Online Pricing with Polluted Offline Data

Yue Wang[1], Zeyu Zheng[2], and Zuo-Jun Max Shen[3]

[1]Department of Computational Medicine, University of California, Los Angeles, yuew@g.ucla.edu
[2]Department of Industrial Engineering and Operations Research, University of California, Berkeley, zyzheng@berkeley.edu
[3]Department of Industrial Engineering and Operations Research, University of California, Berkeley, shen@ieor.berkeley.edu

Abstract

We consider the problem of online pricing with offline data, where the decision maker has in hand pre-existed offline data and then takes online sequential actions to maximize expected cumulative revenue as well as learning the optimal pricing. We focus on a simple and specific problem setting where the distribution of the offline data is "polluted", meaning that their distribution can be different from the online data to be obtained. In this work, we provide explicit answers to two questions. First, if the decision maker does not know that the offline data are polluted, and applies a policy that should be optimal if the offline data are not polluted, how serious is the effect of pollution on the online learning performance? We find the critical pollution level, above which the "optimal policy" with polluted data behaves worse than another policy that does not utilize offline data at all. Second, if the decision maker knows the magnitude of how the offline data are polluted, how to utilize such polluted offline data to better enhance online learning performances? We find a better policy by assigning a smaller weight to each offline data point.

1 INTRODUCTION

In online learning problems, for each period, the decision maker chooses an action, and receives a revenue, which is determined by some unknown system parameters. The aim is to minimize the regret, i.e., the difference between the maximal possible expected cumulative revenue and the actual expected cumulative revenue achieved by the decision maker. The decision maker may need to learn the possible values of such unknown system parameters, and adjust the action-choosing policy dynamically.

In some scenarios, the decision maker has already observed some offline data before facing the online learning problem. Such offline data are supposed to be generated by the same system, therefore containing useful distributional information about the system parameters. With the knowledge from offline data, the decision maker could apply wiser policies, and receive better revenues, compared to scenarios where offline data are not available at all before the online learning problem.

The precise worth of offline data comes from the assumption that the offline data could faithfully reflect the online system properties. In reality, during the time between the offline stage and the online stage, the system might have changed. Then the offline data possess a risk to reflect only outdated system properties. The offline data could also be intentionally distorted or unintentionally contaminated. When the offline data cannot correctly reflect the online system properties, we call such offline data "polluted". The polluted offline data might mislead the decision maker to obtain undesirable regrets for the online stage.

If the pollution is not restricted, e.g., if all offline data are replaced by random numbers independent with the system, the offline data would be totally useless. In this paper, we consider a less extreme and potentially more relevant problem: when the pollution satisfies a certain realistic format, and the pollution level is limited to some degree, what is the precise influence of this pollution on online pricing, and what countermeasures can be adopted?

In this work, we first consider situations when the decision maker does not know the offline data are polluted, and applies a policy that should be optimal if the offline data are not polluted. We show that even if the pollution level is relatively low, the performance of this "optimal" policy might be worse than another policy that does not consider offline data at all. This means the polluted offline data have negative effects for online pricing. In this case, we will search for the critical pollution level, at which the polluted offline data shift from helpful to harmful. We provide explicit quantification for this critical pollution level.

We then consider situations where the decision maker knows the offline data are polluted, but does not know how the offline data are polluted. We show that if the decision maker knows the pollution level is within a limited range, it is possible to distill some useful information from the polluted offline data and achieve better revenues.

In Section 2, we introduce a simplified setup of the online pricing with offline data problem. For policies that utilize offline data, we find the optimal policy π^*. For policies that do not utilize offline data, we find the optimal policy $\tilde{\pi}^*$. We also introduce the format of offline data pollution.

In Section 3, we consider the scenario that the decision maker does not know the existence of offline data pollution, and still applies the "optimal" policy π^*. We study the effect of pollution on the overall regret of π^*, especially when the overall regret of π^* is larger than that of $\tilde{\pi}^*$.

In Section 4, we consider the scenario that the decision maker knows the offline data are polluted within a given range. According to the pollution's properties, we design a policy π_0 that outperforms π^* and $\tilde{\pi}^*$.

We finish with some discussions in Section 5.

1.1 Related Literature

The topic of online pricing, or sometimes referred to as dynamic pricing, has been a keen focus in the online learning literature; see Araman and Caldentey [2009], Besbes and Zeevi [2009], Broder and Rusmevichientong [2012], den Boer and Zwart [2014], den Boer [2014], Keskin and Zeevi [2014], Cheung et al. [2017], den Boer and Keskin [2022], Bu et al. [2020], among others. We refer to den Boer [2015] for a comprehensive review. The online pricing problem can be viewed as a continuous parameter version of the multi-armed bandit problem in the online learning domain. Our work adopts the online pricing framework to illustrate the effect of polluted offline data on the online learning task and regret performances.

