gives

$$\sum_{C^*} B^* \log_a(2 + \Gamma^*/B^*) \le \text{OPT}(X) \cdot \log_a\left(2 + \frac{\Gamma}{\text{OPT}(X)}\right).$$

Thus, summing (4) across the optimal clusters proves the desired bound. □

3.3 Robustness Analysis

Recall from Section 2 that in both noise models, we write $\hat{X} = \hat{Y} \cup \hat{Z}$, where $\hat{Y}$ and $\hat{Z}$ denote the uncorrupted and corrupted parts of the offline sample. Let ζ be the number of clients before the first element of $\hat{Z}$ in the random order on $\hat{X}$ used by RO. (Let $\zeta = n$ if $\hat{Z} = \varnothing$.) Thus, $\hat{X}_{\le\zeta} := (\hat{X}_1, \ldots, \hat{X}_\zeta)$ is an uncorrupted prefix of the random-order Meyerson run (RO) on $\hat{X}$, with cumulative opening and connection cost C^{RO}_ζ.

For each $r \in \{0, \ldots, n\}$, let

$$\Gamma_r := \sum_{i=1}^{n} \kappa_{F^{\text{RO}}_r}(v^i)$$

where F^{RO}_r is the set of facilities opened by RO after the first r clients in the random order on $\hat{X}$ have been processed. Our analysis shows that the uncorrupted prefix has small expected cost and that the facilities it opens make Γ_ζ small in expectation.

Lemma 3.4 (Uncorrupted-prefix bounds)**.** *In either noise model, we have*

$$\mathbb{E}[C^{\text{RO}}_\zeta] \le O(\mathbb{E}_X[\text{OPT}(X)]), \quad \mathbb{E}[\Gamma_\zeta] \le O(\ell)\,\mathbb{E}_X[\text{OPT}(X)]. \tag{5}$$

Proof. **Cost of the uncorrupted prefix.** In both models, we have $\mathbb{E}[\text{OPT}(\hat{Y})] \le \mathbb{E}_X[\text{OPT}(X)]$. Indeed, for adversarial noise, $\hat{Y}$ is a subset of the original uncorrupted sample, which has the same distribution as X. For distributional noise, retain the sample values drawn from P_i and replace each draw from $\hat{Q}_i$ by an independent draw from Q_i. The resulting sample has the same distribution as X and contains $\hat{Y}$, so monotonicity of the optimum gives the inequality.

Now, condition on $\hat{Y}$. Its relative order in the sample permutation is uniform, and $\hat{X}_{\le\zeta}$ is a prefix of $\hat{Y}$ in that order. The random-order guarantee (Meyerson, 2001) therefore gives

$$\mathbb{E}[C^{\text{RO}}_\zeta] \le O(\mathbb{E}[\text{OPT}(\hat{Y})]) \le O(\mathbb{E}_X[\text{OPT}(X)]), \tag{6}$$

which proves the first bound in Lemma 3.4. The proof of the second inequality is more involved; we give a sketch below and defer the full proof to Section B of the appendix.

Distributional residual cost. Recall that each online client $v^i \sim D_i = P_i + Q_i$ is drawn either from the common part P_i or from the distinct part Q_i. We handle these two contributions separately. Clients drawn from the common parts can be compared with the uncorrupted offline samples and charged to the cost incurred by RO on the uncorrupted prefix. The first corrupted sample typically appears after about n/ℓ positions in the random order, leading to an $O(\ell)$ loss in this charging argument. Clients drawn from the distinct parts have total probability mass $\sum_i q_i = \ell/2$, and their contribution is bounded using Lemma B.3. Combining these two arguments with (6) gives $\mathbb{E}[\Gamma_\zeta] \le O(\ell)\,\mathbb{E}_X[\mathrm{OPT}(X)]$.

Adversarial residual cost. Here the adversary observes the offline sample before deciding which entries to corrupt. Thus the uncorrupted samples are no longer independent: conditioning on a random-order prefix containing no corrupted entries can itself bias the distribution of that prefix, making the analysis more involved than in the distributional case.

First, we label each offline sample by its index $i \in [n]$. Fix an uncorrupted labeled prefix h of the offline sample $\hat{X}$ of length t. In the absence of corruption, a (labeled) value (i, x) would appear next with probability $D_i(x)/(n-t)$ in the random order. We call (i, x) *bad* if, conditioned on the observed uncorrupted prefix, its probability of appearing next is less than half this benchmark:

$$\Pr[\hat{X}_{t+1} = (i,x) \mid \hat{X}_{\le t} = h, \zeta > t] < \frac{D_i(x)}{2(n-t)}.$$

Our stochastic-distortion lemma (Lemma B.8) shows that the expected total probability mass of such badly distorted values is only $O(\ell)$. The remaining values occur with probability comparable to their uncorrupted benchmark, so their service cost can be charged to the cost of RO on the uncorrupted prefix. Since this prefix typically has length about n/ℓ, this charging incurs an $O(\ell)$ multiplicative loss. The distorted mass is handled using Lemma B.3, giving $\mathbb{E}[\Gamma_\zeta] \le O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)]$. □

We are now ready to prove the performance guarantee of ROBUST-ONLINE in both noise models.

Proof of Theorem 3.1. Note that our budgeting rule ensures that the cost of the facilities opened from the simulated RO run on $\hat{X}$ is no more than the cost incurred by ON. Let t^* be the first time that $\mathsf{cost}(\mathrm{ON}_{\le t}) = B_t \ge C_\zeta^{\mathrm{RO}}$, if such a time exists. The idea is that after t^*, all the facilities in F_ζ^{RO} are open and available to our algorithm.

Before t^*. Condition on X, the offline sample, and the simulated RO run. The cost strictly before t^* is less than C_ζ^{RO}. Combining this with the one-step Meyerson bound (Lemma B.1) and Lemma 3.4,

$$\mathbb{E}[B_{t^*}] \le \mathbb{E}[C_\zeta^{\mathrm{RO}}] + \mathbb{E}_X[\mathrm{OPT}(X)] \le O(1) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

If t^* does not exist, this already bounds the expected cost of ON.

After t^*. If t^* exists, then $\eta_{t^*} \ge \zeta$ and all facilities in F_ζ^{RO} are open at the end of time t^*. Condition on X, the offline sample, the simulated RO run, and the facilities open immediately after time t^*. Since these facilities include F_ζ^{RO}, the remaining clients have residual cost at most Γ_ζ, while their optimum is at most $\mathrm{OPT}(X)$. Hence, by Lemma 3.3, for every $a \ge 2$ the expected cost incurred by ON for the subsequent clients is at most

$$O\left(a \cdot \mathrm{OPT}(X) + \mathrm{OPT}(X) \cdot \log_a\left(2 + \frac{\Gamma_\zeta}{\mathrm{OPT}(X)}\right)\right).$$

By Lemma 3.4 we have $\mathbb{E}[\Gamma_\zeta] \le O(\ell)\,\mathbb{E}_X[\mathrm{OPT}(X)]$. The function $(x, y) \mapsto x \log_a(2 + y/x)$ is jointly concave for $x > 0$ and $y \ge 0$. Therefore, multivariate Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\left[\mathrm{OPT}(X) \cdot \log_a\left(2 + \frac{\Gamma_\zeta}{\mathrm{OPT}(X)}\right)\right] &\le \mathbb{E}_X[\mathrm{OPT}(X)] \cdot \log_a\left(2 + \frac{\mathbb{E}[\Gamma_\zeta]}{\mathbb{E}_X[\mathrm{OPT}(X)]}\right) \\ &\le O(\log_a(2+\ell)) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].\end{aligned}$$

Thus, we can combine the expected cost incurred by ON before and after t^* to conclude

$$\mathbb{E}[\mathsf{cost}(\mathrm{ON})] = \mathbb{E}[B_n] \le O\left(a + \log_a(2+\ell)\right) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

Choosing $a = \Theta(\log \ell / \log\log \ell)$ proves the theorem. □

4 EXPERIMENTAL EVALUATION

Due to space constraints, we give some experimental results in this section and defer remaining details to Section A of the appendix.

We compare our algorithm ROBUST-ONLINE with the following alternative online algorithms:

- *Greedy* connects each incoming client in the cheapest way possible (which might involve opening a new facility),
- *Fotakis* is the deterministic $O\left(\frac{\log n}{\log \log n}\right)$-competitive algorithm of Fotakis (2008),
- *Fotakis PD* is the deterministic $O(\log n)$-competitive primal-dual algorithm of Fotakis (2007), and
- *Meyerson* is the randomized algorithm of Meyerson (2001).

We used Gurobi 13.0.3 and its mixed-integer linear programming solver to compute optimal facility location solutions. We also compare against a sample-based baseline for each online algorithm $\mathcal{A}$: *Gurobi* + $\mathcal{A}$ computes an exact solution on the offline sample, opens its facilities, and runs $\mathcal{A}$ online.

Instance families. We used four cluster-based instance families and one hierarchical family:

- **Equidistant Clusters** comprises well-separated clusters at pairwise equidistant centers,
- **Circular Clusters** comprises cluster centers around a circle,
- **Exponential Line** comprises cluster centers at exponentially increasing positions $1, 2, 4, \ldots$ on a line,
- **Paired Clusters** arranges pairs of closely spaced clusters (A_i, B_i) around a circle, and
- **Hierarchical Line** is based on the lower-bound tree construction of Fotakis (2008).

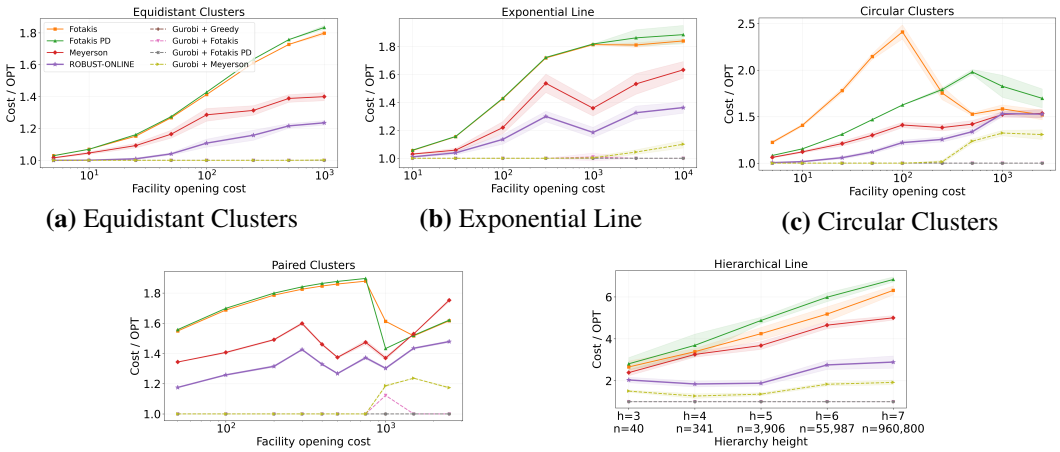
All experiments use uniform facility opening costs. For noisy samples, we use the stronger form of noise, adversarial corruption: we generate a clean sample from the same instance as the online input, then replace selected samples before presenting them to the algorithms.

Accurate samples. We first consider the five instance families above without sample corruption. In panels (a)–(d) of Figure 1, we vary the facility opening cost to test the tradeoff between opening a new facility and connecting to an existing one. For **Hierarchical Line** in panel (e), the lower-bound construction fixes the facility cost as a function of the hierarchy height h, so we instead vary h.

Across these experiments, ROBUST-ONLINE generally performs better than Meyerson, showing that an offline sample can substantially improve performance. The *Gurobi* + $\mathcal{A}$ baselines perform the best, often attaining or nearly attaining the optimum, since an accurate sample closely reflects the online client distribution. However, this comes at a higher computational cost. Since a ratio of 1 is optimal, improvements are more meaningful in terms of excess cost over OPT. For example, in **Equidistant Clusters** at facility cost 100, ROBUST-ONLINE improves the ratio from 1.29 to 1.11, cutting the excess cost by more than half. In panels (a)–(d), ratios generally increase with the facility opening cost, as unnecessary openings become more expensive. At sufficiently high costs, however, all algorithms open facilities more sparingly, so the approximation ratio can decrease again.

Corrupted samples. Next, we adversarially corrupt a varying fraction of the offline sample for the same instance families. Panels (a)–(d) of Figure 2 consider the cluster-based instances. In **Equidistant Cluster Corruption** and **Circular Cluster Corruption**, corruptions can cause the *Gurobi* + $\mathcal{A}$ baselines to incur substantial extra cost from unnecessary facility openings, especially in the former, while ROBUST-ONLINE is much less sensitive. In **Paired Cluster Corruption** and **Exponential Line Corruption**, ROBUST-ONLINE outperforms Meyerson throughout both corruption ranges and, for **Exponential Line**, often outperforms Gurobi-initialized baselines. Panels (e) and (f) consider hierarchical corruptions. In **Hierarchical Branch Corruption**, ROBUST-ONLINE substantially outperforms the online-only algorithms through most of the corruption range, deteriorating only once the sample contains little useful information. In **Hierarchical Component Corruption**, ROBUST-ONLINE and the Gurobi-initialized baselines approach the performance of online-only algorithms as corruption increases.

