Measuring Policy Performance in Online Pricing with Offline Data: Worst-case Perspective and Bayesian Perspective

Yue Wang[1,2], **Zeyu Zheng**[3]

[1]Department of Computational Medicine, University of California, Los Angeles, CA 90095, USA; yuew@g.ucla.edu

[2]Institut des Hautes Études Scientifiques, Bures-sur-Yvette, Essonne 91440, France

[3]Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA; zyzheng@berkeley.edu

Abstract

The problems of online pricing with offline data, among other similar online decision making with offline data problems, aim at designing and evaluating online pricing policies in presence of a certain amount of existing offline data. To evaluate pricing policies when offline data are available, the decision maker can either position herself at the time point when the offline data are already observed and viewed as deterministic, or at the time point when the offline data are not yet generated and viewed as stochastic. We write a framework to discuss how and why these two different positions are relevant to online policy evaluations, from a worst-case perspective and from a Bayesian perspective. We then use a simple online pricing setting with offline data to illustrate the constructions of optimal policies for these two approaches and discuss their differences, especially whether we can decompose the searching for the optimal policy into independent subproblems and optimize separately, and whether there exists a deterministic optimal policy.

Keywords: Online pricing, offline data, performance measure, worst-case approach, Bayesian approach

1 Introduction and Setup

For online learning problems, there are probability distributions or system parameters that are unknown a priori and need to be learned from data that arrive sequentially. Actions by a decision maker can impact the way that data arrive and therefore her own ability to learn the unknown system. A decision maker then dynamically sets actions to learn the unknown system while striving to accumulate rewards. A typical goal in online learning problems is to minimize the so-called *regret*. (We deliberately introduce no formal definition for regret at this point and defer to the later part of the paper.) On the other hand, for offline learning problems, data are already observed, stored offline, and presented to a decision maker. The decision maker uses the data to infer about probability distributions and system parameters. A typical objective function in offline learning problems is the out-of-sample performance. Even though the two sets of problems, online and offline, did not cross much in the literature, there is a rising interest in applications where online learning algorithms

need to be designed for future, while some informative historical offline data are readily available from the past. Thus, in the framework of online learning with offline data, a natural goal is to design online learning policies that can integrate offline data. In order to compare different policies, as a first step, one needs to clearly define how to measure the performance of a policy. Our work focuses largely on analyzing this first step of proposing performance measures, rather than designing policies under a certain performance measure.

To fix ideas, we consider a framework of online pricing problem with offline data (OPOD). We briefly describe the setup of OPOD using a plain and basic form. The online stage has T periods. The decision maker knows the value of T. For period $t = i$, the decision maker sets a price $p_i \in [p_{\min}, p_{\max}]$ with $0 < p_{min} < p_{max} < \infty$. Then the system produces the demand $D_i = a - bp_i + \epsilon_i$, and the decision maker receives a random revenue $p_i D_i$ for this period, whose expectation is $p_i(a - bp_i)$. Here the model parameter $\theta = (a, b)$ in the demand function is unknown, where $\theta \in \Theta = [a_{min}, a_{max}] \times [b_{min}, b_{max}]$ with $0 < a_{min} < a_{max} < \infty$ and $0 < b_{min} < b_{max} < \infty$. The noise ϵ_i has 0 mean, and different $\epsilon_1, \ldots, \epsilon_i, \ldots, \epsilon_T$ are assumed to be independent and identically distributed.

In the offline stage, the decision maker observes some k periods of operations. That is, prices $\hat{p}_1, \ldots, \hat{p}_k$ and corresponding demands $\hat{D}_1, \ldots, \hat{D}_k$ with $\hat{D}_i = a - b\hat{p}_i + \hat{\epsilon}_i$ are observed. Here different $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_i, \ldots, \hat{\epsilon}_k$ are assumed to be independent and identically distributed. The realized demand is typically assumed to be bounded and non-negative in the literature. When offline data are available, it sometimes matters to know how the offline actions have been decided. To this regard, we denote the method that is used to generate the offline prices $\hat{p}_1, \ldots, \hat{p}_k$ as method $\hat{\pi}$, possibly with stochasticity. Define $\hat{H} = \{\hat{p}_1, \ldots, \hat{p}_k, \hat{D}_1, \ldots, \hat{D}_k\}$ to be the collection of all the offline data. Define $\mathbb{H}$ to be the set of all possible offline data, which is determined by $\hat{\pi}$. A policy π for the online stage is a price-determining mechanism, that at each period $t = i$ of online stage, it uses the data from offline stage as well as all previous online periods, to determine the price for the current period (possibly with stochasticity): $p_i = \pi(\hat{H}, p_1, D_1, \ldots, p_{i-1}, D_{i-1})$.

For a given $\theta = (a, b)$, the best price that maximizes the expected single-period revenue $p(a - bp)$ is $p^*(\theta) = a/(2b)$, and the maximal expected single-period revenue is therefore $a^2/(4b)$. The basic online pricing problem assumes that $p^*(\theta) = a/(2b) \in [p_{\min}, p_{\max}]$. For a given price p_0 taken in a period, the expected single-period regret is given by $a^2/(4b) - p_0(a - bp_0)$. For a parameter pair $\theta = (a, b)$, offline data $\hat{H}$, and a policy corresponding to the offline data, $\pi \mid \hat{H}$, the total expected regret is

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

where the expectation is with respect to p_i, which is determined by $\pi \mid \hat{H}$.

We consider two approaches that are commonly used to measure policy performances - a worst-case approach and a Bayesian approach. We first specify the understanding of probability distributions for these two approaches. For the Bayesian approach, we require that the dis-

tribution of ϵ_i is known to belong to a distribution family that can be parameterized by finitely many compact parameters (e.g., Gaussian with unknown variance). For this case, the parameters that are used to describe the distribution family should also be added into θ. For the worst-case approach, this strict requirement is no necessary. A convenient requirement is to assume the noise belongs to a distribution family which is metrizable, and it is compact under this metric. Nevertheless, to unify the discussion, we assume the noise family can be parameterized by finitely many compact parameters.

We introduce the ideas of the worst-case approach and the Bayesian approach and other related references in Section 2.

To define the regret measurement for a policy, we position the decision maker either before or after the concrete offline data are observed. These two scenarios both behave differently under the worst-case approach and the Bayesian approach. For the worst-case approach, when we have observed concrete offline data $\hat{H}$, the value of $\hat{H}$ does not affect this measurement. We discuss this problem in Section 3.

For a given problem, an optimal policy π^* minimizes the regret measurement for the worst-case approach or the Bayesian approach. The existence of π^* is illustrated in Appendix A. We discuss the properties of the optimal policies in Section 4. For the Bayesian approach, the problem of determining the optimal policy can be decomposed into some independent subproblems depending on the offline data, so that the policy that optimizes each subproblem is the optimal policy for the overall problem. Besides, and there always exists a deterministic optimal policy for the Bayesian approach. These properties may not hold for the worst-case approach.

The pricing problem introduced above is a special case in a class of reinforcement learning problems. We provide an example of this general setting in Section 5, so as to further illustrate the properties of the optimal policies under both approaches. We finish this paper with summary and discussions in Section 6. Specifically, we illustrate that most results in this paper also hold for the generalized reinforcement learning problems.

2 Related Works

Online pricing problem or dynamic pricing problem, as a specific online learning problem, has been a prominent online learning topic in the areas of operations research and management science. We refer to den Boer (2015) and Gallego and Topaloglu (2019) for comprehensive reviews. When pre-existing offline data are available prior to the online pricing problem, this framework of OPOD is formally introduced by Bu et al. (2020). Keskin and Zeevi (2014) also discuss the impact of pre-existing information on the online pricing problems. This OPOD framework is a special case of the general reinforcement learning problem with offline data, generally called off-policy reinforcement learning or reinforcement learning with replay buffer (Munos et al., 2016; Fujimoto et al., 2019; Thomas and Brunskill, 2016; Rakelly et al., 2019; Eysenbach et al., 2019; Rolnick et al., 2019). Nevertheless, these papers focus on designing efficient policies, while we focus on comparing policy evaluation measurements.