When offline data are available for the online learning task, [Bu et al., 2020] discuss a phase transition phenomenon to quantify how useful the offline data are for an online pricing problem. Li et al. [2021] and Wang and Zheng [2021] also discuss the settings where offline data are presented for an online learning task. This line of literature differs from our work in the sense that they do not discuss cases where the offline data may have a different distribution from the online data.

Another related line of work concerns dynamic pricing or online pricing problems when there are non-stationarities or distribution shifts during the online phase; see Besbes and Sauré [2014], Chen et al. [2020], Liu et al. [2022]. Their focus is to develop a robust online learning policy that can handle the distribution shift in the online phase. Our work considers a different setting where the distribution shift happens between the offline phase and the online phase, and focuses on understanding how exactly the distribution shift from the offline to online phases affects the online learning performances.

2 SETUP

In this section, we introduce a simplified setting of online pricing with offline data problem, so that the optimal policy and related quantities can be explicitly calculated.

In the analysis of this work, we will use two calculation tricks iteratively without describing them.

1. If X is Gaussian $\mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}(X^2) = \mu^2 + \sigma^2$.
2. For $S = \sum_{i=c}^{d} 1/i$, we have for $c > 1$ and $d > c - 1$,

$$\int_c^{d+1} \frac{1}{x}\,\mathrm{d}x < S < \int_{c-1}^{d} \frac{1}{x}\,\mathrm{d}x,$$

and thus

$$\log\frac{d+1}{c} < S < \log\frac{d}{c-1}.$$

For $U = \sum_{i=c}^{d} 1/i^2$, we have for $c > 1$ and $d > c - 1$,

$$\int_{c}^{d+1} 1/x^2 \ \mathrm{d}x < U < \int_{c-1}^{d} 1/x^2 \ \mathrm{d}x,$$

and thus

$$\frac{1}{c} - \frac{1}{d+1} < U < \frac{1}{c-1} - \frac{1}{d}.$$

2.1 Basic Setting

We first introduce a simple and naive online pricing setting to facilitate the closed-form analysis of the theory. The online stage has T periods. We use t to denote the number of the current period. For an online time period $t = i$ where $i = 1, 2, \cdots, T$, the decision maker chooses a price $p_i \in [p_{\min}, p_{\max}]$ with

$$0 < p_{\min} < p_{\max} < \infty,$$

and observes the demand

$$D_i = a - bp_i + \epsilon_i,$$

where ϵ_i is a mean-zero random noise.

We presume that the decision maker knows the value of $b > 0$, which represents the price elasticity, but the decision maker does not know the value of $a \in [a_{\min}, a_{\max}]$, where

$$0 < a_{\min} < a_{\max} < \infty.$$

The parameter a can be viewed as an intercept term that represents the baseline demand. We presume that the interval $[a_{\min}, a_{\max}]$ is wide enough, so that a Gaussian distribution restricted on this interval can be regarded as a general Gaussian distribution. Different noise terms ϵ_i are presumed to be independent Gaussian $\mathcal{N}(0, \sigma^2)$, and the decision maker knows this distribution.

Before the online stage, the decision maker observes $g(T)$ groups of offline data, denoted as

$$\hat{H} = \{(\hat{p}_1, \hat{D}_1), \ldots, (\hat{p}_{g(T)}, \hat{D}_{g(T)})\},$$

Here $\{\hat{p}_1, \ldots, \hat{p}_{g(T)}\}$ were chosen from a probability distribution on $[p_{\min}, p_{\max}]^{g(T)}$, where the distribution is known to the decision maker. Each $\hat{D}_i$ was determined by

$$\hat{D}_i = a - b\hat{p}_i + \hat{\epsilon}_i.$$

Here noise terms $\hat{\epsilon}_i$ are also presumed to be independent Gaussian distributed $\mathcal{N}(0, \sigma^2)$. Denote the set of all possible offline data as $\mathcal{H}$.

A policy π is a function (possibly stochastic) used by the decision maker to determine each online price p_i for $i = 1, 2, \cdots, T$, based on both offline data and online data that are known till the beginning of the i-th time period. That is,

$$p_i = \pi(\hat{H}, p_1, D_1, \ldots, p_{i-1}, D_{i-1}).$$

For the online period $t = i$, where $i = 1, 2, \cdots, T$, the random revenue is $p_i D_i$, whose expectation is $p_i(a - bp_i)$. For any given $a > 0$, the optimal price that maximizes the revenue is $p^* = a/(2b)$, and the maximal expected revenue is $a^2/(4b)$. For parameter a, offline data $\hat{H}$, and a policy corresponding to the offline data, $\pi \mid \hat{H}$, the overall expected regret is defined as the difference between maximal revenue and actual revenue,

$$R^a_{\pi|\hat{H}} = \sum_{i=1}^{T} \mathbb{E}[a^2/(4b) - p_i(a - bp_i)],$$

where the expectation is with respect to the potential randomness within the policy when deciding each of p_i. When designing a policy, the decision maker knows the total number of online periods, T.