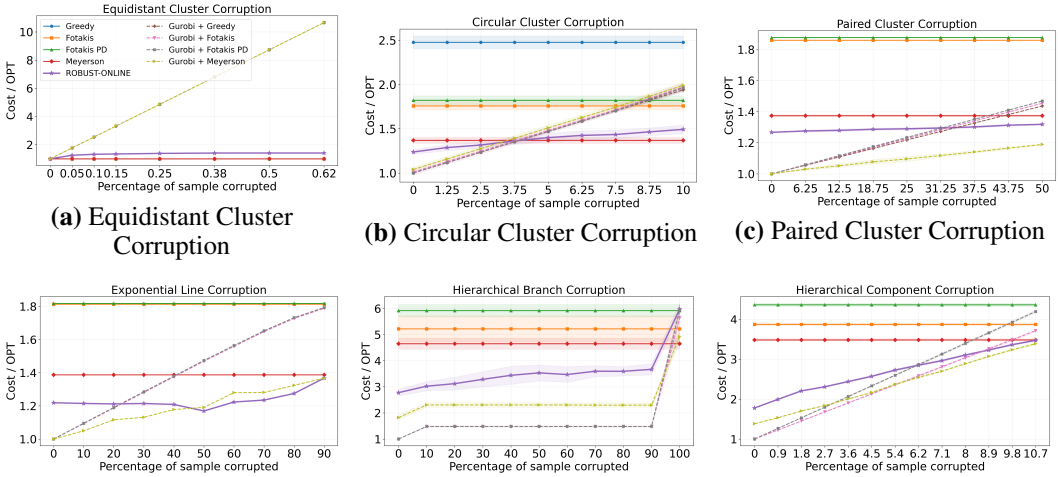
Our experiments suggest the following: when the sample is accurate, an exact sample is useful, but when the sample is noisy, committing fully to it can incur a large overhead. ROBUST-ONLINE in-



(a) Equidistant Clusters **(b)** Exponential Line **(c)** Circular Clusters

(d) Paired Clusters **(e)** Hierarchical Line

Figure 1: **Accurate Samples.** Average approximation ratio versus the facility opening cost in panels (a)–(d), and versus hierarchy height in panel (e). Shaded regions show one standard deviation across ten independently sampled instances. Greedy is omitted for readability; its maximum ratios in panels (a)–(e) are $\approx$ 19, 298, 29, 95, and 71650, respectively. The legend in (a) is shared across panels.



(a) Equidistant Cluster Corruption **(b)** Circular Cluster Corruption **(c)** Paired Cluster Corruption

(d) Exponential Line Corruption **(e)** Hierarchical Branch Corruption **(f)** Hierarchical Component Corruption

Figure 2: **Corruption experiments.** Average approximation ratio versus the percentage of the offline sample corrupted. Shaded regions show one standard deviation across ten independently sampled instances. Greedy is omitted from panels (c)–(f), where its maximum ratios are $\approx$ 15, 193, 4522, and 3165, respectively. Gurobi + Greedy is additionally omitted in panels (d)–(f), where its maximum ratios are $\approx$ 780, 9387, and 311, respectively. The legend in (a) is shared across panels.

stead incorporates sample information gradually, and thus maintains performance comparable to the worst-case online algorithms under corruption while still taking advantage of informative samples.

5 Closing Remarks

In this paper, we gave a robust algorithm for online facility location with samples. We considered two natural forms of sample error, distributional noise and adversarial noise, and gave the ROBUST-ONLINE algorithm whose competitive ratio smoothly and optimally degrades between perfect samples and arbitrarily noisy ones. Our experiments demonstrate the benefits of samples in overcoming worst-case behavior and establish the importance of robustness to noise.

AI USE STATEMENT

We used ChatGPT 6 Astra for minor copy editing and assistance with writing code for the experiments. All AI-assisted code was reviewed and independently checked by the authors, who take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

All algorithmic guarantees in this paper are supported by complete theoretical proofs. We give the full experimental setup and instance constructions needed to reproduce the empirical results. The code used for experiments has been uploaded as supplementary material.

REFERENCES

Matteo Almanza, Flavio Chierichetti, Silvio Lattanzi, Alessandro Panconesi, and Giuseppe Re. Online facility location with multiple advice. 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. 4661–4673, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/250473494b245120a7eaf8b2e6b1f17c-Abstract.html.

Noga Alon, Baruch Awerbuch, Yossi Azar, Niv Buchbinder, and Joseph Naor. A general approach to online network optimization problems. *ACM Transactions on Algorithms (TALG)*, 2(4):640–660, 2006.

Afrouz Jabal Ameli, Laura Sanità, and Moritz Venzin. Learning-augmented online covering problems. In *The 43rd International Conference on Machine Learning, ICML 2026, 6–11 July 2026, Seoul, South Korea*, 2026. doi: 10.48550/ARXIV.2507.06032. URL https://doi.org/10.48550/arXiv.2507.06032. To appear.

Aris Anagnostopoulos, Russell Bent, Eli Upfal, and Pascal Van Hentenryck. A simple and deterministic competitive algorithm for online facility location. *Inf. Comput.*, 194(2):175–202, 2004. doi: 10.1016/J.IC.2004.06.002. URL https://doi.org/10.1016/j.ic.2004.06.002.

Keerti Anand, Rong Ge, Amit Kumar, and Debmalya Panigrahi. Online algorithms with multiple predictions. 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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 582–598. PMLR, 2022. URL https://proceedings.mlr.press/v162/anand22a.html.

Spyros Angelopoulos, Marcin Bienkowski, Christoph Dürr, and Bertrand Simon. Contract scheduling with distributional and multiple advice. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 3652–3660. ijcai.org, 2024. URL https://www.ijcai.org/proceedings/2024/404.

Antonios Antoniadis, Themis Gouleakis, Pieter Kleer, and Pavel Kolev. Secretary and online matching problems with machine learned advice. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/5a378f8490c8d6af8647a753812f6e31-Abstract.html.

C. J. Argue, Alan M. Frieze, Anupam Gupta, and Christopher Seiler. Learning from a sample in online algorithms. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/5a093120ff4776b4f0dc452e3e3b6652-Abstract-Conference.html.

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. doi: 10.1137/1.9781611977073.3. URL https://doi.org/10.1137/1.9781611977073.3.

Étienne Bamas, Andreas Maggiori, and Ola Svensson. The primal-dual method for learning augmented algorithms. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e834cb114d33f729dbc9c7fb0c6bb607-Abstract.html.

Omar Besbes, Will Ma, and Omar Mouchtaki. Beyond IID: data-driven decision-making in heterogeneous environments. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/974ff7b5bf08dbf9400b5d599a39c77f-Abstract-Conference.html.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In Hamed Hatami, Pierre McKenzie, and Valerie King (eds.), *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pp. 47–60. ACM, 2017. doi: 10.1145/3055399.3055491. URL https://doi.org/10.1145/3055399.3055491.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In Timothy M. Chan (ed.), *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pp. 2755–2771. SIAM, 2019. doi: 10.1137/1.9781611975482.171. URL https://doi.org/10.1137/1.9781611975482.171.

Vincent Cohen-Addad, Niklas Hjuler, Nikos Parotsidis, David Saulpic, and Chris Schwiegelshohn. Fully dynamic consistent facility location. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3250–3260, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fface8385abbf94b4593a0ed53a0c70f-Abstract.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*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1126–1134. PMLR, 2021. URL http://proceedings.mlr.press/v130/cohen-addad21a.html.

Qiming Cui and Michael Dinitz. Ski rental with distributional predictions of unknown quality. In *The 43rd International Conference on Machine Learning, ICML 2026, 6–11 July 2026, Seoul, South Korea*, 2026. doi: 10.48550/ARXIV.2602.21104. URL https://doi.org/10.48550/arXiv.2602.21104. To appear.

Arnak S. Dalalyan and Arshak Minasyan. All-in-one robust estimator of the Gaussian mean. *The Annals of Statistics*, 50(2):1193 – 1219, 2022. doi: 10.1214/21-AOS2145. URL https://doi.org/10.1214/21-AOS2145.

Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023. doi: 10.1017/9781108943161. URL https://doi.org/10.1017/9781108943161.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In Irit Dinur (ed.),

IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, Hyatt Regency, New Brunswick, New Jersey, USA, October 9-11, 2016, pp. 655–664. IEEE Computer Society, 2016. doi: 10.1109/FOCS.2016.85. URL https://doi.org/10.1109/FOCS.2016.85.

Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/13ec9935e17e00bed6ec8f06230e33a9-Abstract.html.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, Ali Vakilian, and Nikos Zarifis. Learning online algorithms with distributional advice. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2687–2696. PMLR, 2021. URL http://proceedings.mlr.press/v139/diakonikolas21a.html.

Ilias Diakonikolas, Daniel Kane, Ankit Pensia, and Thanasis Pittas. Near-optimal algorithms for gaussians with huber contamination: Mean estimation and linear regression. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/86c283920335ed1fec3edee227e05fbf-Abstract-Conference.html.

Michael Dinitz, Sungjin Im, Thomas Lavastida, Benjamin Moseley, Aidin Niaparast, and Sergei Vassilvitskii. Binary search with distributional predictions. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a4b293979b8b521e9222d30c40246911-Abstract-Conference.html.

Dimitris Fotakis. A primal-dual algorithm for online non-uniform facility location. *J. Discrete Algorithms*, 5(1):141–148, 2007. doi: 10.1016/J.JDA.2006.03.001. URL https://doi.org/10.1016/j.jda.2006.03.001.

Dimitris Fotakis. On the competitive ratio for online facility location. *Algorithmica*, 50(1):1–57, 2008. doi: 10.1007/S00453-007-9049-Y. URL https://doi.org/10.1007/s00453-007-9049-y.

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.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 26973–26981. AAAI Press, 2025. doi: 10.1609/AAAI.V39I25.34903. URL https://doi.org/10.1609/aaai.v39i25.34903.

Naveen Garg, Anupam Gupta, Stefano Leonardi, and Piotr Sankowski. Stochastic analyses for online combinatorial optimization problems. In Shang-Hua Teng (ed.), *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, California, USA, January 20-22, 2008*, pp. 942–951. SIAM, 2008. URL http://dl.acm.org/citation.cfm?id=1347082.1347185.

Anupam Gupta, Debmalya Panigrahi, Bernardo Subercaseaux, and Kevin Sun. Augmenting online algorithms with ε-accurate predictions. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/0ea048312aa812b2711fe765a9e9ef05-Abstract-Conference.html.

Anupam Gupta, Gregory Kehne, and Roie Levin. Set covering with our eyes wide shut. In David P. Woodruff (ed.), *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pp. 4530–4553. SIAM, 2024. doi: 10.1137/1.9781611977912.160. URL https://doi.org/10.1137/1.9781611977912.160.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Peter J. Huber and Elvezio Ronchetti. *Robust statistics*. Wiley series in probability and statistics. Wiley, Hoboken, N.J, 2nd ed edition, 2009. ISBN 9780470129906.

Shaofeng H.-C. Jiang, Erzhi Liu, You Lyu, Zhihao Gavin Tang, and Yubo Zhang. Online facility location with predictions. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=DSQHjibtgKR.

Daniel M. Kane, Ilias Diakonikolas, Hanshen Xiao, and Sihan Liu. Online robust mean estimation. In David P. Woodruff (ed.), *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pp. 3197–3235. SIAM, 2024. doi: 10.1137/1.9781611977912.115. URL https://doi.org/10.1137/1.9781611977912.115.

Haim Kaplan, David Naori, and Danny Raz. Online weighted matching with a sample. 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. 1247–1272. SIAM, 2022. doi: 10.1137/1.9781611977073.52. URL https://doi.org/10.1137/1.9781611977073.52.

Haim Kaplan, David Naori, and Danny Raz. Almost tight bounds for online facility location in the random-order model. In Nikhil Bansal and Viswanath Nagarajan (eds.), *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pp. 1523–1544. SIAM, 2023. doi: 10.1137/1.9781611977554.CH56. URL https://doi.org/10.1137/1.9781611977554.ch56.

Haim Kaplan, David Naori, and Danny Raz. Competitive analysis with a sample and the secretary problem. *SIAM J. Comput.*, 54(6):1489–1513, 2025. doi: 10.1137/24M1650223. URL https://doi.org/10.1137/24m1650223.

Kevin A. Lai, Anup B. Rao, and Santosh S. Vempala. Agnostic estimation of mean and covariance. In Irit Dinur (ed.), *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, Hyatt Regency, New Brunswick, New Jersey, USA, October 9-11, 2016*, pp. 665–674. IEEE Computer Society, 2016. doi: 10.1109/FOCS.2016.76. URL https://doi.org/10.1109/FOCS.2016.76.

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*, volume 70 of *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.

Silvio Lattanzi, Debmalya Panigrahi, and Ola Svensson. Online rounding and learning augmented algorithms for facility location. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/551868edfc37eddf67662c072bb171ce-Abstract-Conference.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.

Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *J. ACM*, 68(4):24:1–24:25, 2021. doi: 10.1145/3447579. URL https://doi.org/10.1145/3447579.

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. In Tim Roughgarden (ed.), *Beyond the Worst-Case Analysis of Algorithms*, pp. 646–662. Cambridge University Press, 2020. doi: 10.1017/9781108637435.037. URL https://doi.org/10.1017/9781108637435.037.

Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving online algorithms via ML predictions. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9684–9693, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/73a427badebe0e32caa2e1fc7530b7f3-Abstract.html.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In Anna R. Karlin (ed.), *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, Cambridge, MA, USA, January 11-14, 2018*, volume 94 of *LIPIcs*, pp. 45:1–45:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPICS.ITCS.2018.45. URL https://doi.org/10.4230/LIPIcs.ITCS.2018.45.

Alexander Wei and Fred Zhang. Optimal robustness-consistency trade-offs for learning-augmented online algorithms. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/5bd844f11fa520d54fa5edec06ea2507-Abstract.html.