For a policy π under a given parameter vector θ, we can average the total expected regret $R^{\theta}_{\pi|\hat{H}}$ over $\hat{H}$ to obtain the overall regret $\bar{R}^{\theta}_{\pi}$. This regret mapping $\Theta \to \mathbb{R} : \theta \to \bar{R}^{\theta}_{\pi}$ can be used to measure the performance of π. There are two general approaches in the literature of online pricing: the worst-case approach (for example, Keskin and Zeevi (2014), den Boer and Zwart (2015), Ban and Keskin (2021)) and the Bayesian approach (for example, Harrison et al. (2012); Russo and Van Roy (2014); Bastani et al. (2022)). For the worst-case approach (also called the frequentist approach), the measurement is $\max_{\theta\in\Theta} \bar{R}^{\theta}_{\pi}$, namely the L^{∞} norm of $\bar{R}^{\theta}_{\pi}$. (When the noise belongs to a general family, one also needs to take maximum with respect to the noise.) This approach concerns the performance for the worst case among all possible parameters. There is a variant of the worst-case approach: discard the worst η proportion and determine the worst case without taking expectations (Kirschner and Krause, 2018; Zanette et al., 2020). For the Bayesian approach, there is a prior distribution $f(\theta)$ on the space of parameters Θ. The measurement is $\int_{\Theta} f(\theta)\bar{R}^{\theta}_{\pi}\mathrm{d}\theta$, namely the L^1 norm of $\bar{R}^{\theta}_{\pi}$. This approach concerns the average performance. The Bayesian approach measurement is also called Bayes regret (Bastani et al., 2022) or Bayes risk (Russo and Van Roy, 2014). Certainly, how to determine the prior distribution is a tricky problem by itself. These papers generally choose one approach to evaluate the policies. We compare both approaches, especially properties of the optimal policy under both approaches.

When the online pricing is accompanied by offline data, there is another measure, the "instance-dependent regret" (Bu et al., 2020). We briefly describe this measure. Set $\hat{p}$ to be the average of offline prices. Assume we know the true parameter θ^*. Define $\delta = |\hat{p} - p^*(\theta^*)|$, where $p^*(\theta^*)$ is the optimal price for θ^*. The instance-dependent regret considers the worst case in a subset of Θ, which depends on the offline data. The definition is $\max_{\theta\in\Theta}(\bar{R}^{\theta}_{\pi} : |p^*(\theta) - \hat{p}| \in [(1-\xi)\delta, (1+\xi)\delta])$, where $\xi \in (0,1)$ is a constant. The choice of ξ reflects how conservative the target is and may need to be carefully chosen. We will not further discuss the instance-dependent regret in this paper.

3 Evaluation Approaches with Offline Data

3.1 Before the Observation of Concrete Offline Data

In this subsection, we discuss the evaluation measurement standing at the time point when the offline data has not yet been generated. Specifically, at this time point, the concrete offline data has not been observed yet but the method to produce the offline data is known. That is, for the offline stage, the price $\hat{p}_{i+1}$ is determined as a deterministic or stochastic function $\hat{p}_{i+1} = \hat{\pi}(\hat{p}_1, \ldots, \hat{p}_i, \hat{D}_1, \ldots, \hat{D}_i)$, and the function $\hat{\pi}$ is known to the decision maker. The demand $\hat{D}_{i+1}$ depends on $\hat{p}_{i+1}, a, b$, such that $\hat{D}_{i+1} = a - b\hat{p}_{i+1} + \hat{\epsilon}_{i+1}$. As an illustration, in the simplest case, $\hat{p}_1, \ldots, \hat{p}_k$ are fixed, or are drawn independently from a given distribution. The space of possible offline data $\mathbb{H}$ depends on $\hat{\pi}$. Since the decision maker knows the format of $\hat{\pi}$, each $\theta = (a, b)$ determines a known distribution

of the offline data $\hat{H} = (\hat{p}_1, \ldots, \hat{p}_k, \hat{D}_1, \ldots, \hat{D}_k)$.

For each $\theta = (a, b)$, we can calculate the conditional probability density of offline data $\hat{H}$, namely $f_{\hat{\pi}}(\hat{H} \mid \theta)$, depending on the offline price determination mechanism $\hat{\pi}$. Integrating $R^{\theta}_{\pi|\hat{H}}$, the regret of policy π with offline data $\hat{H}$ and parameter θ, we obtain the overall regret of π under θ as $\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}$.

Definition 1. *For the worst-case approach, before observing the concrete offline data $\hat{H}$, the measurement for a policy π is the regret of the worst case for $\theta \in \Theta$: $\max_{\theta\in\Theta}\{\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}\}$. When a policy π does not consider the offline data $\hat{H}$, this measurement degenerates into $\max_{\theta\in\Theta}\{R^{\theta}_{\pi}\}$.*

For the Bayesian approach, the measurement is the regret averaged with respect to the prior distribution $f(\theta)$ of $\theta \in \Theta$, namely $\int_{\Theta} f(\theta) \left[\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}\right] \mathrm{d}\theta$, in which $f(\theta)$ represents the prior density of the parameter θ. Define $f_{\hat{\pi}}(\hat{H}) = \int_{\Theta} f_{\hat{\pi}}(\hat{H} \mid \theta) f(\theta) \mathrm{d}\theta$ to be the density of offline data $\hat{H}$, $f_{\hat{\pi}}(\theta, \hat{H}) = f_{\hat{\pi}}(\hat{H} \mid \theta) f(\theta)$, and $f_{\hat{\pi}}(\theta \mid \hat{H}) = f_{\hat{\pi}}(\theta, \hat{H}) / f_{\hat{\pi}}(\hat{H})$. We can transform the measurement as:

$$
\begin{aligned}
&\int_{\Theta} f(\theta) \left[\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}\right] \mathrm{d}\theta \\
=&\int_{\Theta} \int_{\mathbb{H}} f_{\hat{\pi}}(\theta, \hat{H}) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H} \mathrm{d}\theta \\
=&\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H}) \left[\int_{\Theta} f_{\hat{\pi}}(\theta \mid \hat{H}) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\theta\right] \mathrm{d}\hat{H}
\end{aligned}
$$

Switching the order of integration is legitimized by the Fubini-Tonelli theorem, since the integrand is non-negative. For two $\theta_1, \theta_2 \in \Theta$, we have

$$
\begin{aligned}
&\frac{f_{\hat{\pi}}(\theta_1 \mid \hat{H})}{f_{\hat{\pi}}(\theta_2 \mid \hat{H})} = \frac{f(\theta_1) f_{\hat{\pi}}(\hat{H} \mid \theta_1)}{f(\theta_2) f_{\hat{\pi}}(\hat{H} \mid \theta_2)} \\
=&\frac{f(\theta_1) f_{\hat{\pi}}(\hat{p}_1) f(\hat{D}_1 \mid \theta_1, \hat{p}_1)}{f(\theta_2) f_{\hat{\pi}}(\hat{p}_1) f(\hat{D}_1 \mid \theta_2, \hat{p}_1)} \\
&\times \frac{f_{\hat{\pi}}(\hat{p}_2 \mid \hat{p}_1, \hat{D}_1) f(\hat{D}_2 \mid \theta_1, \hat{p}_2)}{f_{\hat{\pi}}(\hat{p}_2 \mid \hat{p}_1, \hat{D}_1) f(\hat{D}_2 \mid \theta_2, \hat{p}_2)} \times \cdots \\
&\times \frac{f_{\hat{\pi}}(\hat{p}_k \mid \hat{p}_1 \hat{D}_1 \cdots \hat{p}_{k-1} \hat{D}_{k-1}) f(\hat{D}_k \mid \theta_1, \hat{p}_k)}{f_{\hat{\pi}}(\hat{p}_k \mid \hat{p}_1 \hat{D}_1 \cdots \hat{p}_{k-1} \hat{D}_{k-1}) f(\hat{D}_k \mid \theta_2, \hat{p}_k)} \\
=&\frac{f(\theta_1) \prod_{i=1}^{k} f(\hat{D}_i \mid \theta_1, \hat{p}_i)}{f(\theta_2) \prod_{i=1}^{k} f(\hat{D}_i \mid \theta_2, \hat{p}_i)}
\end{aligned}
\tag{2}
$$

Therefore, $f_{\hat{\pi}}(\theta \mid \hat{H})$ only depends on the values of θ and $\hat{H}$, but not on $\hat{\pi}$, and we can denote it by $f(\theta \mid \hat{H})$. Now we can present the transformed definition for the Bayesian measurement.