To evaluate the overall performance of a policy π, we adopt the Bayesian approach of measurement. Assume the decision maker knows the prior distribution of a, $f(a)$, is uniform on $[a_{\min}, a_{\max}]$. For fixed a, the conditional distribution of offline data $\hat{H} \in \mathcal{H}$, $f(\hat{H} \mid a)$, is known. The overall regret of π is averaged over a and $\hat{H}$:

$$\bar{R}_\pi = \int_{a_{\min}}^{a_{\max}} f(a) \int_{\mathcal{H}} f(\hat{H} \mid a) R^a_{\pi|\hat{H}} \, \mathrm{d}\hat{H} \, \mathrm{d}a.$$

The optimal policy π^* is the policy that minimizes the overall regret:

$$\pi^* = \arg\min_\pi \{\bar{R}_\pi\}.$$

2.2 Optimal Policy

At each online period $t = i$, the decision maker sets p_i and observes

$$D_i = a - bp_i + \epsilon_i.$$

Since b is known, the decision maker knows the value of $a + \epsilon_i$. This quantity is independent of p_i, thus the amount of information about a acquired in each period is the same. The "exploration vs. exploitation" dilemma does not hold in this case, since any deliberate "exploration" does not provide extra information. The optimal policy would be pure "exploitation", which means setting the optimal price under the current knowledge. This setting is also discussed in Qiang and Bayati [2016] and den Boer [2015].

For the policies that consider offline data, we can determine the optimal one π^* and calculate the order of its overall regret.

Lemma 1. *The overall expected regret of π^* has order*

$$\Theta\left(1 \vee \log\left[\frac{T}{g(T)}\right]\right).$$

Proof. At the online period $t = i+1$, the decision maker will have observed $g(T)+i$ pairs of samples of (p, D), and we denote their sample mean by $(\bar{p}, \bar{D})$. The posterior distribution of a, $f_i(a)$, is then Gaussian

$$\mathcal{N}(\bar{D} + b\bar{p}, \sigma_i^2).$$

Here σ_i^2 depends on the concrete data. The optimal price p_{i+1}^* should minimize the expected regret

$$\begin{aligned} &\int_{a_{\min}}^{a_{\max}} f_i(a)[a^2/(4b) - p_{i+1}(a - bp_{i+1})]\mathrm{d}a \\ =&\frac{1}{4b}\int_{a_{\min}}^{a_{\max}} f_i(a)(a - 2bp_{i+1})^2\mathrm{d}a. \end{aligned}$$

The last integration is the second moment of a Gaussian random variable

$$\mathcal{N}(\bar{D} + b\bar{p} - 2bp_{i+1}, \sigma_i^2).$$

To minimize this term, the optimal policy π^* should set

$$p_{i+1}^* = \bar{D}/(2b) + \bar{p}/2.$$

For fixed a, since

$$p_{i+1}^* = \frac{1}{g(T)+i}\left[\sum_{j=1}^{g(T)}(\frac{a}{2b} + \frac{\hat{\epsilon}_j}{2b}) + \sum_{j=1}^{i}(\frac{a}{2b} + \frac{\epsilon_j}{2b})\right],$$

p_{i+1}^* is distributed as Gaussian

$$\mathcal{N}\left[\frac{a}{2b}, \frac{\sigma^2}{4b^2}\frac{1}{g(T)+i}\right].$$

Thus the expected regret of period $i+1$ for fixed a is

$$\bar{R}_{\pi^*}^{i+1} = \mathbb{E}\ b(p_{i+1} - \frac{a}{2b})^2 = \frac{\sigma^2}{4b[g(T)+i]},$$

which is independent with a. Therefore, the integration with $f_i(a)$ can be omitted. Summing over i, the overall expected regret

$$\bar{R}_{\pi^*} = \sum_{i=0}^{T-1} \bar{R}_{\pi^*}^{i+1} = \frac{\sigma^2}{4b^2}\sum_{i=0}^{T-1}\frac{1}{g(T)+i}$$

is between

$$\frac{\sigma^2}{4b} \log \frac{g(T)+T+1}{g(T)+1} \quad \text{and} \quad \frac{\sigma^2}{4b} \log \frac{g(T)+T}{g(T)}.$$

Therefore, the order of expected regret is

$$\Theta\left(1 \vee \log\left[\frac{T}{g(T)}\right]\right).$$

Specifically, when $g(T) = \Theta(T)$, the regret is at the order of $\Theta(1)$. □

For the policies that do not consider offline data, we can also find the optimal one $\tilde{\pi}^*$ and calculate the order of its overall regret.