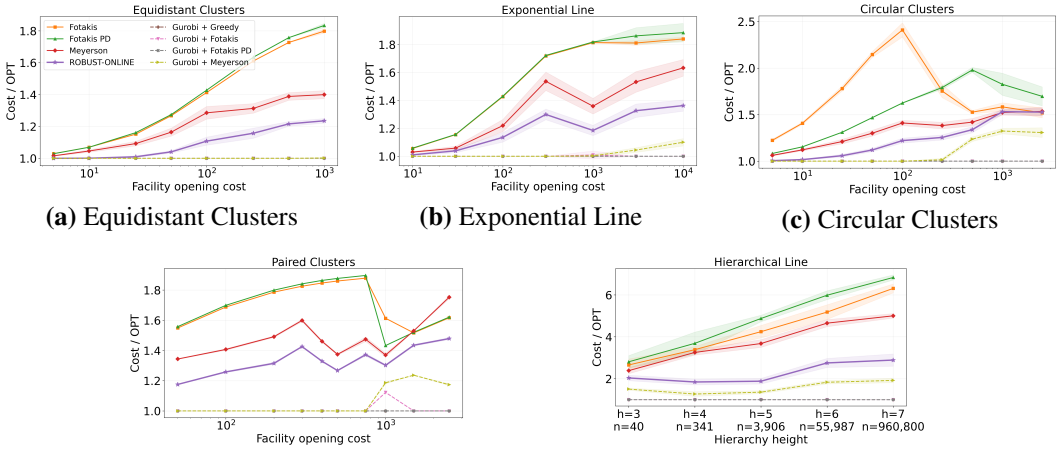
A Detailed Experiments

This appendix gives additional details for the experiments in Section 4, including the instance constructions and experimental setup.

Experimental setup. The experiments were run in Google Colab on an AMD EPYC 7B12 CPU with four physical cores and 50 GiB of RAM, using Gurobi Optimizer 13.0.3 to compute offline reference solutions. In the four cluster-based families, each cluster has a finite support around its center. The support points are sampled uniformly from small Euclidean balls for **Equidistant Clusters**, **Circular Clusters**, and **Paired Clusters**, and from short intervals for **Exponential Line**. Clients are sampled uniformly with replacement from these supports. Within each instance, the online input and clean offline sample are drawn independently from the same supports and with the same cluster counts. For **Equidistant Clusters**, **Exponential Line**, and **Circular Clusters**, we draw cluster proportions from a Dirichlet distribution with parameter 10 per cluster and round the resulting counts while preserving the total input size. In **Paired Clusters**, all sixteen clusters have equal numbers of observations, while in **Hierarchical Line**, the number of clients at each level depends on its depth. Only the cluster centers are candidate facilities in the four cluster-based families; all metric points are candidate facilities in the hierarchical experiments. All facilities have the same opening cost in a given experiment.

The same supports are used for all ten instances, while the online input and clean offline sample are independently resampled for each instance. For each instance and parameter value, Meyerson, ROBUST-ONLINE, and Gurobi + Meyerson are each run five times; Greedy, Fotakis, and Fotakis PD are deterministic on a fixed instance. Each cost is normalized by $\mathrm{OPT}(X)$. We first average the randomized runs within each instance and then average across the ten instances.

A.1 ACCURATE SAMPLES



(a) Equidistant Clusters **(b)** Exponential Line **(c)** Circular Clusters

(d) Paired Clusters **(e)** Hierarchical Line

Figure 3: **Accurate Samples.** Average approximation ratio versus the facility opening cost in panels (a)–(d), and versus hierarchy height in panel (e). Shaded regions show one standard deviation across ten independently sampled instances. Greedy is omitted for readability; its maximum ratios in panels (a)–(e) are $\approx 19, 298, 29, 95$, and 71650, respectively. The legend in (a) is shared across panels.

Equidistant Clusters. The instance has 16 cluster centers with pairwise distance 2. Each center has 100 support points sampled uniformly from the Euclidean ball of radius 0.1 around it. The online input and clean offline sample each contain $20{,}000$ observations with Dirichlet cluster counts. Clients from each cluster arrive consecutively, with no interleaving between clusters. The sixteen centers are the only candidate facilities. For facility costs $5, 10, 25, 50, 100, 250, 500, 1000$, Greedy's omitted ratios are approximately $19, 18, 16, 14, 11, 6.4, 3.9, 2.2$, respectively. ROBUST-ONLINE improves on Meyerson throughout: for instance, at cost 100, ROBUST-ONLINE cuts the excess cost over OPT by more than half. The Gurobi-initialized baselines remain nearly optimal.

Exponential Line. The instance has ten cluster centers at $1, 2, 4, \ldots, 512$. Each center has 100 support points drawn uniformly from the interval of radius 0.05 around it. The online input and clean offline sample each contain $20{,}000$ observations with Dirichlet cluster counts. Clients arrive one cluster at a time, from left to right, and the ten centers are the only candidate facilities. For facility costs $10, 30, 100, 300, 1000, 3000, 10000$, Greedy's omitted ratios are approximately $35, 62, 160, 300, 200, 72, 25$, respectively. ROBUST-ONLINE improves on Meyerson throughout the tested range, while the Gurobi-initialized baselines remain close to optimal.

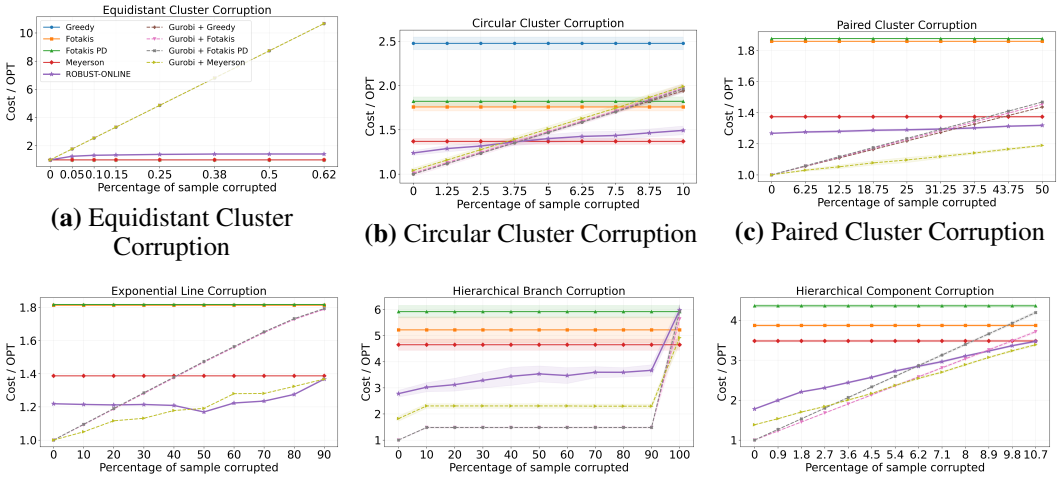
Circular Clusters. The instance has sixteen cluster centers equally spaced on the unit circle in $\mathbb{R}^2$. Each center has 100 support points sampled uniformly from the disk of radius 0.06 around it. The online input and clean offline sample each contain $20{,}000$ observations with Dirichlet cluster counts. Clients arrive one cluster at a time, in order around the circle. The centers are the only candidate facilities. For facility costs $5, 10, 25, 50, 100, 250, 500, 1000, 2500$, Greedy's omitted ratios are approximately $29, 27, 21, 16, 11, 5.4, 3.5, 2.5, 1.7$, respectively. ROBUST-ONLINE improves on Meyerson over most of the tested range, and the Gurobi-initialized baselines give the lowest ratios overall.

Paired Clusters. The instance has eight pairs of cluster centers whose midpoints are equally spaced on a circle of radius 5. Each pair lies along the tangent at its midpoint, and its two centers A_i and B_i are separated by distance 0.8. Each center has five support points sampled uniformly from the disk of radius 0.04 around it and contributes 1250 observations to both the online input and clean offline sample, giving $n = 20{,}000$. The cluster centers are the only candidate facilities. Clients from $A_1, \ldots, A_8$ arrive first, followed by clients from $B_1, \ldots, B_8$; within each cluster, the clients arrive consecutively. This order tests whether the algorithm opens a second facility when subsequent clients arrive near the other center in a pair. For facility costs $50, 100, 200, 300, 400, 500, 750, 1000, 1500, 2500$, Greedy's omitted ratios are approximately $95, 60, 34, 24, 18, 15, 10, 7.9, 6.4, 4.6$, respectively. ROBUST-ONLINE consistently improves on Meyerson, while the Gurobi-initialized baselines give the lowest ratios.

Hierarchical Line. For each $h \in \{3, 4, 5, 6, 7\}$, the instance is based on a complete binary tree of height h embedded on the line, with its root at 0. A vertex at depth $\ell < h$ and position p has children at $p \pm h^{-\ell}$. Each vertex at depth ℓ has five equally spaced support points in $[p - 0.25h^{-\ell}, p + 0.25h^{-\ell}]$. All $5(2^{h+1} - 1)$ support points are candidate facilities, each of cost h. A root-to-leaf path is selected by independent uniformly random left/right choices. At depth ℓ, the online input and clean offline sample each contain h^ℓ independent draws from the selected vertex's support, and clients arrive in increasing order of depth. Thus $n = \sum_{\ell=0}^{h} h^\ell$, giving $40, 341, 3906, 55{,}987, 960{,}800$ clients and $75, 155, 315, 635, 1275$ candidate facilities, respectively. Greedy's omitted ratios are approximately $6, 40, 408, 4435, 71650$ for $h = 3, 4, 5, 6, 7$. ROBUST-ONLINE substantially improves on the algorithms that do not use a sample for the larger hierarchies. The Gurobi-initialized baselines also achieve low cost ratios.

A.2 CORRUPTED SAMPLES

These experiments use adversarial sample corruption: a clean sample is generated and selected observations are then replaced. For each instance, the metric, facility costs, and online input remain fixed as the number of replacements increases. The horizontal axis gives the percentage of sample observations replaced.



(a) Equidistant Cluster Corruption

(b) Circular Cluster Corruption

(c) Paired Cluster Corruption

(d) Exponential Line Corruption

(e) Hierarchical Branch Corruption

(f) Hierarchical Component Corruption

Figure 4: **Corruption experiments.** Average approximation ratio versus the percentage of the offline sample corrupted. Shaded regions show one standard deviation across ten independently sampled instances. Greedy is omitted from panels (c)–(f), where its maximum ratios are ≈ 15, 193, 4522, and 3165, respectively. Gurobi + Greedy is additionally omitted in panels (d)–(f), where its maximum ratios are ≈ 780, 9387, and 311, respectively. The legend in (a) is shared across panels.

Equidistant Cluster Corruption. The instance has five true cluster centers and 125 additional centers, with every pair of centers at distance 2. All 130 centers are candidate facilities and have opening cost 1. Each true center has 100 support points sampled uniformly from the Euclidean ball of radius 0.0004 around it. The online input and clean offline sample each contain 20,000 observations from the five true clusters, with Dirichlet cluster counts. We replace k points in the offline sample with k of the spurious cluster centers, for $k \in \{0, 10, 20, 30, 50, 75, 100, 125\}$. Since the opening cost is smaller than the distance between centers, an optimal solution for the sample opens a facility at each of these additional centers. At 0.625% corruption ($k = 125$), the Gurobi-initialized baselines have ratio about 10.66, compared with 1.41 for ROBUST-ONLINE.

Circular Cluster Corruption. The instance has eight true centers on the unit circle and eight false centers at the same angles on a circle of radius 8. Each center has 100 support points sampled uniformly from the disk of radius 0.06 around it. These sixteen centers are the candidate facilities, all of cost 1250. The online input and clean offline sample each contain 20,000 observations from the true clusters, with Dirichlet cluster counts. Clients arrive one cluster at a time, in order around the circle. Eight disjoint groups of 250 sample observations are selected at random and assigned to distinct false centers in a random order. For $q = 0, \ldots, 8$, the observations in the first q groups are replaced by draws from the support around the corresponding false center, giving corruption levels from 0% to 10% in increments of 1.25%. In these runs, an optimal solution for the sample opens a facility at each false center represented in the corrupted sample. ROBUST-ONLINE deteriorates gradually with corruption, while the Gurobi-initialized baselines worsen more sharply. ROBUST-ONLINE performs comparably to Meyerson throughout, and better at lower corruption levels.

Paired Cluster Corruption. The instance has the same clusters, supports, and arrival order as the accurate-sample Paired Clusters experiment, with 1250 observations per cluster, $n = 20{,}000$, and facility cost 500. The pairs are considered in a fixed random order. For each selected pair, all 1250 sample observations from B_i are replaced by draws from the support around A_i. Corrupting $q = 0, \ldots, 8$ pairs gives corruption levels $0, 6.25, \ldots, 50\%$. ROBUST-ONLINE remains relatively stable and performs better than Meyerson throughout, while the Gurobi-initialized baselines generally worsen as observations from more B_i clusters are replaced. Greedy is omitted from the plot; its ratio is approximately 15 at every corruption level.

Exponential Line Corruption. The centers and supports are as in the accurate-sample Exponential Line experiment: ten centers at $1, 2, 4, \ldots, 512$, each with 100 support points drawn uniformly from an interval of radius 0.05. Each cluster contributes 2000 observations to both the online input and clean offline sample, giving $n = 20{,}000$, and each facility has cost 1000. Clients arrive one cluster at a time, from left to right. For $q = 0, \ldots, 9$, all sample observations from the first q clusters are replaced by draws from the final cluster's support, giving corruption levels $0, 10, \ldots, 90\%$. ROBUST-ONLINE performs better than Meyerson at every tested level. Greedy is omitted from the plot and has ratio approximately 193 at every corruption level. Gurobi + Greedy is also omitted; at corruption levels $0, 10, \ldots, 90\%$, its ratios are approximately $1, 1.1, 1.8, 3.9, 9.9, 25, 61, 147, 341, 780$, respectively.