Definition 2. *For the Bayesian approach, before observing the concrete offline data $\hat{H}$, the measurement for a policy π is $\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H}) \left[\int_{\Theta} f(\theta \mid \hat{H}) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\theta\right] \mathrm{d}\hat{H}$. When a policy π does not consider the offline data $\hat{H}$, this measurement degenerates into $\int_{\Theta} f(\theta) R^{\theta}_{\pi} \mathrm{d}\theta$.*

3.2 After the Observation of Concrete Offline Data

In this subsection, we discuss the evaluation measurement standing at the time point when the concrete offline data has been observed and given. That is, the decision maker has observed concrete prices and demands in the offline data: $\hat{H} = (\hat{p}_1, \ldots, \hat{p}_k, \hat{D}_1, \ldots, \hat{D}_k)$.

Before going into the analysis, we make an assumption about the distribution of the independent and identically distributed random noise ϵ (also $\hat{\epsilon}$). Consider an offline data pair

$(\hat{p}_i, \hat{D}_i)$, which can be generated by a parameter $\theta_1 = (a_1, b_1)$. This means the probability density satisfies $f(\hat{\epsilon}_i = \hat{D}_i - a_1 + b_1\hat{p}_i) > 0$. We assume that for all $\theta_2 = (a_2, b_2) \in \Theta$, $f(\hat{\epsilon}_i = \hat{D}_i - a_2 + b_2\hat{p}_i) > 0$. Therefore, $(\hat{p}_i, \hat{D}_i)$ can be also generated by $\theta_2 = (a_2, b_2)$. This means that if $f(\theta_1 \mid \hat{H}) > 0$ for one θ_1, then $f(\theta_2 \mid \hat{H}) > 0$ for all $\theta_2 \in \Theta$. Thus, with the observation of any $\hat{H}$, the range of possible θ remains Θ. This assumption is aligned with standard motivations of truncated Gaussian or Poisson type of distributions in the literature.

For the worst-case approach, no matter what offline data are concretely observed, the range of possible θ is still Θ, even though the possibility that the offline data are generated from some θ may be inferred to be very small. Thus with concrete $\hat{H}$, the regret measurement degenerates from $\max_{\theta\in\Theta}\{\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}\}$ by removing the integration with $\hat{H}$.

Definition 3. *For the worst-case approach, after observing concrete offline data $\hat{H}$, the measurement for a policy π is* $\max_{\theta\in\Theta} R^{\theta}_{\pi|\hat{H}}$.

We have a seemingly surprise result that different concrete offline data cannot affect the optimal policy.

Proposition 1. *For the worst-case approach, an optimal policy π^* for the scenario without offline data, which minimizes* $\max_{\theta\in\Theta} R^{\theta}_{\pi}$*, is also optimal for the scenario with concrete offline data $\hat{H}$, meaning that it minimizes* $\max_{\theta\in\Theta} R^{\theta}_{\pi|\hat{H}}$*, and this holds for any $\hat{H}$.*

Proof. Define Π to be the space of policies that could consider offline data, and $\tilde{\Pi}$ to be the space of policies that do not consider offline data. For a policy $\tilde{\pi} \in \tilde{\Pi}$ that does not consider offline data, with concrete $\hat{H}$, the worst-case measurement is $\max_{\theta\in\Theta} R^{\theta}_{\tilde{\pi}|\hat{H}} = \max_{\theta\in\Theta} R^{\theta}_{\tilde{\pi}}$, the same as the situation without offline data. For a policy $\pi \in \Pi$ that uses $(\hat{H}, p_1, D_1, \ldots, p_{i-1}, D_{i-1})$ to determine p_i, when $\hat{H}$ is deterministic, this policy $\pi \mid \hat{H}$ can be replicated by a policy $\tilde{\pi} \in \tilde{\Pi}$ that only uses $(p_1, D_1, \ldots, p_{i-1}, D_{i-1})$ to determine p_i. The replication means that for the given deterministic $\hat{H}$, these two policies have the same total expected regret for a given θ: $R^{\theta}_{\pi|\hat{H}} = R^{\theta}_{\tilde{\pi}}$. Thus $\max_{\theta\in\Theta} R^{\theta}_{\pi|\hat{H}} = \max_{\theta\in\Theta} R^{\theta}_{\tilde{\pi}}$. Then for each offline data $\hat{H}$, $\min_{\pi\in\Pi}\max_{\theta\in\Theta} R^{\theta}_{\pi|\hat{H}} = \min_{\tilde{\pi}\in\tilde{\Pi}}\max_{\theta\in\Theta} R^{\theta}_{\tilde{\pi}}$, meaning that with concrete offline data $\hat{H}$ or without any offline data, the worst-case measurements for the optimal policy are the same. □

For the Bayesian approach, the regret measurement before observing offline data is $\int_{\mathbb{H}} f_{\hat{\pi}}(\hat{H})[\int_{\Theta} f(\theta \mid \hat{H}) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\theta]\mathrm{d}\hat{H}$, meaning that we first calculate the total regret $\int_{\Theta} f(\theta \mid \hat{H}) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\theta$ for each $\hat{H}$, and then integrate with $\hat{H}$. In this subsection, since we have concrete offline data, the outer integration over $\hat{H}$ can be omitted.

Definition 4. *For the Bayesian approach, after observing concrete offline data $\hat{H}$, the measurement for a policy π is* $\int_{\Theta} f(\theta \mid \hat{H}) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\theta$.

Compared to the case without offline data, $\int_{\Theta} f(\theta) R^{\theta}_{\tilde{\pi}} \mathrm{d}\theta$, the prior distribution $f(\theta)$ is replaced by the posterior distribution $f(\theta \mid \hat{H})$, due to the offline data $\hat{H}$. Besides, with different values of $\hat{H}$, the posterior distribution $f(\theta \mid \hat{H})$ might differ. Therefore, *the Bayesian measurement for the same policy can depend on the concrete offline data.* Proposition 1 does not hold

for the Bayesian measurement. For different $\hat{H}$, the optimal policy design might also be different. Notice that this measurement depends on the concrete value of $\hat{H}$, but not on $\hat{\pi}$, which determines the distribution of $\hat{H}$.

In the Bayesian measurement for a given policy π, only the $f(\theta \mid \hat{H})$ term depends on $\hat{H}$. From Eq. 2, $f(\theta \mid \hat{H})$ only depends on $f(\hat{D}_i \mid \theta, \hat{p}_i)$. Therefore, only the set $\{(\hat{p}_1, \hat{D}_1), \ldots, (\hat{p}_k, \hat{D}_k)\}$ matters, and we can rearrange the order of offline data, with the Bayesian measurement being invariant.

In Appendix B, we provide an example to illustrate why the knowledge of concrete offline data does not change the problem with the worst-case approach, compared to the problem without any offline data, and why concrete offline data affect the Bayesian measurement.

4 Optimal Policy Analysis

4.1 Optimal Policy for the Worst-case Approach

In this subsection, we consider the worst-case regret measurement calculated before the observation of concrete offline data. An optimal policy $\bar{\pi}^*$ minimizes $\max_{\theta\in\Theta}\{\int_{\mathbb{H}} f(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}\}$.

In general, $\bar{\pi}^*$ can be too complicated to find in closed-form, since we need to design the behavior of $\bar{\pi}^*$ for each possible offline data $\hat{H} \in \mathbb{H}$. It is intuitive to decompose the set of all possible offline data $\mathbb{H}$ into disjoint subsets (possibly uncountable) $\mathbb{H} = \cup_{i=1}^{m} \mathbb{H}_i$, according to different observed prices or demands. Then we design the optimal policy $\bar{\pi}^*_i$ for each subproblem on $\mathbb{H}_i$, meaning that it minimizes $\max_{\theta\in\Theta}\{\int_{\mathbb{H}_i} f(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}\}$. We hope that the optimal policy $\bar{\pi}^*$ for the whole problem can be constructed as follows: apply $\bar{\pi}^*_1$ if $\hat{H} \in \mathbb{H}_1$, ..., apply $\bar{\pi}^*_m$ if $\hat{H} \in \mathbb{H}_m$. If this idea is valid, then we can greedily optimize for each subproblem on $\mathbb{H}_i$. To decompose $\mathbb{H}$, one way is to split by offline prices, i.e., $\mathbb{H}_1 = \{\hat{H} : \hat{p} = \hat{c}_1\}, \ldots, \mathbb{H}_m = \{\hat{H} : \hat{p} = \hat{c}_m\}$. Another way of decomposing $\mathbb{H}$ is to split by both offline prices and demands, i.e., $\mathbb{H} = \cup_{\hat{H}\in\mathbb{H}}\{\hat{H}\}$. In the latter case, we optimize for each $\{\hat{H}\}$, namely minimizing $\max_{\theta\in\Theta}\{R^{\theta}_{\pi|\hat{H}}\}$.