Lemma 2. *The overall expected regret of $\tilde{\pi}^*$ has order*

$$\Theta(\log T).$$

Proof. Similar to the proof of Lemma 1, at online period $t = i+1$, the optimal policy should set

$$p^*_{i+1} = \frac{\bar{D}}{2b} + \frac{\bar{p}}{2},$$

where $\bar{p}$ and $\bar{D}$ are the mean of previous i groups of online prices and demands. The expected regret of period $i+1$ is

$$\bar{R}^{i+1}_{\tilde{\pi}^*} = \frac{\sigma^2}{4bi}.$$

At online period $t = 1$, since there is no available data, due to symmetry and the concavity of the revenue function, the optimal policy $\tilde{\pi}^*$ should set

$$p^*_1 = \frac{a_{\min} + a_{\max}}{4b}.$$

The expected regret of period 1 is

$$\bar{R}^{1}_{\tilde{\pi}^*} = \frac{(a_{\max} - a_{\min})^2}{48b},$$

denoted as C_0. The overall expected regret is

$$\bar{R}_{\tilde{\pi}^*} = \sum_{i=0}^{T-1} \bar{R}^{i+1}_{\tilde{\pi}^*} = C_0 + \frac{\sigma^2}{4b} \sum_{i=2}^{T} \frac{1}{i},$$

which is between

$$C_0 + \frac{\sigma^2}{4b} \log \frac{T}{2}$$

and

$$C_0 + \frac{\sigma^2}{4b}\log T.$$

The regret is then noted as

$$\frac{\sigma^2}{4b}\log T + \mathcal{O}(1),$$

and the order of $\bar{R}_{\tilde{\pi}^*}$ is $\Theta(\log T)$. □

2.3 Offline Data Pollution

Regarding data pollution, one can add large pollution terms to a small proportion of offline data points, or add small pollution terms to a large proportion of offline data points. Since the optimal policy only utilizes the average of offline prices and demands, different formats of pollution are equivalent, as long as they have the same total level of pollution. We describe this level formally as follows.

For simplicity, we assume each offline demand $\hat{D}_i$ is added an extra independent Gaussian noise

$$\mathcal{N}\left[\frac{h(T)}{g(T)}, \hat{\sigma}^2 - \sigma^2\right],$$

where

$$h(T) \geq 0,$$
$$\hat{\sigma}^2 \geq \sigma^2.$$

Since this noise should be bounded, we stipulate that

$$h(T) = \mathcal{O}[g(T)].$$

Equivalently, we can assume the offline data generating system has

$$a' = a + \frac{h(T)}{g(T)}$$

and

$$\hat{\epsilon}_i \sim \mathcal{N}(0, \hat{\sigma}^2).$$

Here $h(T)$ is the total pollution level, and $h(T)/g(T)$ is the average pollution level. Now each

$$\frac{\hat{D}_i}{2b} + \frac{\hat{p}_i}{2}$$

is not

$$\mathcal{N}\left(\frac{a}{2b}, \frac{\sigma^2}{4b^2}\right),$$

but

$$\mathcal{N}\left[\frac{a+h(T)/g(T)}{2b}, \frac{\hat{\sigma}^2}{4b^2}\right].$$

If there is no pollution, with the help of offline data, the overall regret of π^* is smaller than that of $\tilde{\pi}^*$. With offline data pollution, if the decision maker still applies π^*, then the overall regret of π^* might be larger than that of $\tilde{\pi}^*$. This means that the polluted data are "worse than nothing". We will compare the overall regrets of π^* and $\tilde{\pi}^*$ to determine the significance of offline data pollution. In the following, we assume T is sufficiently large, and vary the orders of $g(T)$ and $h(T)$.

3 CASES WHEN DECISION MAKER DOES NOT KNOW THE EXISTENCE OF DATA POLLUTION

Assume the offline demand data $\hat{D}_i$ are polluted by an extra Gaussian noise

$$\mathcal{N}\left[\frac{h(T)}{g(T)}, \hat{\sigma}^2 - \sigma^2\right].$$

If the decision maker does not know that the pollution exists, and still applies π^* that sets

$$p^*_{i+1} = \frac{\bar{D}}{2b} + \frac{\bar{p}}{2},$$

then p^*_{i+1} is Gaussian:

$$\mathcal{N}\left\{\frac{a}{2b} + \frac{h(T)}{2b[g(T)+i]}, \frac{1}{4b^2}\frac{i\sigma^2 + g(T)\hat{\sigma}^2}{[g(T)+i]^2}\right\},$$

and the expected regret is

$$\bar{R}^{i+1}_{\pi^*} = \mathbb{E}\ b(p_{i+1} - \frac{a}{2b})^2 = \frac{1}{4b}\frac{h^2(T) + i\sigma^2 + g(T)\hat{\sigma}^2}{[g(T)+i]^2}.$$

Then the overall regret $\bar{R}_{\pi^*}$ is between

$$\frac{\sigma^2}{4b}\log\frac{g(T)+T+1}{g(T)+1} + \frac{1}{4b}\frac{h^2(T)T + g(T)T(\hat{\sigma}^2 - \sigma^2)}{[g(T)+1][g(T)+T+1]}$$

and

$$\frac{\sigma^2}{4b}\log\frac{g(T)+T}{g(T)} + \frac{1}{4b}\frac{h^2(T)T + g(T)T(\hat{\sigma}^2 - \sigma^2)}{g(T)[g(T)+T]}.$$

In the following, we compare $\bar{R}_{\pi^*}$ and

$$\bar{R}_{\tilde{\pi}^*} = \frac{\sigma^2}{4b}\log T + \mathcal{O}(1),$$

which is not affected by the offline data pollution, and find the critical average pollution level $h(T)/g(T)$, at which $\bar{R}_{\pi^*}$ is at the same order as $\bar{R}_{\tilde{\pi}^*}$. If $\bar{R}_{\pi^*}$ is larger, then the pollution would make the harm of the offline data dominate the benefit of the offline data, in terms of its support on the online pricing part. In this sense, the use of offline data would be worse than completely ignoring the offline data.