Hierarchical Branch Corruption. The instance is the **Hierarchical Line** construction with $h = 6$ and $n = \sum_{\ell=0}^{6} 6^\ell = 55{,}987$. All points are available as candidate facilities. The true root-to-leaf path is selected at random. A second path starts at the same root and takes the opposite left/right choice at every depth. At depth ℓ, the online input and clean offline sample each contain 6^ℓ draws from the selected vertex's support, and clients arrive in increasing order of depth. For $\epsilon \in \{0, 0.1, \ldots, 1\}$, approximately $\epsilon \cdot 6^\ell$ observations at each depth $\ell \geq 1$ are replaced by draws from the second path's support at that depth. At $\approx 100\%$ corruption, every non-root sample observation has been replaced, so 55,986 of the 55,987 observations differ from the clean sample. ROBUST-ONLINE deteriorates moderately from 0% through 90% corruption and then worsens sharply at 100%, when the sample contains no observations from the true path beyond the root. Greedy is omitted from the plot and has ratio approximately 4522 at every corruption level. Gurobi + Greedy is also omitted; at corruption levels $0, 10, \ldots, 100\%$, its ratios are approximately $1, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 9387$, respectively. The jump at 100% occurs because Gurobi then opens facilities along the wrong path, far away from the true online clients.

Hierarchical Component Corruption. The instance consists of thirteen **Hierarchical Line** components with $h = 5$, scaled by 0.25 and placed with pairwise center distance 2. Twelve components have weight 1 and one has weight 100. Each component has one support point at each of its vertices, and all points are available as candidate facilities. A root-to-leaf path is selected independently in each component. A component of weight w contributes $w \cdot 5^{\ell}$ observations at depth ℓ. Clients from the large component arrive first, followed by the smaller components one at a time; within each component, clients arrive in increasing order of depth. Each smaller component contributes 3906 observations and the large component contributes 390,600, giving $n = 437{,}472$ in each sequence. Because each vertex has a single support point, the online input and clean offline sample coincide once the paths have been selected. For $q = 0, \ldots, 12$, the sample observations from q smaller components are replaced by draws with replacement from the large component's clean sample. At 10.7% corruption ($q = 12$), no sample observations remain in the twelve smaller components. ROBUST-ONLINE gradually loses its advantage over Meyerson as observations from more of the smaller components are replaced, and the Gurobi-initialized baselines also deteriorate. Greedy is omitted from the plot and has ratio approximately 3165 at every corruption level. Gurobi + Greedy is also omitted; for $q = 0, \ldots, 12$, its ratios are approximately $1, 23, 53, 75, 101, 128, 154, 179, 208, 233, 257, 284, 311$, respectively.

A.3 RUNNING TIME AND SCALABILITY

Let $N = |V|$ be the number of points in the metric space (V, d) and $M = |\mathcal{F}|$ the number of candidate facilities. Constructing the $N \times N$ distance matrix takes $O(N^2)$ time. Let $L = 1 + \max_{f \in \mathcal{F}} \lfloor \log_2 c_f \rfloor$ be the number of facility-type indices considered by the implementation. For each point in V, we find the closest facility of each type and precompute the distances to the closest facilities of types at most j, for every j. These computations take $O(NM + NL)$ time.

A single execution of ROBUST-ONLINE simulates Meyerson on a random ordering of the n samples and then processes the n online clients. Scanning all L facility types for each observation takes $O(nL)$ time across the two passes. Whenever a facility opens, the implementation updates the distance from every point in V to its nearest open facility, taking $O(N)$ time. Each pass opens at most M facilities, so these updates take $O(NM)$ time in total. Thus, including preprocessing, the total running time is $O(N^2 + NM + NL + nL)$.

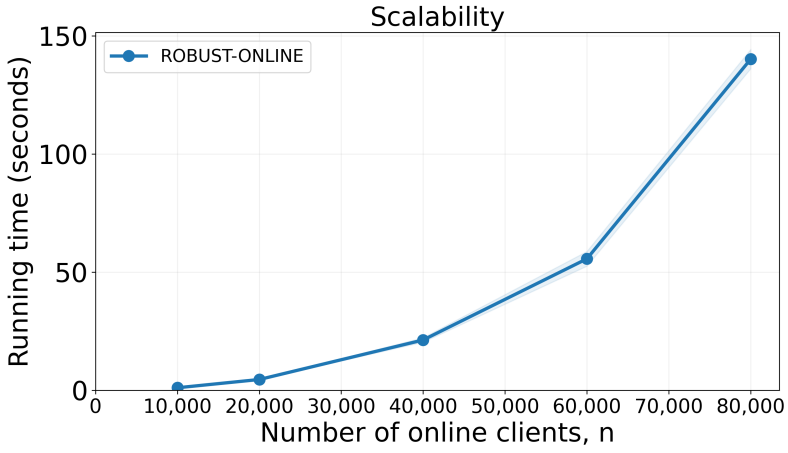


Figure 5: Average running time of ROBUST-ONLINE versus the number of online clients in **Equidistant Clusters**, with all n support points available as facilities at unit opening cost. Each point shows the mean over five runs on the same instance with the same random seed; shaded regions show one standard deviation.

We evaluate the running time of ROBUST-ONLINE on **Equidistant Clusters** instances with $n \in \{10{,}000, 20{,}000, 40{,}000, 60{,}000, 80{,}000\}$ to test scalability. The instance has ten cluster centers in $\mathbb{R}^{10}$ with pairwise distance 2. Each center has $n/10$ support points sampled uniformly from the Euclidean ball of radius 0.1 around it. The online input and clean offline sample each contain n

observations, drawn independently and uniformly with replacement from these supports, with $n/10$ observations per cluster. Clients from each cluster arrive consecutively, and all n support points are candidate facilities, each with opening cost 1.

For each input size, we average the running time over five executions on the same instance with the same random seed. The measured running time includes construction of the distance matrix, facility-type preprocessing, the offline simulation, and the online phase. We compile the Numba routines on a small instance before measuring running times. Figure 5 shows the mean running time, with shaded regions indicating one standard deviation across the five runs. Here $N = M = n$ and $L = 1$, so the running-time bound above reduces to $O(n^2)$. A log–log fit to the mean running times yields a slope of 2.299, consistent with the theoretical running time bound.

B Omitted Proofs from Section 3

We begin with the two estimates for Meyerson's algorithm used in the proof of Lemma 3.3. Throughout these proofs, $F_{<v}$ includes all facilities open immediately before v arrives. We first prove the single time-step bound for Meyerson's algorithm.

Lemma B.1 (Single time-step Meyerson bound)**.** *Let v be an arriving client, and let $F_{<v}$ be the facilities open just before v arrives. The expected cost incurred by Meyerson's algorithm on v, conditioned on v and $F_{<v}$, is at most*

$$O(\kappa_{F_{<v}}(v)).$$

Proof. Recall that $\kappa_{F_{<v}}(v) = \min_{f\in\mathcal{F}}\{\mathbf{1}\{f \notin F_{<v}\} \cdot c_f + d(f, v)\}$ is the cheapest way to serve v when all the facilities in $F_{<v}$ are open. Let

$$f' = \operatorname{argmin}_{f\in\mathcal{F}} \{\mathbf{1}\{f \notin F_{<v}\} \cdot c_f + d(f, v)\}.$$

For $-1 \le j \le s$, let $\delta_j = d(\mathcal{F}_{\le j} \cup F_{<v}, v)$. For $0 \le j \le s$, let $p_j = \min\left(\frac{\delta_{j-1}-\delta_j}{2^j}, 1\right)$ be the probability that ON opens a facility of type j when v arrives.

- First suppose $f' \in F_{<v}$. Then $\kappa_{F_{<v}}(v) = d(f', v)$ and $\delta_{-1} = d(F_{<v}, v) \le d(f', v) = \kappa_{F_{<v}}(v)$. The connection cost is at most δ_{-1}. The expected facility-opening cost is at most

$$\sum_{j=0}^{s} 2^{j+1} p_j \le \sum_{j=0}^{s} 2^{j+1} \cdot \frac{\delta_{j-1} - \delta_j}{2^j} = 2\sum_{j=0}^{s}(\delta_{j-1} - \delta_j) \le 2\delta_{-1} \le 2\kappa_{F_{<v}}(v).$$

  Thus, the total expected cost is at most $3\kappa_{F_{<v}}(v)$ in this case.

- Now suppose $f' \notin F_{<v}$. Let i be the type of f', so $2^i \le c_{f'} < 2^{i+1}$. Then $\kappa_{F_{<v}}(v) = c_{f'} + d(f', v) \ge 2^i + d(f', v)$, and since $f' \in \mathcal{F}_{\le i}$, we have $\delta_i \le d(f', v)$. The expected facility-opening cost is at most

$$\sum_{j=0}^{i} 2^{j+1} p_j + \sum_{j=i+1}^{s} 2^{j+1} p_j \le \sum_{j=0}^{i} 2^{j+1} + \sum_{j=i+1}^{s} 2(\delta_{j-1} - \delta_j) \le 2^{i+2} + 2\delta_i = O(\kappa_{F_{<v}}(v)).$$

  For the connection cost, let $r = \max(\{-1\} \cup \{j \le i : p_j = 1\})$. If $r \ge 0$, then after processing v there is an open facility at distance at most δ_r from v; if $r = -1$, the connection cost is at most δ_{-1}. In either case, because $p_j < 1$ for every $r < j \le i$, we have $\delta_{j-1} - \delta_j < 2^j$ for every $r < j \le i$, and therefore

$$\delta_r \le \delta_i + \sum_{j=r+1}^{i} (\delta_{j-1} - \delta_j) \le \delta_i + \sum_{j=r+1}^{i} 2^j \le \delta_i + 2^{i+1} = O(\kappa_{F_{<v}}(v)).$$

  Thus, the total expected cost is $O(\kappa_{F_{<v}}(v))$ in the second case as well.

□

To prove the ball-cost lemma, we will need the following elementary waiting-time bound.

Lemma B.2 (Waiting-time bound). *Consider a sequence of rounds $1, 2, \ldots$, and let $\mathcal{H}_{i-1}$ denote the history before round i. Conditioned on $\mathcal{H}_{i-1}$, a success occurs in round i with probability $p(\mathcal{H}_{i-1})$. Successful rounds incur a cost of at most A, for some constant $A > 0$, and unsuccessful rounds incur zero cost. Then the expected total cost incurred up to and including the first successful round is at most A.*

Proof. Let τ be the index of the first successful round, setting $\tau = \infty$ if no round succeeds. The event $I_i = \{\tau \geq i\}$ is determined by the history before round i. Therefore,

$$\mathbb{E}[I_i \cdot p(\mathcal{H}_{i-1})] = \Pr[\tau = i].$$

Consequently, the expected total cost incurred up to and including the first successful round is at most

$$\mathbb{E}\left[\sum_{i=1}^{\infty} I_i \cdot A \cdot p(\mathcal{H}_{i-1})\right] = A\sum_{i=1}^{\infty} \mathbb{E}[I_i \cdot p(\mathcal{H}_{i-1})] = A\sum_{i=1}^{\infty} \Pr[\tau = i] \leq A. \qquad \square$$

We can now prove the ball-cost lemma stated in the main text.

Lemma 3.2 (Ball cost). *Let F_0 be the initial open set of facilities. Fix a facility $g \in \mathcal{F}$ and a subsequence $S \subseteq X$ of clients satisfying $d(v, g) \leq R$ for every $v \in S$. Then*

$$\mathbb{E}[\mathsf{cost}_{\mathrm{ON}}(S)] \leq O\left(\mathbf{1}\{g \notin F_0\} \cdot c_g + \sum_{v \in S} d(v, g) + |S|R\right).$$

The clients in S need not arrive consecutively. Here $\mathsf{cost}_{\mathrm{ON}}(S)$ is the cost of facility openings and connections made by ON *in response to clients in S.*

Proof. If $g \in F_0$, then applying Lemma B.1 for each $v \in S$ gives

$$\mathbb{E}[\mathsf{cost}_{\mathrm{ON}}(S)] \leq O\left(\sum_{v \in S} d(v, g)\right),$$

so suppose that $g \notin F_0$. Let r_g be the type of g. Recall that $\mathcal{F}_{\leq r}$ is the set of facilities of type at most r, and $F_{<v}$ is the set of facilities open just before $v \in S$ arrives. For every $r \in \{-1, 0, \ldots, s\}$, let

$$\delta_r(v) = d(\mathcal{F}_{\leq r} \cup F_{<v}, v)$$

denote the distance from v to the closest facility that is of type at most r or is already open. When $v \in S$ arrives, ON opens the closest facility of type r to v with probability $\min\left(\frac{\delta_{r-1}(v) - \delta_r(v)}{2^r}, 1\right)$. Consequently, the expected facility cost incurred due to the arrival of v (conditioned on $F_{<v}$) is at most

$$F(v) \leq \sum_{r=0}^{s} 2^{r+1} \min\left(\frac{\delta_{r-1}(v) - \delta_r(v)}{2^r}, 1\right) \leq \sum_{r=0}^{s} 2(\delta_{r-1}(v) - \delta_r(v)) \leq 2\delta_{-1}(v). \tag{7}$$