In Section 4.1.1, we show that for the worst-case approach, the decomposition of $\mathbb{H}$ does not work. We cannot optimize for each subproblem on $\mathbb{H}_i$ and combine the optimal policy for each subproblem to obtain the optimal policy for the overall problem. This means that different subproblems are not independent, and greedily improve one subproblem might not improve the overall problem. For the same $\hat{H} \in \mathbb{H}_1$, the globally optimal policy might behave differently, depending on other subproblems on $\mathbb{H}_2, \ldots, \mathbb{H}_m$, or more specifically, on $\hat{\pi}$, which determines the distribution of offline data $\hat{H}$.

Another question is whether there always exists a deterministic optimal policy. If so, the problem of searching for the optimal policy is further simplified, since there are many more stochastic policies. In Section 4.1.2, we prove that there might not exist a deterministic optimal policy.

4.1.1 The Optimal Policy for the Overall Problem Might Not Be Optimal for Subproblems.

We present an example, where a policy π is optimal for each subproblem of minimizing $\max_{\theta\in\Theta}\{\int_{\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$ for $\mathbb{H}_1 = \{\hat{H} : \hat{p} = \hat{c}_1\}$ and $\mathbb{H}_2 = \{\hat{H} : \hat{p} = \hat{c}_2\}$, but does not minimize $\max_{\theta\in\Theta}\{\int_{\mathbb{H}_1\cup\mathbb{H}_2} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$. Besides, the optimal policy that minimizes $\max_{\theta\in\Theta}\{R^{\theta}_{\pi|\hat{H}}\}$ for each $\hat{H} \in \mathbb{H}$ does not minimize $\max_{\theta\in\Theta}\{\int_{\mathbb{H}} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$. Therefore, we cannot decompose the problem of minimizing $\max_{\theta\in\Theta}\{\int_{\mathbb{H}} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$ into subproblems of minimizing $\max_{\theta\in\Theta}\{\int_{\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$ or subproblems of minimizing $\max_{\theta\in\Theta}\{R^{\theta}_{\pi|\hat{H}}\}$.

This example is not very surprising. Consider two policies π_1, π_2. Assume the set of offline data is decomposed into $\mathbb{H} = \cup_{i=1}^{m}\mathbb{H}_i$. Define

$$g_i(\theta) = \int_{\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi_1|\hat{H}}\mathrm{d}\hat{H}$$

and

$$h_i(\theta) = \int_{\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi_2|\hat{H}}\mathrm{d}\hat{H}$$

The worst-case measurement of π_1 for the subproblem on $\mathbb{H}_i$ is the L^∞ norm of $g_i(\theta)$:

$$||g_i||_\infty = \max_{\theta\in\Theta}\{\int_{\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$$

The worst-case measurement of π_1 for the overall problem on $\mathbb{H}$ is the L^∞ norm of $\sum_{i=1}^{m} g_i(\theta)$:

$$\begin{aligned}||\sum_{i=1}^{m} g_i||_\infty =& \max_{\theta\in\Theta}\{\sum_{i=1}^{m}\int_{\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}\\ =& \max_{\theta\in\Theta}\{\int_{\mathbb{H}=\cup_{i=1}^{m}\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}\end{aligned}$$

Assume π_1 is better than π_2 for each subproblem, meaning that $||f_i||_\infty < ||g_i||_\infty$ for $i = 1, \ldots, m$. This does not necessarily imply that π_1 is better than π_2 for the overall problem, and we might have $||\sum f_i||_\infty > ||\sum g_i||_\infty$. This is a natural property of the L^∞ norm.

Proposition 2. *For the worst-case approach, (a) when we split the set of possible offline data $\mathbb{H}$ into disjoint subsets $\mathbb{H}_i$ according to offline prices, a policy π that minimizes $\max_{\theta\in\Theta}\{\int_{\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$ for each $\mathbb{H}_i$ may not minimize $\max_{\theta\in\Theta}\{\int_{\mathbb{H}=\cup\mathbb{H}_i} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$. (b) When we split the set of possible offline data $\mathbb{H}$ according to both offline prices and demands, namely decomposing into each individual $\hat{H}$, a policy that minimizes $\max_{\theta\in\Theta} R^{\theta}_{\pi|\hat{H}}$ for each $\hat{H}$ might not minimize $\max_{\theta\in\Theta}\{\int_{\mathbb{H}} f(\hat{H} \mid \theta)R^{\theta}_{\pi|\hat{H}}\mathrm{d}\hat{H}\}$.*

Proof. We use the following example to prove this proposition.

Example 1. *Set $\theta = (a, b) \in \Theta = [8, 10] \times [4, 5]$ and $p \in [0.7, 1.25]$. The noise ϵ is Gaussian with mean 0 and variance 10^{-20}. The online stage has one period. In the offline stage, possible prices we can observe are $\hat{p} = \hat{c}_1 = 0.8$ and $\hat{p} = \hat{c}_2 = 1.25$. We only observe once in the offline stage. Since $\hat{D} = a - b\hat{p} + \hat{\epsilon}$, the offline data can tell the approximate value of $a - 0.8b$ or $a - 1.25b$ with very small errors. We consider three distributions (i.e., three different $\hat{\pi}$) of the offline data $(\hat{p}, \hat{D})$: (1) only observe $\hat{p} = \hat{c}_1 = 0.8$; (2) only observe $\hat{p} = \hat{c}_2 = 1.25$; (3) observe $\hat{p} = \hat{c}_1 = 0.8$ with probability 0.5, and observe $\hat{p} = \hat{c}_2 = 1.25$ with probability 0.5.*

When we consider the offline data (before they are explicitly observed), since the online stage only runs for one period, the optimal pol-

icy must be deterministic. For example, consider a stochastic policy that sets $p = c_1$ with probability P_1, and sets $p = c_2$ with probability $1 - P_1$, then its expected revenue $p(a - bp)$ is less than that of another policy that sets $p = P_1c_1 + (1 - P_1)c_2$ with probability 1, for any $\theta = (a, b)$. The reason is that the revenue function $p(a-bp)$ is convex up, and we can apply Jensen's inequality.

The space of deterministic policy in this example is simple enough, and we can use brute force to search for the optimal policy.

In case (1), we know $(\hat{p}, \hat{D}) = (0.8, a - 0.8b + \hat{\epsilon})$. A θ near the line segment $a - 0.8b = \hat{D}$ is more likely to produce this $\hat{D}$. The optimal policy is to set $p = 0.1713\hat{D} + 0.1148$ when $\hat{D} < 4.8$, $p = 0.1117\hat{D} + 0.4008$ when $4.8 \le \hat{D} \le 6$, and $p = 0.2238\hat{D} - 0.2718$ when $\hat{D} > 6$. The largest regret $R = 0.0251$ is achieved by $\theta_1 = (10, 5)$ and $\theta_2 = (9.2, 4)$. Denote this policy by π_1.

In case (2), we know $(\hat{p}, \hat{D}) = (1.25, a - 1.25b + \hat{\epsilon})$. The optimal policy is to set $p = 0.1280\hat{D} + 0.5760$ when $\hat{D} < 3$, $p = 0.1120\hat{D} + 0.6240$ when $3 \le \hat{D} \le 3.75$, and $p = 0.1648\hat{D} + 0.4260$ when $\hat{D} > 3.75$. The largest regret $R = 0.0098$ is achieved by $\theta_1 = (10, 5)$ and $\theta_3 = (8.75, 4)$. Denote this policy by π_2.

In case (3), consider the policy: if $\hat{p} = 0.8$, apply π_1; if $\hat{p} = 1.25$, apply π_2. Denote this policy by π_3. For this policy, the largest regret is only achieved by $\theta_1 = (10, 5)$, with $R = 0.0174$.