Proposition 1. *When $g(T) \leq \mathcal{O}(T)$, for any small positive number δ, the pollution level*

$$\frac{h(T)}{g(T)} = \Theta\left[\sqrt{\frac{(\log T)^{1+\delta}}{g(T)}}\right],$$

is enough for $\bar{R}_{\pi^*} > \bar{R}_{\tilde{\pi}^*}$ in the sense of order, meaning that $\bar{R}_{\tilde{\pi}^*}/\bar{R}_{\pi^*} = o(1)$.

Proof. When $g(T) \leq \mathcal{O}(T)$,

$$\bar{R}_{\pi^*} = \frac{\sigma^2}{4b}\log\frac{T}{g(T)} + \frac{1}{4b}\frac{h^2(T)/g^2(T)}{[1/T + 1/g(T)]} + \mathcal{O}(1).$$

The critical average pollution level for $\bar{R}_{\pi^*} = \bar{R}_{\tilde{\pi}^*}$ is

$$\frac{h(T)}{g(T)} = \sqrt{\sigma^2\left[\frac{1}{T} + \frac{1}{g(T)}\right]\log[g(T)]} + o(1),$$

which is

$$\Theta\left\{\sqrt{\frac{\log[g(T)]}{g(T)}}\right\}.$$

For $\bar{R}_{\pi^*} > \bar{R}_{\tilde{\pi}^*}$ in the sense of order, we just need

$$\frac{h(T)}{g(T)} > \mathcal{O}\left[\sqrt{\frac{\log T}{g(T)}}\right],$$

such as

$$\frac{h(T)}{g(T)} = \Theta\left[\sqrt{\frac{(\log T)^{1+\delta}}{g(T)}}\right],$$

where δ is any small positive number. □

When $g(T) \leq \mathcal{O}(\log T)$, due to the restriction $h(T) = \mathcal{O}[g(T)]$, we cannot have

$$\frac{h(T)}{g(T)} > \mathcal{O}\left[\sqrt{\frac{\log T}{g(T)}}\right],$$

meaning that there are not enough offline data to affect the online pricing in the sense of order.

Proposition 2. *When $g(T) > \mathcal{O}(T)$, the pollution level*

$$\frac{h(T)}{g(T)} = \Theta(T^{-0.5+\delta})$$

is enough for $\bar{R}_{\pi^*} > \bar{R}_{\tilde{\pi}^*}$ in the sense of order, where δ is any small positive number.

Proof. When $g(T) > \mathcal{O}(T)$,

$$\bar{R}_{\pi^*} = \frac{1}{4b} \frac{h^2(T)/g^2(T)}{[1/T + 1/g(T)]} + \mathcal{O}(1).$$

The critical average pollution level for $\bar{R}_{\pi^*} = \bar{R}_{\tilde{\pi}^*}$ is

$$\frac{h(T)}{g(T)} = \sqrt{\sigma^2 \left[\frac{1}{T} + \frac{1}{g(T)}\right] \log T} + o(1),$$

which is

$$\Theta\left[\sqrt{\frac{\log T}{T}}\right].$$

For $\bar{R}_{\pi^*} > \bar{R}_{\tilde{\pi}^*}$ in the sense of order, we just need

$$\frac{h(T)}{g(T)} > \mathcal{O}\left[\sqrt{\frac{\log T}{T}}\right],$$

such as

$$\frac{h(T)}{g(T)} = \Theta(T^{-0.5+\delta}),$$

where δ is any small positive number. □

See Figure 1 for the critical average pollution level (blue) as a function of $g(T)$ and two approximations (red, green) for $g(T) < T$ and $g(T) > T$. Specifically, if $g(T) = T$, the average pollution level $h(T)/g(T)$ just needs to be $T^{-0.5+\delta}$. If $h(T)/g(T) = \Theta(1)$, the number of offline data points $g(T)$ just needs to be $(\log T)^{1+\delta}$. Low-level pollution can already make offline data more harmful than useful.

When $g(T) < \mathcal{O}(T)$, the decision maker will finally learn the correct p^* during the online stage. Thus adding new polluted offline data points have an essential effect on $\bar{R}_{\pi^*}$, and the critical average pollution level

$$\Theta\left\{\sqrt{\frac{\log[g(T)]}{g(T)}}\right\}$$

decreases with $g(T)$.