Moreover, we claim that the expected assignment cost $A(v)$ due to v is at most the sum of the expected facility cost due to v and $d(v, g)$, i.e.,

$$A(v) \leq F(v) + d(v, g).$$

Indeed, let p be the highest index such that $\delta_{p-1}(v) - \delta_p(v) > 2^p$, setting $p = -1$ if no such index exists. Then, we can always assign v to a facility at distance at most $\delta_p(v)$. The expected facility cost incurred by v due to types at least $p + 1$ is at least

$$\sum_{r=p+1}^{s} 2^r \cdot \frac{\delta_{r-1}(v) - \delta_r(v)}{2^r} = \delta_p(v) - \delta_s(v) \geq \delta_p(v) - d(v, g).$$

As a result, we have

$$\sum_{v \in S}(F(v) + A(v)) \leq 2\sum_{v \in S} F(v) + \sum_{v \in S} d(v, g). \tag{8}$$

We now bound the expected facility cost incurred due to clients in S. Label the clients in S sequentially according to their order of arrival as $v_1, v_2, \ldots, v_{|S|}$. The expected facility cost incurred by clients in S due to types at least $r_g + 1$ is at most

$$\sum_{v_i \in S} \sum_{r=r_g+1}^{s} 2^{r+1} \cdot \frac{\delta_{r-1}(v_i) - \delta_r(v_i)}{2^r} \leq \sum_{v_i \in S} 2\delta_{r_g}(v_i) \leq \sum_{v_i \in S} 2d(v_i, g). \tag{9}$$

Let $\rho_r = d(\mathcal{F}_{\leq r}, g)$ denote the distance from g to the closest facility $\widetilde{f}_r$ of type at most r. Define $\mathcal{E}_r$, for $0 \leq r \leq r_g$, to be the event that ON opens a facility f such that

$$d(f, g) \leq \rho_r + 2R.$$

Also define $\mathcal{E}$ to be the event that ON opens a facility f such that

$$d(f, g) \leq 12R.$$

Note that if any $v_i \in S$ opens a facility f of type at least r, then $\mathcal{E}_r$ occurs. Indeed, if the type of f is $r' \geq r$, then f must be the closest facility of type at most r' to v_i (by definition of ON). Thus, using the triangle inequality,

$$d(v_i, f) \leq d(v_i, \widetilde{f}_r) \leq d(v_i, g) + d(g, \widetilde{f}_r) \leq R + \rho_r,$$

and hence

$$d(f, g) \leq d(f, v_i) + d(v_i, g) \leq R + \rho_r + R = \rho_r + 2R.$$

Let $\widetilde{F}(r)$ denote the expected facility cost incurred by clients in S due to facilities of type r up to and including the first occurrence of $\mathcal{E}_r$. Since the opening cost of a facility of type r is at most 2^{r+1}, by Lemma B.2 we have $\widetilde{F}(r) \leq 2^{r+1}$. Consequently,

$$\sum_{r=0}^{r_g} \widetilde{F}(r) \leq \sum_{r=0}^{r_g} 2^{r+1} \leq 2^{r_g+2} \leq 4c_g. \tag{10}$$

Note that once $\mathcal{E}_{r_g}$ occurs, the event $\mathcal{E}$ must also have occurred, since $\rho_{r_g} = d(\mathcal{F}_{\leq r_g}, g) = 0$ and hence the facility opened when $\mathcal{E}_{r_g}$ occurs is within distance $2R$ of g. Consider any client $v_i \in S$ arriving after $\mathcal{E}_r$ has occurred but before $\mathcal{E}$ has occurred.

The expected facility cost incurred due to v_i from types at most r is at most

$$\sum_{j=0}^{r} 2^{j+1} \cdot \min\left(\frac{\delta_{j-1}(v_i) - \delta_j(v_i)}{2^j}, 1\right) \leq 2\delta_{-1}(v_i) - 2\delta_r(v_i). \tag{11}$$

The expected facility cost incurred due to v_i from types greater than r is at least

$$\sum_{j=r+1}^{r_g} 2^j \cdot \min\left(\frac{\delta_{j-1}(v_i) - \delta_j(v_i)}{2^j}, 1\right) \tag{12}$$

If there is no facility of type at most r at distance $\delta_r(v_i)$ from v_i, $\delta_r(v_i) = \delta_{-1}(v_i)$ and all expected facility cost due to v_i comes from types greater than r. Otherwise, suppose that there is a facility of type at most r at distance $\delta_r(v_i)$ from v_i. This facility is at distance at most $\delta_r(v_i) + R$ away from g, and hence $\delta_r(v_i) + R \geq \rho_r$.

Since $\mathcal{E}_r$ has occurred, there is an open facility within distance $\rho_r + 2R$ of g. Therefore,

$$\delta_{-1}(v_i) = d(F_{<v_i}, v_i) \leq d(F_{<v_i}, g) + d(g, v_i) \leq \rho_r + 3R.$$

Since $\mathcal{E}_r$ has occurred but $\mathcal{E}$ has not, we have $\rho_r + 2R > 12R$, which implies that $\rho_r > 10R$. Consequently, (11) is at most $2(\rho_r + 3R) - 2(\rho_r - R) \leq 8R$. If $\delta_{j-1}(v_i) - \delta_j(v_i) \leq 2^j$ for every $r < j \leq r_g$, then (12) is at least $\delta_r(v_i) - \delta_{r_g}(v_i) \geq (\rho_r - R) - R > 8R$. Thus, in this case the expected facility cost incurred due to v_i from types at most r is at most the cost incurred from types greater than r. Otherwise, suppose $\delta_{j-1}(v_i) - \delta_j(v_i) > 2^j$ for some $r < j \leq r_g$. Then a facility of type j is always opened, and incurs cost at least 2^j. On the other hand, the expected facility cost due to types at most r is less than $\sum_{k=0}^{r} 2^{k+1} \leq 2^{r+2} \leq 2^{j+1}$. Thus, in this case the expected facility

cost incurred due to v_i from types at most r is at most twice the cost incurred from types greater than r.

Now, consider a client v_i arriving before $\mathcal{E}$ occurs. The contribution to the expected facility cost from types larger than r_g is accounted for in (9), so it suffices to bound the contribution from types at most r_g. If none of the events $\mathcal{E}_0, \ldots, \mathcal{E}_{r_g}$ has occurred, then all of its facility cost due to types at most r_g is charged before the corresponding events $\mathcal{E}_r$ and is accounted for in (10). Otherwise, let r be the largest index such that $\mathcal{E}_r$ has occurred. At least a third of the expected facility cost due to v_i is charged to types at least $r+1$. For every $r < t \le r_g$, the event $\mathcal{E}_t$ has not yet occurred, so the type-t contribution is again accounted for in (10). Therefore, the expected facility cost incurred by clients in S before $\mathcal{E}$ occurs is at most

$$3\left(4c_g + \sum_{v_i \in S} 2d(v_i, g)\right) = 12c_g + \sum_{v_i \in S} 6d(v_i, g). \tag{13}$$

Using (7) and the definition of the event $\mathcal{E}$, the expected facility cost due to $v_i \in S$ after $\mathcal{E}$ has occurred is at most

$$2\delta_{-1}(v_i) = 2d(F_{<v_i}, v_i) \le 2d(F_{<v_i}, g) + 2d(v_i, g) \le 24R + 2d(v_i, g).$$

Combining this with (13), the total expected facility cost due to clients in S is at most

$$12c_g + \sum_{v_i \in S} 6d(v_i, g) + \sum_{v_i \in S} (24R + 2d(v_i, g)) = 12c_g + \sum_{v_i \in S} (24R + 8d(v_i, g)).$$

Combining this bound with (8) proves Lemma 3.2. □

We next turn to the proof of the uncorrupted-prefix bounds. Both noise models use the following lemma to control the cost of ON on clients that cannot be charged directly to the random-order run on the uncorrupted prefix.

Lemma B.3. *Let $X = (v^1, \ldots, v^n)$ be a sequence of independent random clients, where $v^i \sim D_i$ and each D_i is supported on V. For each $i \in [n]$, let $\tilde{D}_i$ be a nonnegative vector over V such that $\tilde{D}_i(u) \le D_i(u)$ for every $u \in V$, and let $\gamma = \sum_{i=1}^n \|\tilde{D}_i\|_1$. Then*

$$\mathbb{E}\left[\sum_{i=1}^n \mathrm{OPT}(\{v^i\}) \cdot \mathbf{1}\{v^i \sim \tilde{D}_i\}\right] \le (1+\gamma) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

Proof. For $t \ge 0$, let $a_i(t) = \Pr[v^i \sim \tilde{D}_i,\ \mathrm{OPT}(\{v^i\}) > t]$. Hence $\sum_{i=1}^n a_i(t) \le \sum_{i=1}^n \|\tilde{D}_i\|_1 = \gamma$ for every $t \ge 0$. For every nonnegative random variable Y, we have $\mathbb{E}[Y] = \int_0^\infty \Pr[Y > t]dt$. Consequently,

$$\mathbb{E}\left[\sum_{i=1}^n \mathrm{OPT}(\{v^i\}) \cdot \mathbf{1}\{v^i \sim \tilde{D}_i\}\right] = \int_0^\infty \sum_{i=1}^n a_i(t)dt.$$

We now lower bound $\mathbb{E}_X[\mathrm{OPT}(X)]$. Let $\tilde{a}_i(t) = \Pr[\mathrm{OPT}(\{v^i\}) > t]$. Since any feasible facility location solution for X must serve every client v^i, we have $\mathrm{OPT}(X) \ge \max_{i \in [n]} \mathrm{OPT}(\{v^i\})$. Therefore,

$$\mathbb{E}_X[\mathrm{OPT}(X)] \ge \int_0^\infty \Pr\left[\max_{i \in [n]} \mathrm{OPT}(\{v^i\}) > t\right] dt = \int_0^\infty \left(1 - \prod_{i=1}^n (1 - \tilde{a}_i(t))\right) dt.$$

Since $\tilde{a}_i(t) \ge a_i(t)$ and $1 - x \le e^{-x}$,

$$1 - \prod_{i=1}^n (1 - \tilde{a}_i(t)) \ge 1 - \prod_{i=1}^n (1 - a_i(t)) \ge 1 - \exp\left(-\sum_{i=1}^n a_i(t)\right). \tag{14}$$

Finally, $1 - e^{-x} \ge x/(1+x)$ for $x \ge 0$, and $\sum_{i=1}^n a_i(t) \le \gamma$. Therefore,

$$1 - \exp\left(-\sum_{i=1}^n a_i(t)\right) \ge \frac{1}{1+\gamma} \sum_{i=1}^n a_i(t). \tag{15}$$

Substituting (14) and (15) into the lower bound for $\mathbb{E}_X[\mathrm{OPT}(X)]$, we conclude

$$\mathbb{E}_X[\mathrm{OPT}(X)] \geq \frac{1}{1+\gamma}\int_0^\infty \sum_{i=1}^n a_i(t)dt = \frac{1}{1+\gamma}\mathbb{E}\left[\sum_{i=1}^n \mathrm{OPT}(\{v^i\})\cdot \mathbf{1}\{v^i \sim \tilde{D}_i\}\right].$$

Rearranging proves the lemma. □

We will also repeatedly use the following immediate consequence of the definition of κ.

Observation B.4. *For every $0 \leq r < n$, the cost incurred by* RO *when processing $\hat{X}_{r+1}$ is at least $\kappa_{F_r^{\mathrm{RO}}}(\hat{X}_{r+1})$.*

Proof. This follows directly from the definition of κ: RO serves $\hat{X}_{r+1}$ starting from the facility set F_r^{RO}. □

We first prove the second bound in Lemma 3.4 for the distributional noise model.

Proof of the distributional residual bound in Lemma 3.4. Recall that G_i is the indicator that the offline sample $\hat{v}^i$ is drawn from the (normalized) distinct part $\hat{Q}_i$, and let $J = \sum_{i=1}^n G_i$. Condition on the vector $G = (G_1, \ldots, G_n)$. Thus, there are exactly J corrupted offline samples and $n - J$ uncorrupted offline samples. We split each online draw according to whether it comes from the common part P_i or the distinct part Q_i:

$$\mathbb{E}[\Gamma_\zeta \mid G] = \mathbb{E}\left[\sum_{i=1}^n \kappa_{F_\zeta^{\mathrm{RO}}}(v^i) \mid G\right] \leq \sum_{i=1}^n \mathbb{E}\left[\kappa_{F_\zeta^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G\right] + \sum_{i=1}^n \mathbb{E}\left[\mathrm{OPT}(\{v^i\})\cdot \mathbf{1}\{v^i \sim Q_i\} \mid G\right]. \tag{16}$$

We bound the two terms in (16) separately.