Consider a policy π_0 that directly sets $p = 1$, regardless of offline data. With π_0, the regret for $\theta_1 = (10, 5)$ is 0, although it behaves badly for θ that is not close to $\theta_1 = (10, 5)$. We construct a new policy π_4: apply π_3 with probability 0.99, and apply π_0 with probability 0.01. The regret measurement $R = 0.0173$, meaning that π_4 is better than π_3, and π_3 is not the optimal policy for case (3).

In Example 1, if we do not consider the offline data, then the optimal policy $\tilde{\pi}$ that minimizes $\max_{\theta\in\Theta} R^{\theta}_{\tilde{\pi}}$ is to set $p = 1.0125$, and the largest regret $R = 0.2257$ is achieved by $\theta_4 = (8, 5)$ and $\theta_5 = (10, 4)$. By Proposition 1, this $\tilde{\pi}$ also minimizes $\max_{\theta\in\Theta} R^{\theta}_{\pi|\hat{H}}$ for any concrete $\hat{H}$. The regret of $\tilde{\pi}$ is much larger than the regret in any of the above situations. Thus it does not minimizes $\max_{\theta\in\Theta}\{\int_{\mathbb{H}} f(\hat{H} \mid \theta) R^{\theta}_{\pi|\hat{H}} \mathrm{d}\hat{H}\}$. □

See Figure 1 for the contour plots of regret function for $\pi_1, \pi_2, \pi_3, \pi_0$ in Example 1. Brighter colors represent larger regrets. The x-axis is parameter a, and the y-axis is parameter b. Red circles are the global maximal points. Policy π_0 behaves well near the worst cases of π_3. Thus a mixture of π_3 and π_0 is better than pure π_3.

The following lemma explains why we can construct π_4 to defeat π_3, whose largest regret is only achieved by a single θ_1.

Lemma 1. *Under the worst-case approach, for a policy π_1, if the total expect regret function R^{θ}_1 has a unique maximal point $\theta_0 = (a_0, b_0)$, then this policy is not optimal.*

Proof of Lemma 1. Under any policy, the distribution of price-demand (online and offline) sequence is continuous with θ under total variation distance, then the total expected regret R^{θ}, as a continuous function on price-demand sequence, is also a continuous function on the compact set Θ. Consider a policy π_2 that always sets $p = a_0/(2b_0)$. The total regret R^{θ}_2 of π_2 takes 0

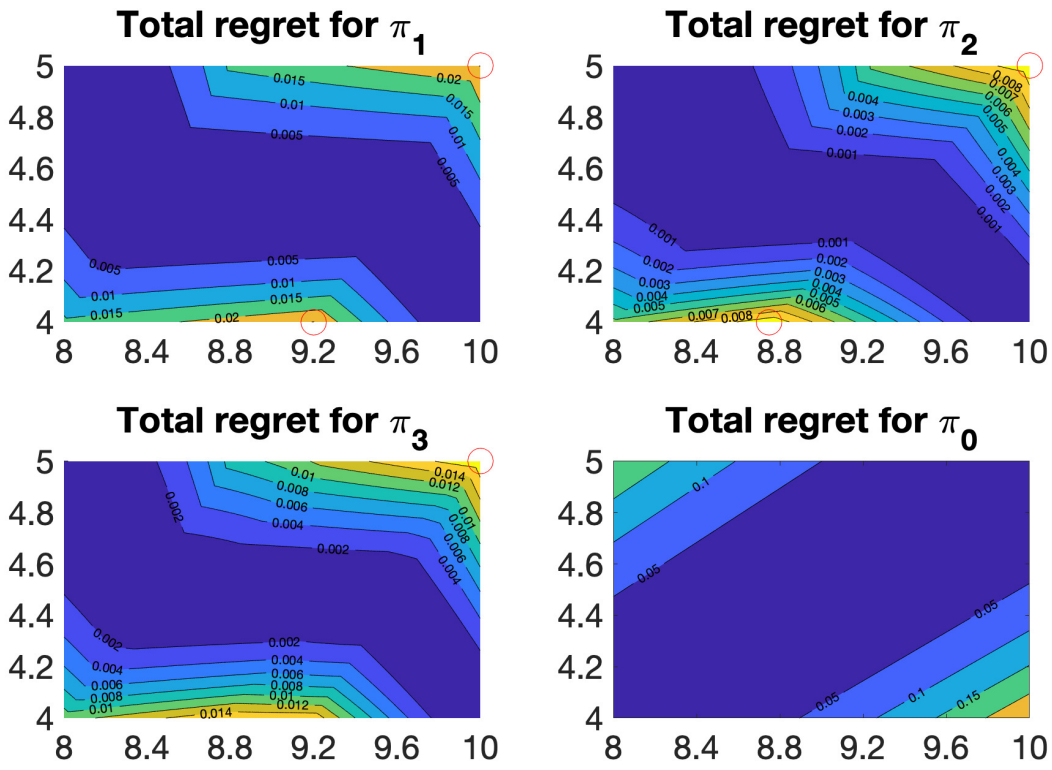

Figure 1: Contour Plots of Regret Function for $\pi_1, \pi_2, \pi_3, \pi_0$ in Example 1

at $\theta_0 = (a_0, b_0)$, and there is a neighborhood Ω of $\theta_0 = (a_0, b_0)$, in which R_2^θ is strictly smaller than R_1^θ. Construct the policy π_3: apply π_1 with probability $1-\delta$, and apply π_2 with probability δ. Denote the maximum of π_1 on Θ by Z_1, the maximum of π_1 on $\Theta \backslash \Omega$ by Z_1', the maximum of π_2 on Θ by Z_2, the maximum of π_3 on Θ by Z_3, the maximum of π_3 on $\Theta \backslash \Omega$ by Z_3'. Since $Z_1' < Z_1$, when $\delta < (Z_1 - Z_1')/Z_2$, we have $Z_3' \leq (1-\delta)Z_1' + \delta Z_2 < Z_1$. Inside Ω, the total regret R_3^θ of π_3 is strictly smaller than R_1^θ. In sum, we have $Z_3 < Z_1$. □

4.1.2 There Might Not Be a Deterministic Optimal Policy.

We present another result that the optimal policy under the worst-case approach might have to be stochastic.

Proposition 3. *For the worst-case approach, there might not exist a deterministic optimal policy that minimizes* $\max_{\theta \in \Theta} \{ \int_{\mathbb{H}} f(\hat{H} \mid \theta) R^\theta_{\pi|\hat{H}} \mathrm{d}\hat{H} \}$.

Proof. We use the following example to prove this proposition.

Example 2. *The setting of this example is almost the same as Example 1, except that the offline price can be chosen by the decision maker.*

An equivalent description is to consider no offline stage and two online periods, but the first online period does not count in the regret.

First consider deterministic policies. We can use brute force to search for the optimal policy. The best choice for the offline stage is to set $\hat{p} = 1.25$. For the online stage, the largest regret 0.0098 corresponds to $\theta = (10, 5)$ and $\theta = (8.75, 4)$, the same as case (2) of Example 1. Denote this policy by π_1.

Consider another policy π_2: for offline stage, sets $\hat{p} = 1$; for online stage, if the offline demand $\hat{D} = a - b\hat{p} + \hat{\epsilon}$ satisfies $\hat{D} < 4.9$, sets $p = 1.09375$ (optimal for $\theta = (8.75, 4)$), otherwise sets $p = 1$ (optimal for $\theta = (10, 5)$). Policy π_2 has almost 0 regret near $\theta = (10, 5)$ and $\theta = (8.75, 4)$.

Now consider another policy π_3 that applies π_1 with probability 0.99 and applies π_2 with probability 0.01. The regret measurement $R = 0.0097$, meaning that π_3 is better than π_1. Since π_1 is the best deterministic policy, the optimal policy has to be stochastic.

□

See Figure 2 for the contour plots of regret function for π_1, π_2 in Example 2. Brighter colors are for larger regrets. The x-axis is parameter a, and the y-axis is parameter b. Red circles are the global maximal points. Policy π_2 behaves well near the worst cases of π_1. Thus a mixture of π_1 and π_2 is better than pure π_1.

In the proof of Proposition 1, we illustrate that the optimal policy should be deterministic for the last period. However, in Example 2, it practically has two periods that can be controlled by the decision maker. Thus the optimal policy can be stochastic.

The regret of π_1 has two global maximal points, so that we cannot naively improve both. However, using such offline data, we can almost guarantee that these two points do not occur simultaneously, and we can improve them separately.