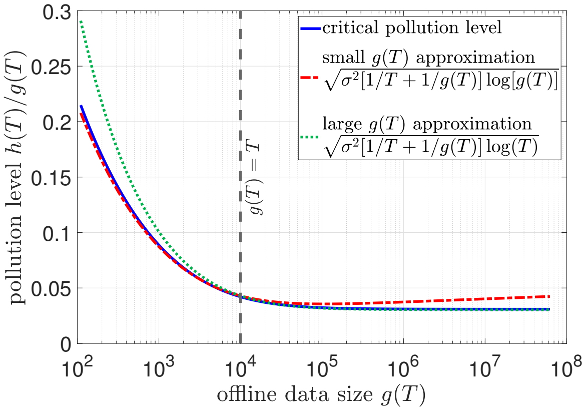

Figure 1: For the system with $a = 2$, $a_{\min} = 1$, $a_{\max} = 3$, $b = 1$, $\sigma^2 = \hat{\sigma}^2 = 1$, $T = 10000$, we plot the critical pollution level (blue solid line, calculated from directly summing over $\bar{R}^i_{\pi^*}$), the approximation for $g(T) < T$ (red dash-dotted line), and the approximation for $g(T) > T$ (green dotted line), as a function of $g(T)$. The black dashed vertical line indicates the boundary $g(T) = T = 10000$ between the two scenarios. When $g(T) < T$, the small $g(T)$ approximation (red) fits well with the critical pollution level. When $g(T) > T$, the large $g(T)$ approximation (green) fits well with the critical pollution level.

When $g(T) > \mathcal{O}(T)$, the effect of polluted offline data is saturated, meaning that the decision maker will always set

$$p^* = \frac{a}{2b} + \frac{1}{2b}\frac{h(T)}{g(T)} + o(1)$$

during the online stage. Thus the critical average pollution level is fixed at

$$\Theta\left[\sqrt{\frac{\log T}{T}}\right],$$

and further increasing $g(T)$ without changing $h(T)/g(T)$ has no effect on the order of $\bar{R}_{\pi^*}$.

4 CASES WHEN DECISION MAKER KNOWS THE EXISTENCE OF DATA POLLUTION

Assume the offline demand data $\hat{D}_i$ are polluted by an extra Gaussian noise:

$$\mathcal{N}\left[\frac{h'(T)}{g(T)}, \hat{\sigma}'^2 - \sigma^2\right].$$

In this scenario, the decision maker knows the existence of the pollution, and knows that the pollution magnitude is bounded:

$$|h'(T)| \leq h(T),$$

$$\sigma^2 \leq \hat{\sigma}'^2 \leq \hat{\sigma}^2.$$

The decision maker knows the values of $h(T)$ and $\hat{\sigma}^2$, but not the values of $h'(T)$ and $\hat{\sigma}'^2$. Otherwise, the pollution can be partially removed.

The decision maker has different choices:

(1) fully abandon the offline data and apply $\tilde{\pi}^*$;

(2) ignore the pollution and apply π^*;

(3) apply π^* first, and switch to $\tilde{\pi}^*$ at a proper time point (when $\bar{R}^i_{\pi^*} = \bar{R}^i_{\tilde{\pi}^*}$).

Choice (3) could perform better than (1) and (2), since at the beginning, although the offline data provide polluted information, they still lead to more accurate prices. Later, when the accurate online data have accumulated enough, the advantage of the offline data is gradually dominated by the disadvantage, and the offline data can be abandoned.

Inspired by Lemma 1, we consider policies that set p_{i+1} as a weighted average of $\hat{D}_j/(2b) + \hat{p}_j/2$ and $D_j/(2b) + p_j/2$. Denote the space of such policies as Π. If we regard the relative weight of each online data point as 1, then these approaches assign different weights to each offline data point: choice (1) assigns weight 0; choice (2) assigns weight 1; choice (3) assigns weight 1, then 0.

These choices are not optimal. We can explicitly calculate the optimal offline data weight. We need to optimize the regret for the worst case, namely when $|h'(T)| = h(T)$ and $\hat{\sigma}'^2 = \hat{\sigma}^2$.

Proposition 3. *When the decision maker knows the pollution magnitude, the optimal policy $\pi_0 \in \Pi$ is to assign weight 1 to each online data point, and weight*

$$r^* = \frac{g(T)\sigma^2}{h^2(T) + g(T)\hat{\sigma}^2}$$

to each offline data point. This means setting

$$\begin{aligned} p_{i+1} = &\sum_{j=1}^{g(T)} \left(\frac{\hat{D}_j}{2b} + \frac{\hat{p}_j}{2} \right) \frac{r^*}{i + r^* g(T)} \\ &+ \sum_{j=1}^{i} \left(\frac{D_j}{2b} + \frac{p_j}{2} \right) \frac{1}{i + r^* g(T)}, \end{aligned}$$

at online period $i + 1$.