Claim B.5. *For every fixed vector G,*

$$\sum_{i=1}^n \mathbb{E}\left[\kappa_{F_\zeta^{\mathrm{RO}}}(v^i)\cdot \boldsymbol{1}\{v^i \sim P_i\} \mid G\right] \leq (J+1)\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G] + J\cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

Proof. Let $\sigma(i)$ be the position of $\hat{v}^i$ in the random order of $\hat{X}$. We split the indices into three groups. First consider indices with $G_i = 0$ and $\sigma(i) \leq \zeta$. These correspond to uncorrupted samples that appear in $\hat{X}_{\leq\zeta}$. Since F_ζ^{RO} contains $F_{\sigma(i)-1}^{\mathrm{RO}}$ for every such index i, monotonicity gives

$$\mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)\leq\zeta} \kappa_{F_\zeta^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G\right] \leq \mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)\leq\zeta} \kappa_{F_{\sigma(i)-1}^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G\right].$$

For every fixed index i with $G_i = 0$, conditioned on the event $\sigma(i) \leq \zeta$, the random variable $\hat{v}^i$ has distribution $P_i/(1-q_i)$, while the online draw v^i contributes to the left-hand side below only when it is drawn from P_i, in which case it has the same conditional distribution. Since $\hat{v}^i$ is independent of $F_{\sigma(i)-1}^{\mathrm{RO}}$,

$$\mathbb{E}\left[\kappa_{F_{\sigma(i)-1}^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G, \sigma(i) \leq \zeta\right] = (1-q_i)\cdot\mathbb{E}\left[\kappa_{F_{\sigma(i)-1}^{\mathrm{RO}}}(\hat{v}^i) \mid G, \sigma(i) \leq \zeta\right] \leq \mathbb{E}\left[\kappa_{F_{\sigma(i)-1}^{\mathrm{RO}}}(\hat{v}^i) \mid G, \sigma(i) \leq \zeta\right].$$

Summing this across indices i such that $G_i = 0$ and $\sigma(i) \leq \zeta$,

$$\begin{aligned}\mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)\leq\zeta} \kappa_{F_{\sigma(i)-1}^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G\right] &\leq \mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)\leq\zeta} \kappa_{F_{\sigma(i)-1}^{\mathrm{RO}}}(\hat{v}^i) \mid G\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{\zeta} \kappa_{F_{j-1}^{\mathrm{RO}}}(\hat{X}_j) \mid G\right] \\ &\leq \mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G],\end{aligned}$$

where the last inequality follows from Observation B.4.

Next consider indices with $G_i = 0$ and $\sigma(i) > \zeta$. These correspond to uncorrupted samples that do not appear in $\hat{X}_{\le\zeta}$. If $J = 0$, there are no such indices, so assume $J \ge 1$. We decompose according to the value of ζ:

$$\mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)>\zeta} \kappa_{F_\zeta^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G\right]$$
$$= \sum_{q=0}^{n-J-1} \Pr[\zeta = q \mid G]\cdot \mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)>q} \kappa_{F_q^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G, \zeta = q\right]. \tag{17}$$

Fix q. Under the event $\zeta = q$, the first q positions are uncorrupted and position $q+1$ is corrupted. Thus, the indices in the inner sum are the $n-J-q$ uncorrupted indices that have not appeared in the first q positions. Now fix the first q positions and one such index i. Since $G_i = 0$, the offline draw $\hat{v}^i$ is distributed as $P_i/(1-q_i)$ and is independent of F_q^{RO}. The online draw v^i contributes only when it is drawn from P_i, in which case it has the same distribution. Thus, the expected contribution of index i is at most the expected contribution of the corresponding offline draw $\hat{v}^i$. The distribution of the first q uncorrupted positions is the same under $\zeta = q$ and $\zeta > q$. Conditioned on G and $\zeta > q$, the index of $\hat{X}_{q+1}$ is uniform among the $n-J-q$ uncorrupted indices not appearing in the first q positions, so we have

$$\mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)>q} \kappa_{F_q^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \,\middle|\, G, \zeta = q\right] \le \mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)>q} \kappa_{F_q^{\mathrm{RO}}}(\hat{v}^i) \,\middle|\, G, \zeta = q\right]$$
$$= \mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)>q} \kappa_{F_q^{\mathrm{RO}}}(\hat{v}^i) \,\middle|\, G, \zeta > q\right]$$
$$= (n-J-q)\cdot \mathbb{E}\left[\kappa_{F_q^{\mathrm{RO}}}(\hat{X}_{q+1}) \,\middle|\, G, \zeta > q\right].$$

Substituting this into the previous equation, we get

$$\mathbb{E}\left[\sum_{i:G_i=0,\ \sigma(i)>\zeta} \kappa_{F_\zeta^{\mathrm{RO}}}(v^i)\cdot \mathbf{1}\{v^i \sim P_i\} \mid G\right] \le \sum_{q=0}^{n-J-1} \Pr[\zeta = q \mid G]\cdot(n-J-q)\cdot \mathbb{E}[\kappa_{F_q^{\mathrm{RO}}}(\hat{X}_{q+1}) \mid G, \zeta > q]. \tag{18}$$

We now compare this to the cost incurred by RO on the uncorrupted prefix. For each q, the client $\hat{X}_{q+1}$ is present in $\hat{X}_{\le\zeta}$ on the event $\zeta > q$. By Observation B.4, the cost incurred by RO in round $q+1$ is at least $\kappa_{F_q^{\mathrm{RO}}}(\hat{X}_{q+1})$. Therefore,

$$\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G] \ge \sum_{q=0}^{n-J-1} \Pr[\zeta > q \mid G]\cdot \mathbb{E}[\kappa_{F_q^{\mathrm{RO}}}(\hat{X}_{q+1}) \mid G, \zeta > q]. \tag{19}$$

Note that for every $0 \le q \le n-J-1$, we have $\Pr[\zeta = q \mid G]\cdot(n-J-q) = J\cdot\Pr[\zeta > q \mid G]$. Indeed,

$$\Pr[\zeta = q \mid G] = \frac{(n-J)_q}{(n)_q}\cdot\frac{J}{n-q}, \qquad \Pr[\zeta > q \mid G] = \frac{(n-J)_{q+1}}{(n)_{q+1}},$$

where $(a)_t = a(a-1)\cdots(a-t+1)$ denotes the falling factorial. Combining this with (18) and (19), the contribution of indices with $G_i = 0$ and $\sigma(i) > \zeta$ is at most $J\cdot\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G]$. Finally, there are exactly J indices with $G_i = 1$. For each such index,

$$\kappa_{F_\zeta^{\mathrm{RO}}}(v^i) \le \mathrm{OPT}(\{v^i\}) \le \mathrm{OPT}(X).$$

Thus, their total contribution is at most $J\cdot\mathbb{E}_X[\mathrm{OPT}(X)]$. Combining the three bounds proves the claim. □

Claim B.6. *For every fixed vector G,*

$$\sum_{i=1}^{n} \mathbb{E}\left[\mathrm{OPT}(\{v^i\}) \cdot \mathbf{1}\{v^i \sim Q_i\} \mid G\right] \le O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

Proof. The online input is independent of G. Since $\sum_i \|Q_i\|_1 = \ell/2$, the claim follows from Lemma B.3. □

Claim B.7. *We have*

$$\mathbb{E}_G[J \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G]] \le \frac{\ell}{2} \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}}].$$

Proof. We use a monotonicity property of $\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G]$ as a function of the vector G. For each index i, draw a potential sample Y_i from the common part P_i and a potential sample Z_i from the distinct part $\hat{Q}_i$ of $\hat{D}_i$. The vector G decides whether the offline sample $\hat{v}^i$ is the uncorrupted value Y_i or the corrupted value Z_i. Fix all coordinates of G except G_i, as well as all potential samples and randomness of RO. If G_i is changed from 0 to 1, then the ith sample $\hat{v}^i$ is corrupted. If $\hat{v}^i$ appeared before the first corrupted sample under $G_i = 0$, then the uncorrupted prefix becomes a truncation of the original prefix. If it appeared after the first corrupted sample, the uncorrupted prefix is unchanged. Thus, for every fixed realization of the potential samples and randomness of RO, the cost C_ζ^{RO} is non-increasing in G_i. Taking expectations over these realizations, $\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G]$ is also non-increasing in G_i. Since G_i is independent of the remaining coordinates of G, it follows that

$$\mathbb{E}_G[\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G] \mid G_i = 1] \le \mathbb{E}_G[\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G] \mid G_i = 0],$$

and hence $\mathbb{E}_G[\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G] \mid G_i = 1] \le \mathbb{E}[C_\zeta^{\mathrm{RO}}]$. Therefore,

$$\begin{aligned}
\mathbb{E}_G[J \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G]] &= \sum_{i=1}^{n} \mathbb{E}_G[G_i \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G]] && (\text{using } J = \sum_{i=1}^{n} G_i) \\
&= \sum_{i=1}^{n} q_i \cdot \mathbb{E}_G[\mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G] \mid G_i = 1] && (\text{using } \mathbb{E}[G_i] = \|\hat{Q}_i\|_1 = q_i) \\
&\le \left(\sum_{i=1}^{n} q_i\right) \mathbb{E}[C_\zeta^{\mathrm{RO}}] = \frac{\ell}{2}\mathbb{E}[C_\zeta^{\mathrm{RO}}]. && (\text{using } \sum_{i=1}^{n} q_i = \ell/2)
\end{aligned}$$

□

Applying Claims B.5 and B.6 to (16) and taking expectations over G gives

$$\mathbb{E}[\Gamma_\zeta] \le \mathbb{E}[C_\zeta^{\mathrm{RO}}] + \mathbb{E}_G[J \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}} \mid G]] + \mathbb{E}_G[J] \cdot \mathbb{E}_X[\mathrm{OPT}(X)] + O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

Since $\mathbb{E}_G[J] = \sum_i \mathbb{E}_G[G_i] = \sum_i q_i = \ell/2$, Claim B.7 gives

$$\mathbb{E}[\Gamma_\zeta] \le O(\ell) \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}}] + O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

Finally, (6) gives

$$\mathbb{E}[C_\zeta^{\mathrm{RO}}] \le O(1) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

This proves the distributional residual bound. □

We next turn to the adversarial noise model. The main difficulty is that conditioning on a prefix being uncorrupted can distort the distribution of the prefix.

For a labeled prefix h of length t, let $U(h)$ be the set of labels in $[n]$ that do not appear in h. For each $i \in U(h)$, define

$$\mathrm{Bad}_i(h) := \left\{ x \in V : \Pr[\hat{X}_{t+1} = (i, x) \mid \hat{X}_{\le t} = h, \zeta > t] < \frac{D_i(x)}{2(n-t)} \right\}.$$

Lemma B.8 (Stochastic distortion)**.** *Suppose $\ell \le n/4$. For every integer $0 \le t \le n/(4\ell)$,*

$$\mathbb{E}_{\hat{X}_{\le t}}\left[\sum_{i \in U(\hat{X}_{\le t})} D_i(\mathrm{Bad}_i(\hat{X}_{\le t})) \,\middle|\, \zeta > t \right] \le O(\ell).$$

Proof of Lemma B.8. Let $r = n - \ell$ be the number of uncorrupted samples, and let $\mathcal{H}_t$ be the set of all possible uncorrupted prefixes of size t. For $h \in \mathcal{H}_t$, define $\mu(h) = \sum_{i \in U(h)} D_i(\mathrm{Bad}_i(h))$. Also define $\mathcal{S}_h$ to be the event that the original offline draw $\tilde{X}$ realizes the values specified by h, and define $\mathcal{T}_h$ to be the event that all labels appearing in h survive corruption. That is, if $h = \{(i, x_i) : i \in P\}$, then

$$\mathcal{S}_h = \{\tilde{v}^i = x_i \text{ for all } i \in P\}, \qquad \mathcal{T}_h = \{P \subseteq \hat{Y}\}.$$

Let

$$\alpha_h := \Pr[\mathcal{T}_h \mid \mathcal{S}_h].$$

Let M_t be the benchmark distribution that first chooses a set $P \subseteq [n]$ of size t uniformly at random, and then independently draws $x_i \sim D_i$ for each $i \in P$. Thus, for $h = \{(i, x_i) : i \in P\}$, $\mathrm{M}_t(h) = \frac{1}{\binom{n}{t}} \prod_{i \in P} D_i(x_i)$. The event $\{\hat{X}_{\le t} = h,\ \zeta > t\}$ occurs when h is realized by $\tilde{X}$, the labels in P survive corruption, the first t labels in the random order are exactly P, and the next label is also uncorrupted. Therefore,

$$\Pr[\hat{X}_{\le t} = h,\ \zeta > t] = \mathrm{M}_t(h) \cdot \alpha_h \cdot \frac{r - t}{n - t}.$$

Also, $\Pr[\zeta > t] = \frac{(r)_{t+1}}{(n)_{t+1}}$, where $(a)_t = a(a-1)\cdots(a-t+1)$. Hence

$$\mathbb{E}[\mu(\hat{X}_{\le t}) \mid \zeta > t] = \sum_{h \in \mathcal{H}_t} \frac{\Pr[\hat{X}_{\le t} = h,\ \zeta > t]}{\Pr[\zeta > t]} \cdot \mu(h) = \frac{(n)_t}{(r)_t} \sum_{h \in \mathcal{H}_t} \mathrm{M}_t(h) \alpha_h \mu(h). \tag{20}$$

We first bound $\frac{(n)_t}{(r)_t}$. Since $t \le n/(4\ell)$ and $\ell \le n/4$, for every $0 \le j \le t-1$ we have $r - j = n - \ell - j \ge n/2$. Therefore,

$$\frac{(n)_t}{(r)_t} = \prod_{j=0}^{t-1} \left(1 + \frac{\ell}{r-j}\right) \le \left(1 + \frac{2\ell}{n}\right)^t \le \left(1 + \frac{2\ell}{n}\right)^{n/(4\ell)} \le \sqrt{e}.$$

Thus, it remains to show that $\alpha_h \mu(h) = O(\ell)$ for every prefix h.

Fix a prefix $h = \{(i, x_i) : i \in P\}$. If $\mu(h) \le 4\ell$, then $\alpha_h \mu(h) \le 4\ell$, so assume $\mu(h) > 4\ell$. We prove the stronger bound $\alpha_h \mu(h) = O(1)$ in this case.