4.2 Optimal Policy for the Bayesian Approach

In this subsection, we consider the Bayesian regret measurement calculated before the observation of concrete offline data. An optimal policy $\bar{\pi}^*$ minimizes

$$\int_\Theta f(\theta)\left[\int_{\mathbb{H}} f(\hat{H} \mid \theta) R^\theta_{\pi|\hat{H}} \mathrm{d}\hat{H}\right]\mathrm{d}\theta$$
$$= \int_{\mathbb{H}} f(\hat{H})\left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta\right]\mathrm{d}\hat{H}$$

We study whether the optimization problem can be decomposed into independent subproblems and whether a deterministic optimal policy always exists.

Proposition 4. *When we decompose the set of possible offline data into $\mathbb{H} = \cup_{i=1}^m \mathbb{H}_i$, if the policy π minimizes*

$$\int_{\mathbb{H}_i} f(\hat{H})\left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta\right]\mathrm{d}\hat{H}$$

for each $\mathbb{H}_i$, then π minimizes

$$\int_{\mathbb{H}=\cup\mathbb{H}_i} f(\hat{H})\left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta\right]\mathrm{d}\hat{H}$$

Besides, if the policy π minimizes

$$\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta$$

for each $\hat{H}$, then π minimizes

$$\int_{\mathbb{H}} f(\hat{H})\left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta\right]\mathrm{d}\hat{H}$$

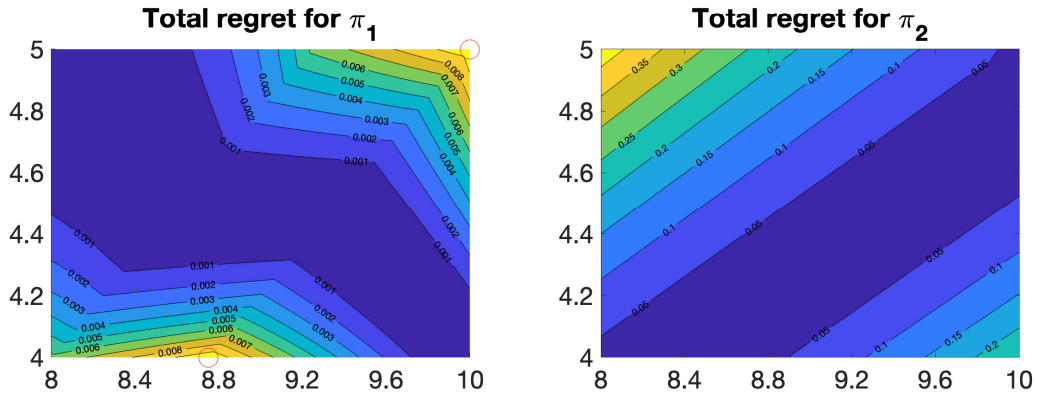

Figure 2: Contour Plots of Regret Function for π_1, π_2 in Example 2

Proof. Just notice that

$$\int_{\mathbb{H}=\cup\mathbb{H}_i} f(\hat{H}) \left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta\right] \mathrm{d}\hat{H}$$
$$= \sum_{i=1}^m \int_{\mathbb{H}_i} f(\hat{H}) \left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta\right] \mathrm{d}\hat{H}$$

A policy that minimizes each term on the right hand side also minimizes the left hand side. Besides, since $f(\hat{H}) \geq 0$, a policy that minimizes $\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta$ for each $\hat{H}$also minimizes $\int_{\mathbb{H}} f(\hat{H}) \left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi|\hat{H}} \mathrm{d}\theta\right] \mathrm{d}\hat{H}$. □

Define $g_i(\theta) = \int_{\mathbb{H}_i} f(\hat{H} \mid \theta) R^\theta_{\pi_1|\hat{H}} \mathrm{d}\hat{H}$, then its Bayesian measurement is its L^1 norm, with prior distribution $f(\theta)$ as the measure: $\int_\Theta f(\theta) g_i(\theta) \mathrm{d}\theta$. Since $g_i(\theta) \geq 0$, $||\sum_i g_i||_1 = \sum_i ||g_i||_1$. Thus optimizing $||\sum_i g_i||_1$ is equivalent to optimizing all $||g_i||_1$. The optimization problem can be decomposed into many independent subproblems according to offline data $\hat{H}$. An improvement for a subproblem, even not reaching optimum, is also an improvement for the global problem. Besides, the distribution of $\hat{H}$ does not affect which policy should be applied if a certain $\hat{H}$ is observed.

Proposition 5. *For the Bayesian approach, there always exists an optimal policy which is deterministic. If there exists a stochastic optimal policy, then it is a combination of some deterministic optimal policies.*

Proof. Consider a policy π_0 that has stochasticity in determining prices of online stage. We can decompose it into a combination of deterministic policies with different probabilities, $\pi_0 = \sum_{i=1}^k P_i \pi_i$, which means applying policy π_i with probability P_i. From the definition of regret, Eq. 1, we have $R^\theta_{\pi_0|\hat{H}} = \sum_{i=1}^k P_i R^\theta_{\pi_i|\hat{H}}$. Furthermore, for the Bayesian measurement, we have

$$\int_{\mathbb{H}} f(\hat{H}) \left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi_0|\hat{H}} \mathrm{d}\theta\right] \mathrm{d}\hat{H}$$
$$= \sum_{i=1}^k P_i \int_{\mathbb{H}} f(\hat{H}) \left[\int_\Theta f(\theta \mid \hat{H}) R^\theta_{\pi_i|\hat{H}} \mathrm{d}\theta\right] \mathrm{d}\hat{H}$$

This means that the overall regret of this stochastic policy is the average overall regret of these decomposed deterministic policies. If a stochastic policy is optimal, then all the deterministic policies in its decomposition are also optimal, in the sense that they render the same minimal regret. Thus we have a deterministic optimal policy. □

For the Bayesian approach, to search for the

optimal policy, we only need to focus on deterministic policies. This is fundamentally different from the worst-case approach.

For the Bayesian approach, we can search for the optimal policy for each $\hat{H}$, which should be deterministic, and combine them to obtain the overall optimal policy. We explicitly determine the optimal policy for Examples 1,2.

For Example 1, assume the prior distribution $f(\theta)$ is uniform on Θ. If $\hat{p} = 0.8$, then the posterior distribution is concentrated on a line segment with slope 1.25 in Θ. Through straightforward calculation, we can determine that for each $\hat{D}$, the online price p should be the optimal price for the middle point of this segment. Therefore, for case (1) with $\hat{p} = 0.8$, the optimal policy π_1 should set $p = \hat{D}/6 + 2/15$ for $\hat{D} < 4.8$, $p = \hat{D}/9 + 2/5$ for $4.8 \leq \hat{D} \leq 6$, and $p = 11\hat{D}/48 - 37/120$ for $\hat{D} > 6.8$. The overall Bayesian measurement is 0.0014. For case (2) with $\hat{p} = 1.25$, similarly, the optimal policy π_2 should set $p = 19\hat{D}/150 + 347/600$ for $\hat{D} < 1.75$, $p = \hat{D}/9 + 5/8$ for $1.75 \leq \hat{D} \leq 3$, and $p = \hat{D}/6 + 5/12$ for $\hat{D} > 3$. The overall Bayesian measurement is 0.0005. Therefore, for case (3) of Example 1, the optimal policy should be applying π_1 when $\hat{p} = 0.8$, and applying π_2 when $\hat{p} = 1.25$. For other cases, whenever $\hat{p} = 0.8$, applying π_1 is always optimal. Notice that for the same scenario, the Bayesian measurement is smaller than the worst-case measurement, since the former averages over all θ, and the latter only counts the worst θ.

For Example 2, assume the prior distribution $f(\theta)$ is uniform on Θ. When the offline $\hat{p}$ is chosen, it is the same as Example 1, and the Bayesian measurement can be calculated simply. We observe that the Bayesian measurement $\bar{R}$ for the online stage decreases with $\hat{p}$: $\bar{R}(\hat{p} = 0.7) = 0.0016$, $\bar{R}(\hat{p} = 0.8) = 0.0014$, $\bar{R}(\hat{p} = 0.9) = 0.0011$, $\bar{R}(\hat{p} = 1) = 0.0009$, $\bar{R}(\hat{p} = 1.1) = 0.0007$, $\bar{R}(\hat{p} = 1.25) = 0.0005$. Since the offline stage does not count in the regret, the optimal policy for Example 2 is to set $\hat{p} = 1.25$ for the offline stage. For the online stage, set $p = 19\hat{D}/150 + 347/600$ for $\hat{D} < 1.75$, $p = \hat{D}/9 + 5/8$ for $1.75 \leq \hat{D} \leq 3$, and $p = \hat{D}/6 + 5/12$ for $\hat{D} > 3$.