Proof. At online period $i+1$, if we set the weight of each online data point as 1, and the weight of each offline data point as r, then the online price is

$$p_{i+1} = \sum_{j=1}^{g(T)} \left(\frac{\hat{D}_j}{2b} + \frac{\hat{p}_j}{2} \right) \frac{r}{i + rg(T)} + \sum_{j=1}^{i} \left(\frac{D_j}{2b} + \frac{p_j}{2} \right) \frac{1}{i + rg(T)},$$

and its distribution is Gaussian:

$$\mathcal{N} \left\{ \frac{a}{2b} + \frac{1}{2b} \frac{rh'(T)}{i + rg(T)}, \frac{1}{4b^2} \frac{i\sigma^2 + r^2 g(T) \hat{\sigma}'^2}{[i + rg(T)]^2} \right\}.$$

The expected regret of period $i+1$ is

$$\bar{R}^{i+1} = \frac{1}{4b} \frac{i\sigma^2 + r^2 h'^2(T) + r^2 g(T) \hat{\sigma}'^2}{[i + rg(T)]^2}.$$

Set $|h'(T)| = h(T)$ and $\hat{\sigma}'^2 = \hat{\sigma}^2$, then $\bar{R}^{i+1}$ can be transformed into

$$\frac{1}{4b} \left\{ \frac{\sigma^2}{i} - \frac{2g(T)\sigma^2}{i} \frac{r}{i + rg(T)} + \left[\frac{g^2(T)\sigma^2}{i} + h^2(T) + g(T)\hat{\sigma}^2 \right] \left[\frac{r}{i + rg(T)} \right]^2 \right\}.$$

Its minimum is achieved at

$$\frac{r}{i + rg(T)} = \frac{g(T)\sigma^2/i}{g^2(T)\sigma^2/i + h^2(T) + g(T)\hat{\sigma}^2},$$

meaning that

$$r^* = \frac{g(T)\sigma^2}{h^2(T) + g(T)\hat{\sigma}^2}.$$

□

Since $\sigma^2 \leq \hat{\sigma}^2$, when $h(T) > 0$, $0 < r^* < 1$. This is intuitive: offline data are polluted and less informative, thus should have a smaller weight.

Notice that the weight of offline data r^* only depends on $g(T), h(T), \sigma^2$, $\hat{\sigma}^2$, but not on i or T. This means the weight of offline data is independent with the online stage (current period or total period).

Proposition 4. *The overall expected regret of π_0 has order*

$$\Theta\left\{\log\left[1 \vee \frac{T}{g(T)} \vee \frac{h^2(T)T}{g^2(T)}\right]\right\}.$$

When the pollution level $h(T) > 0$, π_0 is better than fully believing in or ignoring offline data, meaning that $\bar{R}_{\pi_0} < \bar{R}_{\pi^*}$ and $\bar{R}_{\pi_0} < \bar{R}_{\tilde{\pi}^*}$. Besides, these inequalities could hold in the sense of order.

Proof. The expected regret of π_0 at period $i+1$ is

$$\bar{R}_{\pi_0}^{i+1} = \frac{\sigma^2}{4b}\frac{1}{i + \frac{g^2(T)\sigma^2}{h^2(T)+g(T)\hat{\sigma}^2}}.$$

We can directly see that $\bar{R}_{\pi_0}^{i+1} < \bar{R}_{\tilde{\pi}^*}^{i+1} = \sigma^2/(4bi)$. Thus $\bar{R}_{\pi_0} < \bar{R}_{\tilde{\pi}^*}$.

To prove $\bar{R}_{\pi_0} < \bar{R}_{\pi^*}$, we just need $\bar{R}_{\pi_0}^{i+1} < \bar{R}_{\pi^*}^{i+1}$. Since

$$\bar{R}_{\pi^*}^{i+1} = \frac{1}{4b}\frac{h^2(T) + i\sigma^2 + g(T)\hat{\sigma}^2}{[g(T)+i]^2},$$

$\bar{R}_{\pi_0}^{i+1} < \bar{R}_{\pi^*}^{i+1}$ is equivalent to

$$[ih^2(T) + ig(T)\hat{\sigma}^2 + g^2(T)\sigma^2][h^2(T) + g(T)\hat{\sigma}^2 + ig^2(T)]$$

$$> \sigma^2[h^2(T) + g(T)\hat{\sigma}^2][g(T)+i]^2,$$

which means

$$ih^4(T) + 2ig(T)h^2(T)(\hat{\sigma}^2 - \sigma^2) + ig^2(T)(\hat{\sigma}^2 - \sigma^2)^2 > 0.$$

Since $h(T) > 0$, $\hat{\sigma}^2 \geq \sigma^2$, the above inequality holds.