Condition on $\mathcal{S}_h$. For each $i \in U(h)$, define $I_i = \mathbf{1}\{\tilde{v}^i \in \mathrm{Bad}_i(h)\}$, and let $K(h) = \sum_{i \in U(h)} I_i$. Under the conditioning on $\mathcal{S}_h$, the remaining samples are still independent and $\tilde{v}^i \sim D_i$ for every $i \in U(h)$. Hence $\mathbb{E}[K(h) \mid \mathcal{S}_h] = \mu(h)$. Since the indicators I_i are independent Bernoulli random variables, we also have

$$\mathrm{Var}(K(h) \mid \mathcal{S}_h) = \sum_{i \in U(h)} \mathrm{Var}(I_i \mid \mathcal{S}_h) \le \sum_{i \in U(h)} D_i(\mathrm{Bad}_i(h)) = \mu(h).$$

Let $\mathrm{Bad}(h) = \{(i, x) : i \in U(h),\ x \in \mathrm{Bad}_i(h)\}$. Using the definition of $\mathrm{Bad}_i(h)$,

$$\begin{aligned}\Pr[\hat{X}_{t+1} \in \mathrm{Bad}(h) \mid \hat{X}_{\le t} = h,\ \zeta > t] &= \sum_{i \in U(h)} \sum_{x \in \mathrm{Bad}_i(h)} \Pr[\hat{X}_{t+1} = (i, x) \mid \hat{X}_{\le t} = h,\ \zeta > t] \\ &\le \sum_{i \in U(h)} \sum_{x \in \mathrm{Bad}_i(h)} \frac{D_i(x)}{2(n-t)} = \frac{\mu(h)}{2(n-t)}. \end{aligned} \tag{21}$$

Now let $L(h)$ be the number of uncorrupted surviving labels outside the prefix whose uncorrupted values lie in their corresponding bad sets:

$$L(h) = \left| \left\{ i \in U(h) : (i, \tilde{v}^i) \in \hat{Y},\ \tilde{v}^i \in \mathrm{Bad}_i(h) \right\} \right|.$$

Conditioned on $\hat{X}_{\le t} = h$ and $\zeta > t$, the next label is uniform among the $r - t$ surviving uncorrupted labels outside the prefix. The distribution of $L(h)$ under this conditioning is the same as conditioning

on $\mathcal{S}_h$ and $\mathcal{T}_h$, since the random order is independent of the sample values and of the adversary's choices. Hence

$$\Pr[\hat{X}_{t+1} \in \mathrm{Bad}(h) \mid \hat{X}_{\le t} = h,\ \zeta > t] = \frac{\mathbb{E}[L(h) \mid \mathcal{S}_h, \mathcal{T}_h]}{r-t}.$$

Combining this with (21) gives $\mathbb{E}[L(h) \mid \mathcal{S}_h, \mathcal{T}_h] \le (r-t) \cdot \frac{\mu(h)}{2(n-t)} \le \frac{\mu(h)}{2}$. Every label counted by $K(h)$ either survives corruption and is counted by $L(h)$, or is one of the ℓ corrupted labels. Therefore $K(h) \le L(h) + \ell$ deterministically. Since $\mu(h) > 4\ell$, we get

$$\mathbb{E}[K(h) \mid \mathcal{S}_h, \mathcal{T}_h] \le \mathbb{E}[L(h) \mid \mathcal{S}_h, \mathcal{T}_h] + \ell \le \frac{\mu(h)}{2} + \ell \le \frac{3\mu(h)}{4}.$$

We now apply the following elementary inequality. Let K be a random variable with $\mathbb{E}[K] = \mu$ and $\mathrm{Var}(K) \le \mu$, and let T be an event with $\Pr[T] = \alpha$ and $\mathbb{E}[K \mid T] \le 3\mu/4$. Since $f(x) = (x-\mu)^2$ is convex,

$$\begin{aligned}
\mu \ge \mathrm{Var}(K) = \mathbb{E}[(K-\mu)^2] &\ge \alpha \cdot \mathbb{E}[(K-\mu)^2 \mid T] \\
&\ge \alpha \cdot (\mathbb{E}[K-\mu \mid T])^2 \quad \text{(Jensen's inequality)} \\
&= \alpha \cdot (\mu - \mathbb{E}[K \mid T])^2 \\
&\ge \alpha \cdot \left(\frac{\mu}{4}\right)^2.
\end{aligned}$$

Rearranging gives $\alpha\mu \le 16$. Apply this inequality in the probability space conditioned on $\mathcal{S}_h$, with $K = K(h)$, $T = \mathcal{T}_h$, $\alpha = \alpha_h$, and $\mu = \mu(h)$. We obtain $\alpha_h \mu(h) \le 16$. Therefore $\alpha_h \mu(h) = O(\ell)$ for every $h \in \mathcal{H}_t$. Substituting this into (20) and using $(n)_t/(r)_t \le \sqrt{e}$ proves the lemma. □

With the stochastic-distortion lemma in hand, we can now prove the second bound in Lemma 3.4 for the adversarial noise model.

Proof of the adversarial residual bound in Lemma 3.4. If $\ell > n/4$, then the desired bound follows immediately from

$$\Gamma_\zeta \le \Gamma_0 = \sum_{i=1}^{n} \kappa_{F_0^{\mathrm{RO}}}(v^i) \le \sum_{i=1}^{n} \mathrm{OPT}(\{v^i\}) \le n \cdot \mathrm{OPT}(X) \le 4\ell \cdot \mathrm{OPT}(X).$$

Thus, assume $\ell \le n/4$, and assume $T := n/(4\ell)$ is an integer for simplicity. This only affects constants. We first reduce $\mathbb{E}[\Gamma_\zeta]$ to an average over short uncorrupted prefixes. Recall that Γ_t is monotone in t, i.e., $\Gamma_t \le \Gamma_{t-1}$. We claim that

$$\mathbb{E}[\Gamma_\zeta] \le \ell \cdot \mathbb{E}_X[\mathrm{OPT}(X)] + \frac{C}{T} \sum_{t=0}^{T-1} \Pr[\zeta > t] \cdot \mathbb{E}[\Gamma_t \mid \zeta > t], \tag{22}$$

for an absolute constant $C > 0$. Indeed, we do a case analysis on where the first corrupted sample appears in the random order.

- If the first corrupted sample appears in position 1, then $\zeta = 0$. We have $\Pr[\zeta = 0] = \ell/n$, and $\Gamma_0 \le \sum_{i=1}^{n} \mathrm{OPT}(\{v^i\}) \le n \cdot \mathrm{OPT}(X)$. Hence, this case contributes at most $\ell \cdot \mathbb{E}_X[\mathrm{OPT}(X)]$.
- Suppose the first corrupted sample appears in position $q+1$, where $1 \le q < T$. Equivalently, $\zeta = q$. By monotonicity, $\Gamma_q \le \Gamma_{q-1}$. Thus, the contribution from these cases is at most $\sum_{q=1}^{T-1} \Pr[\zeta = q] \cdot \mathbb{E}[\Gamma_{q-1} \mid \zeta = q]$. Let $t = q-1$. The random variable Γ_t depends only on the online input X, the first t samples in the random order, and the internal randomness of RO. Hence, its distribution under the condition $\zeta = t+1$ is the same as under $\zeta > t$. Also,

  $$\Pr[\zeta = t+1 \mid \zeta > t] = \frac{\ell}{n-t-1} = O\left(\frac{\ell}{n}\right) = O\left(\frac{1}{T}\right),$$

  since $t < T \le n/4$. Therefore, the contribution from $1 \le \zeta < T$ is at most $O\left(\frac{1}{T}\right) \sum_{t=0}^{T-1} \Pr[\zeta > t] \cdot \mathbb{E}[\Gamma_t \mid \zeta > t]$.

- Finally, suppose the first corrupted sample appears after position T, i.e., $\zeta \geq T$. By monotonicity, for every $0 \leq t \leq T-1$ we have $\Gamma_\zeta \leq \Gamma_t$, and hence $\Gamma_\zeta \leq \frac{1}{T}\sum_{t=0}^{T-1}\Gamma_t$. Thus, the contribution from this case is at most $\frac{1}{T}\sum_{t=0}^{T-1}\Pr[\zeta \geq T]\cdot\mathbb{E}[\Gamma_t \mid \zeta \geq T]$. For each $t < T$, conditional on the online input X and the internal randomness of RO, the random variable Γ_t depends only on the first t samples in the random order. These samples have the same distribution under the conditions $\zeta \geq T$ and $\zeta > t$. Therefore, $\mathbb{E}[\Gamma_t \mid \zeta \geq T] = \mathbb{E}[\Gamma_t \mid \zeta > t]$. Since $\Pr[\zeta \geq T] \leq \Pr[\zeta > t]$, this contribution is at most

$$\frac{1}{T}\sum_{t=0}^{T-1}\Pr[\zeta > t]\cdot\mathbb{E}[\Gamma_t \mid \zeta > t].$$

Combining the three cases proves (22).

For every t, write $H = \hat{X}_{\leq t}$. We split Γ_t into processed and unprocessed labels:

$$\Gamma_t = \sum_{i\notin U(H)} \kappa_{F_t^{\mathrm{RO}}}(v^i) + \sum_{i\in U(H)} \kappa_{F_t^{\mathrm{RO}}}(v^i).$$

We bound these two contributions separately.

Claim B.9 (Unprocessed labels)**.** *We have*

$$\frac{C}{T}\sum_{t=0}^{T-1}\Pr[\zeta > t]\cdot\mathbb{E}\left[\sum_{i\in U(H)}\kappa_{F_t^{\mathrm{RO}}}(v^i) \mid \zeta > t\right] \leq O(\ell)\cdot\mathbb{E}[C_\zeta^{\mathrm{RO}}] + O(\ell)\cdot\mathbb{E}_X[\mathrm{OPT}(X)].$$

Proof. Fix t and condition on $H = h$, where $h = \{(i, x_i) : i \in P\}$ and $|P| = t$. We split the contribution of the labels in $U(h)$ according to whether the corresponding online draw lies in the bad set $\mathrm{Bad}_i(h)$:

$$\begin{aligned}\sum_{i\in U(h)}\mathbb{E}\left[\kappa_{F_t^{\mathrm{RO}}}(v^i) \mid H = h, \zeta > t\right] &= \sum_{i\in U(h)}\mathbb{E}\left[\kappa_{F_t^{\mathrm{RO}}}(v^i)\cdot\mathbf{1}\{v^i \notin \mathrm{Bad}_i(h)\} \mid H = h, \zeta > t\right] \\ &\quad + \sum_{i\in U(h)}\mathbb{E}\left[\kappa_{F_t^{\mathrm{RO}}}(v^i)\cdot\mathbf{1}\{v^i \in \mathrm{Bad}_i(h)\} \mid H = h, \zeta > t\right].\end{aligned}$$

For the non-bad part, the definition of $\mathrm{Bad}_i(h)$ implies that for every $i \in U(h)$ and every $x \notin \mathrm{Bad}_i(h)$,

$$D_i(x) \leq 2(n-t)\Pr[\hat{X}_{t+1} = (i, x) \mid H = h, \zeta > t].$$

Since the online input is independent of the offline random order, and F_t^{RO} is determined by the prefix h and the randomness of RO, we obtain

$$\begin{aligned}\sum_{i\in U(h)}\mathbb{E}\left[\kappa_{F_t^{\mathrm{RO}}}(v^i)\cdot\mathbf{1}\{v^i \notin \mathrm{Bad}_i(h)\} \mid H = h, \zeta > t\right] &= \sum_{i\in U(h)}\sum_{x\notin\mathrm{Bad}_i(h)} D_i(x)\cdot\mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(x) \mid H = h, \zeta > t] \\ &\leq 2n\cdot\mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(\hat{X}_{t+1}) \mid H = h, \zeta > t].\end{aligned}$$

Taking expectations over H conditioned on $\zeta > t$, multiplying by $\Pr[\zeta > t]$, summing over $0 \leq t \leq T-1$, and using $T = n/(4\ell)$, we get that the total non-bad contribution satisfies

$$\frac{C}{T}\sum_{t=0}^{T-1}\Pr[\zeta > t]\cdot\mathbb{E}\left[\sum_{i\in U(H)}\kappa_{F_t^{\mathrm{RO}}}(v^i)\cdot\mathbf{1}\{v^i \notin \mathrm{Bad}_i(H)\} \mid \zeta > t\right] \leq O(\ell)\sum_{t=0}^{T-1}\Pr[\zeta > t]\cdot\mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(\hat{X}_{t+1}) \mid \zeta > t].$$

Moreover, whenever $\zeta > t$, the client $\hat{X}_{t+1}$ belongs to the uncorrupted prefix $\hat{X}_{\leq\zeta}$, and the cost incurred by RO in round $t+1$ is at least $\kappa_{F_t^{\mathrm{RO}}}(\hat{X}_{t+1})$ by Observation B.4. Therefore,

$$\sum_{t=0}^{T-1}\Pr[\zeta > t]\cdot\mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(\hat{X}_{t+1}) \mid \zeta > t] \leq \mathbb{E}[C_\zeta^{\mathrm{RO}}]. \tag{23}$$

Combining (23) with the previous equation, the total non-bad contribution in Claim B.9 is $O(\ell) \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}}]$. For the bad part, $\kappa_{F_t^{\mathrm{RO}}}(v^i) \le \mathrm{OPT}(\{v^i\})$. Hence,

$$\sum_{i \in U(h)} \kappa_{F_t^{\mathrm{RO}}}(v^i) \cdot \mathbf{1}\{v^i \in \mathrm{Bad}_i(h)\} \le \sum_{i \in U(h)} \mathrm{OPT}(\{v^i\}) \cdot \mathbf{1}\{v^i \in \mathrm{Bad}_i(h)\}.$$