To search for the optimal policy under the Bayesian approach, we can greedily search each period backwards. When we have had the optimal policy for the last m periods (for each posterior distribution), we can search the $m+1$ period from the last, to find the optimal policy for the last $m+1$ periods. This is a standard dynamic programming method. However, when the system is complicated enough, especially when the number of online periods is too large, it is not numerically feasible to determine the optimal policy.

Remark 1. *To define the optimal policy under the Bayesian approach, we need two factors: one is the current distribution (belief) of θ, the other is the time left, i.e., the number of remaining online periods. A policy that only considers the distribution of θ, but not the time left, is called "stationary". If we restrict to stationary policies, then there might not exist deterministic optimal policies, different from Proposition 5. In the paper by Russo and Van Roy (2018), they constructed a model (Example 1 in that paper) that any deterministic stationary policy behaves much worse than some stochastic stationary policies, such as the Thompson sampling (Russo*

et al., 2018). Here for each period, the Thompson sampling randomly chooses a possible value of the parameter θ according to the current posterior distribution of θ, and applies the optimal action for this θ. The meaning of considering the time left is simple: if there is almost no time left, it is better to exploit (choosing the most proper prices) according to current knowledge; if there is plenty of time, it is better to explore (testing extreme prices) more to find potential better actions.

Remark 2. *Although the time left is essential in the optimal policy, the current time should not be explicitly considered. Consider a policy that always explores for the first two online periods regardless of offline data, and exploits for the rest periods. Running this policy for two periods, then next we shall only exploit. However, at this point, if we regard the first two periods as offline, and run this policy again, then we need to explore for two more periods. Thus it is not optimal. In fact, this is a necessary condition for optimal policy: when we have run the optimal policy for some periods, restarting this policy and continuing it should be the same. Thompson sampling satisfies this condition, but it is stochastic, thus is not optimal in general.*

5 An Example with Discrete Setting

We have discussed some properties of the optimal policy for online pricing problem with offline data. However, the setting of this problem is very complicated, such that in most cases, the optimal policy is difficult to explicitly determine. We introduce an example with discrete setting, similar to the multi-armed bandit problem. This example is simple enough, so that the optimal policy can be calculated explicitly. We use this example to illustrate different properties of the optimal policies under the worst-case approach and the Bayesian approach. The details of the example can be found in Appendix C.

6 Summary and Discussion

6.1 Summary of Difference Between Two Approaches

Consider the scenario that the decision maker is positioned at the time point when the offline data are already observed and viewed as deterministic. Under the worst-case approach, different concrete offline data do not affect the optimization problem, and it is the same to a scenario without any offline data. In contrast, the Bayesian approach can absorb such concrete data into a posterior distribution, and different offline data correspond to different optimization problems.

Consider the scenario that the decision maker is positioned at the time point before the offline data are generated. The overall optimization problem can be decomposed into different subproblems according to offline data $\hat{H}$. Under the worst-case approach, the globally optimal policy may not be optimal for each subproblem, and the behavior of the globally optimal policy for each subproblem depends on the distribution of $\hat{H}$. Thus we cannot solve the global optimization problem by optimizing each local subproblem. Under the Bayesian approach, the globally

optimal policy is also optimal for each subproblem, regardless of the distribution of the offline data $\hat{H}$.

For the worst-case approach, deterministic optimal policy may not exist. For Bayesian approach, there always exists a deterministic optimal policy.

We note that the Bayesian approach needs to assign a prior distribution on the space of parameters. Besides, it requires that the explicit form of the noise is known (at most with finitely many unknown parameters). The worst-case approach does not need such assumptions.

In summary, the two approaches, Bayesian and worst-case, have different properties regarding different perspectives as described above. This summary may provide information to assist practitioners to make a selection based on their needs.

6.2 Discussion

The pricing problem introduced in this paper is a special case in a class of reinforcement learning problems. The general form is: The system has an unknown parameter vector θ. For each online period, the decision maker chooses an action $a_i \in \mathbb{A}$, and receives a revenue (depends on θ, a_i, and possibly a stochastic noise) from the system. There are some offline data, i.e., action-revenue pairs that contain information of θ. The goal is to minimize the regret, the difference between optimal revenue and actual revenue. Assume that if the offline data $\hat{H}$ can be generated by parameter θ_1, then $\hat{H}$ also can be generated by any other parameter θ_2. Also, as discussed in Remark 1, the decision maker should know the number of online periods left. We assume that offline data and online data are generated from the same system parameters. If not, the optimal policy might differ (Wang et al., 2023).

Although we only consider a simple linear model, most results in this paper work for the general reinforcement learning problem. All the results in Section 3, especially Proposition 1, apply for the general reinforcement learning problem. The results in Section 4.1, Propositions 2,3, are derived on Examples 1,2, which depend on the concrete form of this linear model. Nevertheless, Proposition 2 is based on a property of the L^∞ norm: $||f_i||_\infty \leq ||g_i||_\infty$ for $i = 1, \ldots, m$ does not imply $||\sum f_i||_\infty \geq ||\sum g_i||_\infty$. Besides, Lemma 1 holds in general. Thus for a general reinforcement learning model, we believe that Proposition 2 still holds. When the general reinforcement learning model is not too simple, the argument of Proposition 3 should also hold. The results in Section 4.2, Propositions 4,5, do not depend on the specific linear model, and hold for general reinforcement learning models.

In many situations, the system is too complicated that the optimal policy is difficult to determine, especially when there are many online periods. Therefore, researchers often propose methods that are easy to compute and have near-optimal performance in the asymptotic sense, such as the upper confidence bound algorithm (Srinivas et al., 2012) and Thompson sampling (Russo et al., 2018). Although we mainly discuss the optimal policy, the worst-case approach and the Bayesian approach are different in searching for a near-optimal policy: for the Bayesian approach, an improvement on a specific $\hat{H}$ or $\mathbb{H}_i$, even not reaching optimum, always benefits the

overall regret (Proposition 4). This means we can greedily optimize each subproblem. For the worst-case approach, since we only care about the worst case, an improvement for other cases might not affect the measurement, and sometimes might even harm the measurement (Proposition 2). Besides, the worst-case measurement has a property: after observing concrete offline data, the information in the offline data does not help with reducing the regret (Proposition 1). If a researcher wants to apply information theoretical or causal tools in this scenario (Wang and Wang, 2020), then the worst-case approach might not be proper.

For a purely online pricing problem, if we have run a policy for several online periods, then we can regard the past as new offline data, and take this as an online pricing problem with concrete offline data. Therefore, our discussion about the incorporation of concrete offline data under different approaches is also related to pure online pricing problems.

A common method for evaluating policies is to let the number of periods $T \to \infty$, and consider the growth order of the total expected regret. Assume θ^* has the largest growth order $g(T)$, and for any $\epsilon > 0$, θ^* has a neighborhood Θ_ϵ^* that $\forall \theta \in \Theta_\epsilon^*$ has growth order at least $g^{1-\epsilon}(T)$. Then the worst-case measurement is $g(T)$. The Bayesian measurement is between $\int_{\Theta_\epsilon^*} f(\theta) g^{1-\epsilon}(T) \mathrm{d}\theta$ and $\int_\Theta f(\theta) g(T) \mathrm{d}\theta$, and the order is between $g^{1-\epsilon}(T)$ and $g(T)$. Therefore, when we consider the limiting behavior, two approaches are almost the same.

The total regret is the difference between the maximal revenue and the actual revenue for a policy. Under the worst-case approach, denote the maximal revenue function on Θ by $F(\theta)$, and denote the total revenue of policy π by $G^\pi(\theta)$, then the regret function is $F(\theta) - G^\pi(\theta)$. Since $F(\theta)$ is not a constant function, generally we do not have

$$\arg\min_\theta [G^\pi(\theta)] = \arg\max_{\theta \in \Theta} [F(\theta) - G^\pi(\theta)]$$

Also, the optimal $\pi \in \Pi$ that minimizes $\max_{\theta \in \Theta} [F(\theta) - G^\pi(\theta)]$ might not maximize $\min_\theta [G^\pi(\theta)]$. Therefore, under the worst-case approach, minimizing the total regret and maximizing the total revenue are not equivalent. Under the Bayesian approach, the maximal revenue is a fixed number, thus the optimal policy that minimizes the total regret also maximizes the total revenue.