Summing over $\bar{R}_{\pi_0}^{i+1}$, we can see that the overall expected regret $\bar{R}_{\pi_0}$ is between

$$\frac{\sigma^2}{4b}\log\left[\frac{h^2(T)T + g(T)T\hat{\sigma}^2}{g^2(T)\sigma^2 + h^2(T) + g(T)\hat{\sigma}^2} + \frac{g^2(T)\sigma^2 + h^2(T) + g(T)\hat{\sigma}^2}{g^2(T)\sigma^2 + h^2(T) + g(T)\hat{\sigma}^2}\right]$$

and

$$\frac{\sigma^2}{4b}\log\left[\frac{h^2(T)T + g(T)T\hat{\sigma}^2 + g^2(T)\sigma^2}{g^2(T)\sigma^2}\right].$$

Thus the order of $\bar{R}_{\pi_0}$ is

$$\Theta\left\{\log\left[1 \vee \frac{T}{g(T)} \vee \frac{h^2(T)T}{g^2(T)}\right]\right\}.$$

When

$$g(T) = T^{3/2},$$
$$h(T) = T\sqrt{\log T},$$

we have

$$\bar{R}_{\pi_0} = \Theta(\log \log T),$$

which has a smaller order than

$$\bar{R}_{\pi^*} = \Theta(\log T)$$

and

$$\bar{R}_{\tilde{\pi}^*} = \Theta(\log T).$$

□

5 Discussion

In this paper, we introduce the online pricing with polluted offline data problem, where the offline data cannot faithfully reflect the online system. When the decision maker does not know the existence of pollution, and applies a policy that should be optimal if the offline data are not polluted, we study when the pollution makes the offline data harmful to online pricing. When the decision maker knows the pollution level is within a limited range, we design a policy that utilizes the polluted offline data in a better way.

The scenarios discussed in this paper can be generalized to other learning problems that utilize offline data, such as the multi-armed bandit problem. When the system is complex enough, different pollution formats (such as adding $\Theta(1)$ pollution on a small proportion of offline data points) have different effects, and should be discussed separately. Correspondingly, the decision maker has more countermeasures, such as abandoning suspicious data points.

Assume the decision maker knows the pollution level is within a limited range, but the system is complicated, so that the optimal offline data weight r^* cannot be explicitly determined. In this scenario, we could adopt choice (3), which starts with a policy that fully trusts offline data, and switch to a policy that does not utilize offline data.

From the viewpoint of information theory, offline data provide information gain for the online pricing problem. Correspondingly, data pollution can be regarded as information loss.

If the data pollution is from a malicious polluter, another question is: under limitations on the pollution level, what is the most effective pollution strategy, under which the online stage regret is maximized? If the decision maker does not know the existence of pollution, and applies a policy that should be optimal if the offline data are not polluted, this is a counterpart of the scenario in Section 4. If the decision maker knows the existence of pollution, this becomes a game theory problem.

In sum, we have a 2×2 problem matrix: the online pricing policy is fixed, or can be chosen by a decision maker that is aware of pollution; the pollution format is fixed, or can be chosen by a malicious maker that aims at increasing the overall regret.

In this paper, the pollution only performs on the offline stage. There are some generalizations about online data pollution. Consider a scenario, where the online demands observed by the decision maker are polluted, although the final overall regret is calculated with the true demands. Consider another scenario, where the online prices determined by the decision maker are polluted before sending into the system to generate the corresponding online demands, and the decision maker does not know the online prices after pollution.

References

Victor F Araman and René Caldentey. Dynamic pricing for nonperishable products with demand learning. *Operations Research*, 57(5):1169–1188, 2009.

Omar Besbes and Denis Sauré. Dynamic pricing strategies in the presence of demand shifts. *Manufacturing & Service Operations Management*, 16(4):513–528, 2014.

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Jinzhi Bu, David Simchi-Levi, and Yunzong Xu. Online pricing with offline data: Phase transition and inverse square law. In *International Conference on Machine Learning*, pages 1202–1210. PMLR, 2020.

Ningyuan Chen, Chun Wang, and Longlin Wang. Learning and optimization with seasonal patterns. *arXiv preprint arXiv:2005.08088*, 2020.

Wang Chi Cheung, David Simchi-Levi, and He Wang. Dynamic pricing and demand learning with limited price experimentation. *Operations Research*, 65(6):1722–1731, 2017.

Arnoud V den Boer. Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of Operations Research*, 39(3):863–888, 2014.

Arnoud V den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1): 1–18, 2015.

Arnoud V den Boer and N Bora Keskin. Dynamic pricing with demand learning and reference effects. *Management Science*, 2022.

Arnoud V den Boer and Bert Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, 60(3):770–783, 2014.

N Bora Keskin and Assaf Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014.

Ye Li, Hong Xie, Yishi Lin, and John CS Lui. Unifying offline causal inference and online bandit learning for data driven decision. In *Proceedings of the Web Conference 2021*, pages 2291–2303, 2021.

Po-Yi Liu, Chi-Hua Wang, and Heng-Hsui Tsai. Non-stationary dynamic pricing via actor-critic information-directed pricing. *arXiv preprint arXiv:2208.09372*, 2022.

Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates. *Available at SSRN 2765257*, 2016.

Yue Wang and Zeyu Zheng. Measuring policy performance in online pricing with offline data. *Available at SSRN 3729003*, 2021.