Conditioned on $H = h$ and $\zeta > t$, the online input remains independent of the offline sample and still has distribution $v^i \sim D_i$. Moreover, once h is fixed, the sets $\mathrm{Bad}_i(h)$ are deterministic. Therefore, we can apply Lemma B.3 with $\tilde{D}_i(x) = D_i(x) \cdot \mathbf{1}\{x \in \mathrm{Bad}_i(h)\}$ for $i \in U(h)$, and $\tilde{D}_i = \mathbf{0}$ otherwise. Thus,

$$\sum_{i \in U(h)} \mathbb{E}\left[\mathrm{OPT}(\{v^i\}) \cdot \mathbf{1}\{v^i \in \mathrm{Bad}_i(h)\} \mid H = h, \zeta > t\right] \le O\left(1 + \sum_{i \in U(h)} D_i(\mathrm{Bad}_i(h))\right) \cdot \mathbb{E}_X[\mathrm{OPT}(X)]. \tag{24}$$

Taking expectation over $H = \hat{X}_{\le t}$ conditioned on $\zeta > t$ and using Lemma B.8, this is at most $O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)]$. Therefore, the contribution to Claim B.9 from the bad parts is

$$\frac{C}{T} \sum_{t=0}^{T-1} \Pr[\zeta > t] \cdot \mathbb{E}\left[\sum_{i \in U(H)} \kappa_{F_t^{\mathrm{RO}}}(v^i) \cdot \mathbf{1}\{v^i \in \mathrm{Bad}_i(H)\} \mid \zeta > t\right] \le O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

Combining the contributions from the non-bad and bad parts proves the claim. □

Claim B.10 (Processed labels)**.** *We have*

$$\frac{C}{T} \sum_{t=0}^{T-1} \Pr[\zeta > t] \cdot \mathbb{E}\left[\sum_{i \notin U(H)} \kappa_{F_t^{\mathrm{RO}}}(v^i) \mid \zeta > t\right] \le O(\mathbb{E}[C_\zeta^{\mathrm{RO}}]) + O(\mathbb{E}_X[\mathrm{OPT}(X)]).$$

Proof. Let $\sigma(i)$ denote the position of the labeled sample $\hat{v}^i$ in the random order of $\hat{X}$. Define

$$\mathcal{C}_{\mathrm{proc}} := \frac{C}{T} \sum_{t=0}^{T-1} \Pr[\zeta > t] \cdot \mathbb{E}\left[\sum_{i \notin U(H)} \kappa_{F_t^{\mathrm{RO}}}(v^i) \mid \zeta > t\right],$$

where $H = \hat{X}_{\le t}$. If $i \notin U(H)$, then $\sigma(i) \le t$. Since F_t^{RO} contains $F_{\sigma(i)-1}^{\mathrm{RO}}$, monotonicity gives $\kappa_{F_t^{\mathrm{RO}}}(v^i) \le \kappa_{F_{\sigma(i)-1}^{\mathrm{RO}}}(v^i)$. Therefore,

$$\mathcal{C}_{\mathrm{proc}} \le \frac{C}{T} \sum_{t=0}^{T-1} \Pr[\zeta > t] \cdot \sum_{r=1}^{t} \mathbb{E}[\kappa_{F_{r-1}^{\mathrm{RO}}}(v^{\sigma^{-1}(r)}) \mid \zeta > t].$$

For $r \le t$, the random variable $\kappa_{F_{r-1}^{\mathrm{RO}}}(v^{\sigma^{-1}(r)})$ depends only on the online input, the first r positions of the random order, and the internal randomness of RO. The conditional distribution of the first r positions is the same under $\zeta > t$ and $\zeta > r-1$, as the random order is independent of the adversary's choices. Hence,

$$\mathbb{E}[\kappa_{F_{r-1}^{\mathrm{RO}}}(v^{\sigma^{-1}(r)}) \mid \zeta > t] = \mathbb{E}[\kappa_{F_{r-1}^{\mathrm{RO}}}(v^{\sigma^{-1}(r)}) \mid \zeta > r-1].$$

Using the above equality and changing the order of summation gives

$$\begin{aligned}
\mathcal{C}_{\mathrm{proc}} &\le \frac{C}{T} \sum_{r=1}^{T-1} \left(\sum_{t=r}^{T-1} \Pr[\zeta > t]\right) \mathbb{E}[\kappa_{F_{r-1}^{\mathrm{RO}}}(v^{\sigma^{-1}(r)}) \mid \zeta > r-1] \\
&\le C \sum_{r=1}^{T-1} \Pr[\zeta > r-1] \cdot \mathbb{E}[\kappa_{F_{r-1}^{\mathrm{RO}}}(v^{\sigma^{-1}(r)}) \mid \zeta > r-1] \\
&= C \sum_{t=0}^{T-2} \Pr[\zeta > t] \cdot \mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(v^{\sigma^{-1}(t+1)}) \mid \zeta > t].
\end{aligned} \tag{25}$$

We now bound the expectation in (25). Fix t and condition on $H = h$ and $\zeta > t$. Then the sample in position $t+1$ is uncorrupted, so $\sigma^{-1}(t+1)$ is the label of the next uncorrupted sample after the prefix h. For each fixed realization of the surviving uncorrupted labels, any label $i \in U(h)$ is selected as $\sigma^{-1}(t+1)$ with probability $1/(n-\ell-t)$ if it survives corruption, and with probability 0 otherwise, since there are exactly $n-\ell-t$ surviving uncorrupted labels outside the prefix h. We then split according to the bad sets $\mathrm{Bad}_{\sigma^{-1}(t+1)}(h)$.

For the non-bad part,

$$\mathbb{E}\left[\kappa_{F_t^{\mathrm{RO}}}(v^{\sigma^{-1}(t+1)}) \cdot \mathbf{1}\{v^{\sigma^{-1}(t+1)} \notin \mathrm{Bad}_{\sigma^{-1}(t+1)}(h)\} \mid H = h, \zeta > t\right] \leq \frac{1}{n-\ell-t} \sum_{i \in U(h)} \sum_{x \notin \mathrm{Bad}_i(h)} D_i(x) \cdot \mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(x) \mid H = h, \zeta > t].$$

Using the definition of $\mathrm{Bad}_i(h)$, $D_i(x) \leq 2(n-t) \cdot \Pr[\hat{X}_{t+1} = (i,x) \mid H = h, \zeta > t]$ for $i \in U(h)$ and $x \notin \mathrm{Bad}_i(h)$. Thus, this is at most

$$\frac{2(n-t)}{n-\ell-t} \cdot \mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(\hat{X}_{t+1}) \mid H = h, \zeta > t].$$

Since $t < T = n/(4\ell)$ and $\ell \leq n/4$, we have $n-\ell-t = \Omega(n)$ and therefore $\frac{2(n-t)}{n-\ell-t} = O(1)$. Taking expectations over H conditioned on $\zeta > t$, the contribution to (25) from the non-bad part is $O(1)\sum_{t=0}^{T-2} \Pr[\zeta > t] \cdot \mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(\hat{X}_{t+1}) \mid \zeta > t]$. For the bad part, using the same $1/(n-\ell-t)$ bound on the probability of each label in $U(h)$,

$$\mathbb{E}\left[\kappa_{F_t^{\mathrm{RO}}}(v^{\sigma^{-1}(t+1)}) \cdot \mathbf{1}\{v^{\sigma^{-1}(t+1)} \in \mathrm{Bad}_{\sigma^{-1}(t+1)}(h)\} \mid H = h, \zeta > t\right] \leq \frac{1}{n-\ell-t} \sum_{i \in U(h)} \mathbb{E}\left[\mathrm{OPT}(\{v^i\}) \cdot \mathbf{1}\{v^i \in \mathrm{Bad}_i(h)\} \mid H = h, \zeta > t\right].$$

Taking expectations over H conditioned on $\zeta > t$, using (24), Lemma B.8, and $n-\ell-t = \Omega(n)$, this is at most $O\left(\frac{\ell}{n}\right) \cdot \mathbb{E}_X[\mathrm{OPT}(X)]$. Substituting the expected contribution from the non-bad and bad parts into (25),

$$\mathcal{C}_{\mathrm{proc}} \leq O(1) \sum_{t=0}^{T-2} \Pr[\zeta > t] \cdot \mathbb{E}[\kappa_{F_t^{\mathrm{RO}}}(\hat{X}_{t+1}) \mid \zeta > t] + O\left(\frac{\ell}{n}\right) \sum_{t=0}^{T-2} \Pr[\zeta > t] \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

The first term is at most $O(\mathbb{E}[C_\zeta^{\mathrm{RO}}])$ by (23). The second term is at most $O(\mathbb{E}_X[\mathrm{OPT}(X)])$, since there are at most $T = n/(4\ell)$ terms in the sum and $\Pr[\zeta > t] \leq 1$. This proves the claim. □

Substituting Claims B.9 and B.10 into (22), we obtain

$$\mathbb{E}[\Gamma_\zeta] \leq O(\ell) \cdot \mathbb{E}[C_\zeta^{\mathrm{RO}}] + O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$

By (6), we have $\mathbb{E}[C_\zeta^{\mathrm{RO}}] \leq O(1) \cdot \mathbb{E}_X[\mathrm{OPT}(X)]$. Hence,

$$\mathbb{E}[\Gamma_\zeta] \leq O(\ell) \cdot \mathbb{E}_X[\mathrm{OPT}(X)].$$ □

Finally, we show that the dependence on ℓ in Theorem 3.1 is optimal up to constant factors.

Theorem B.11 (Lower bound)**.** *In either noise model, any algorithm for online facility location with samples has competitive ratio*

$$\Omega\left(\frac{\log \ell}{\log \log \ell}\right).$$

Proof. We adapt the hard instance of Fotakis (2008), which shows that no randomized algorithm for online facility location can achieve a competitive ratio better than

$$\Omega\left(\frac{\log n}{\log \log n}\right)$$

against an oblivious adversary, even with uniform facility costs. Let $(C_{\mathrm{hard}}, \mathcal{F}_{\mathrm{hard}})$ denote the underlying collection of clients and facilities in this instance, and let X_{hard} be the sequence of clients that arrives online. We choose $|X_{\mathrm{hard}}| = \Theta(\ell)$.

Our clients and facilities consist of $(C_{\mathrm{hard}}, \mathcal{F}_{\mathrm{hard}})$, together with a single client v and a facility f_v co-located with v. We place v sufficiently far from every client in C_{hard}. The observed offline sample $\hat{X}$ consists only of copies of v, while the online input X consists of copies of v together with X_{hard}.

For the distributional noise model, we choose the true and predicted distributions so that the coordinates corresponding to X_{hard} are point masses on the hard clients under D_i and point masses on v under $\hat{D}_i$. Thus, choosing $|X_{\mathrm{hard}}| = \ell/2$ gives

$$\sum_i \|D_i - \hat{D}_i\|_1 = \ell.$$

For the adversarial noise model, the corresponding entries of an initially clean sample are replaced by copies of v, so that $|X_{\mathrm{hard}}| = \ell$ entries are corrupted.

In either case, the observed sample provides no information about X_{hard}. Even if the algorithm serves all copies of v using f_v, it must incur expected cost

$$\Omega\left(\frac{\log |X_{\mathrm{hard}}|}{\log\log |X_{\mathrm{hard}}|} \cdot \mathrm{OPT}(X_{\mathrm{hard}})\right)$$

on the hard portion of the online sequence. Since $|X_{\mathrm{hard}}| = \Theta(\ell)$, this gives the desired

$$\Omega\left(\frac{\log \ell}{\log\log \ell}\right)$$

lower bound. □

C Noise Sensitivity of Gupta et al. (2024)

We also note that the sample-based algorithm of Gupta et al. (2024) is not robust to noise.

Lemma C.1. *The sample-based online facility location algorithm of Gupta et al. (2024), which runs a random-order algorithm on the offline sample and then serves the online clients greedily, can have competitive ratio $\Omega(\ell)$ under either noise model.*

Proof. Recall that Gupta et al. (2024) consider the uniform case, where all facilities have the same opening cost. Normalize all facility opening costs to 1. Let the online input consist of n copies of a single point x, with a facility colocated with x. Hence $\mathrm{OPT}(X) = 1$.

For the adversarial noise model, corrupt ℓ entries of the offline sample and replace them by distinct points $z_1, \ldots, z_\ell$ that are also available as candidate facilities. Place the points so that every pair of distinct points in $\{x, z_1, \ldots, z_\ell\}$ is at distance at least 2.

Recall that the 3-competitive random-order algorithm of Kaplan et al. (2023) opens a facility at an arriving client with probability $\min\{d/2, 1\}$, where d is the distance to the currently open facilities. Therefore, whenever a point z_i is processed, every previously opened facility is at distance at least 2 from z_i, and so a facility opens at z_i with probability one (if it is not already open). Thus, the random-order run opens all ℓ such facilities. The algorithm of Gupta et al. (2024) adds these facilities to its solution, and therefore incurs cost at least ℓ, while $\mathrm{OPT}(X) = 1$. Its competitive ratio is therefore $\Omega(\ell)$.

For the distributional noise model, we use the same construction with $\ell/2$ distinct points $z_1, \ldots, z_{\ell/2}$. For $1 \le i \le \ell/2$, let the true distribution D_i be a point mass at x and the predicted distribution $\hat{D}_i$ be a point mass at the corresponding z_i. Let the remaining true and predicted distributions be point masses at x. The offline sample therefore contains z_i for $1 \le i \le \ell/2$, and the same argument shows that the random-order run opens a facility at each of them. Hence the algorithm incurs cost at least $\ell/2 = \Omega(\ell) \cdot \mathrm{OPT}(X)$. □