A The Existence of an Optimal Policy

We prove the existence of an optimal policy for both approaches, with or without offline data. First, we introduce the Lévy-Prokhorov metric (Prokhorov, 1956). Denote the set of all Borel sets in $[p_{\min}, p_{\max}]$ by $\mathcal{B}$. For $A \in \mathcal{B}$ and $\epsilon > 0$, define

$$A^\epsilon = \{p \in [p_{\min}, p_{\max}] \mid \exists q \in A,\ \mathrm{d}(p, q) < \epsilon\}$$

where $\mathrm{d}(p, q)$ is the Euclidean metric. Then on the space of probability measures on $[p_{\min}, p_{\max}]$, we can define the Lévy-Prokhorov metric:

$$\mathrm{D}_{\mathrm{LP}}(\mu, \nu) = \inf\{\epsilon > 0 \mid \mu(A) \le \nu(A^\epsilon) + \epsilon \text{ and } \nu(A) \le \mu(A^\epsilon) + \epsilon,\ \forall A \in \mathcal{B}\}$$

For the worst-case approach, we need to consider all policies, whether deterministic or

stochastic. For simplicity, assume we have concrete offline data $\hat{H}$, and the regret measurement is $R^{\theta}_{\pi|\hat{H}}$. Since $R^{\theta}_{\pi|\hat{H}}$ is bounded below by 0, $R_0 = \inf_\pi R^{\theta}_{\pi|\hat{H}}$ is finite. We can find a sequence of policies $\pi_1, \ldots, \pi_n, \ldots$ with $R^{\theta}_{\pi_n|\hat{H}} \to R_0$. The question is to find a policy π^* with $R^{\theta}_{\pi^*|\hat{H}} = R_0$.

(1) For each period, a (stochastic) policy is a probability measure on $[p_{\min}, p_{\max}]$. By a corollary of the Prokhorov's theorem (Durrett, 2019), a sequence of probability measures $\pi_1, \ldots, \pi_n, \ldots$ on a compact subset of $\mathbb{R}^n$ has a subsequence that weakly converges to a probability measure π^*. (2) Consider the Lévy-Prokhorov metric for probability measures on $[p_{\min}, p_{\max}]$. In this case, the convergence of probability measures under the Lévy-Prokhorov metric is equivalent to weak convergence (Billingsley, 2013). Thus a subsequence of $\pi_1, \ldots, \pi_n, \ldots$ converges to π^* under the Lévy-Prokhorov metric. (3) The worst-case measurement is a continuous function for probability measures on $[p_{\min}, p_{\max}]$ under the Lévy-Prokhorov metric, since the range of θ, price, and demand are all compact in $\mathbb{R}^n$. Thus a subsequence of $R^{\theta}_{\pi_n|\hat{H}}$ converges to $R^{\theta}_{\pi^*|\hat{H}}$, which is R_0.

For the Bayesian approach, similarly, we can find an optimal policy π^* in the set of deterministic policies. The convergence argument can be much simplified. As illustrated in Proposition 5, we cannot find a stochastic policy with a smaller regret than π^*. Thus π^* is an optimal policy in the set of any policies.

B A Discrete Example Illustrating the Role of Offline Data

We provide an example to illustrate why the knowledge of concrete offline data does not change the problem under the worst-case approach, compared to the problem without any offline data.

Example 3. *There are two possible parameters θ_1, θ_2, two possible sets of offline data $\hat{H}_1, \hat{H}_2$, and two possible basic policies π_1, π_2. $\mathbb{P}(\hat{H}_1 \mid \theta_1) = \mathbb{P}(\hat{H}_2 \mid \theta_2) = 0.9$, and $\mathbb{P}(\hat{H}_1 \mid \theta_2) = \mathbb{P}(\hat{H}_2 \mid \theta_1) = 0.1$, meaning that for any $i = 1, 2$, observing $\hat{H}_i$ is more likely under parameter θ_i. The regret of a policy π only depends on the true parameter θ, not on the offline data $\hat{H}$: $R^{\theta_1}_{\pi_1} = R^{\theta_2}_{\pi_2} = 0$, and $R^{\theta_1}_{\pi_2} = R^{\theta_2}_{\pi_1} = 1$. Thus for any $i = 1, 2$, the policy π_i performs better under parameter θ_i.*

Under the worst-case approach, if there is no offline data, the regret measurement of pure π_1 and pure π_2 are both 1, and the optimal policy is a mixture of half π_1 half π_2, with regret measurement 0.5. If the concrete $\hat{H}_1$ is observed, then θ_1 is more possible. However, if the decision maker is induced to apply π_1 more often, although the performance on θ_1 is better, the regret of the less likely θ_2 is sacrificed. Since this approach directly considers the regret of the worst case, without considering its probability density, applying π_1 more often would make the regret measurement increase. Thus the optimal policy is still "applying π_1 with probability 0.5; applying π_2 with probability 0.5". Under the Bayesian approach, since the posterior probabil-

ity density is concentrated on θ_1, the advantage of π_1 on θ_1 can overcome the disadvantage of π_1 on θ_2.

When we evaluate the policies before the observation of concrete $\hat{H}$, the regret $R^{\theta}_{\pi|\hat{H}}$ is averaged over all possible $\hat{H}$. Therefore, "applying π_1 if $\hat{H}_1$ is observed; applying π_2 if $\hat{H}_2$ is observed" is the optimal policy even under the worst-case approach, with regret measurement 0.1. If the true θ is θ_1, the we are more likely to observe $\hat{H}_1$, and then more likely to apply π_1. The risk of "applying π_2 when $\hat{H}_2$ is observed, although the true parameter is θ_1" ($\mathbb{P} = 0.1$) is covered by the advantage of "applying π_1 when $\hat{H}_1$ is observed, and the true parameter is θ_1" ($\mathbb{P} = 0.9$). In total, this policy improves the expected (with respect to $\hat{H}$) performance for each θ. Specifically, under the worst-case approach, the optimal policy before the observation of concrete $\hat{H}$ is different from the optimal policy after the observation of concrete $\hat{H}$.

C A Discrete Example with Complicated Optimal Policies

We have discussed some properties of the optimal policy for online pricing problem with offline data. However, the setting of this problem is very complicated, such that in most cases, the optimal policy is difficult to determine. In this section, we introduce an example with discrete settings, similar to the multi-armed bandit problem. This example is simple enough, so that the optimal policy can be calculated explicitly. We use this example to illustrate different behaviors of the optimal policies under the worst-case approach and the Bayesian approach.

Example 4. *There are two possible parameters θ_1, θ_2, two possible sets of offline data $\hat{H}_1, \hat{H}_2$, and two possible basic policies π_1, π_2. Two probabilities Q_1, Q_2 are used to describe the generation of offline data. With given θ and $\hat{H}$, each policy has a corresponding total expected regret $R(\pi \mid \theta, \hat{H})$. See Table 1 for model details.*

For this problem, all the possible policies are the combinations of π_1, π_2: when $\hat{H}_1$ is observed, apply π_1 with probability P_1, and apply π_2 with probability $1 - P_1$; when $\hat{H}_2$ is observed, apply π_1 with probability P_2, and apply π_2 with probability $1 - P_2$.

For the worst-case approach, if only $\hat{H}_1$ is observed, then π_1 is better than π_2; if only $\hat{H}_2$ is observed, then π_2 is better than π_1. This implies that the optimal policy should be: apply π_1 if $\hat{H}_1$ is observed, and apply π_2 if $\hat{H}_2$ is observed. We shall see that the truth is much more complicated.

For different values of Q_1, Q_2, the optimal policies for the worst-case approach can be calculated directly but tediously, by minimizing

$$\begin{aligned}\max\{&2 + 2Q_1 - Q_1P_1 - (1 - Q_1)P_2,\\ &3 - 2Q_2 + Q_2P_1 + (1 - Q_2)P_2\}\end{aligned}$$

The optimal policies in all cases are presented in Table 2. In this table, uniqueness describes whether the optimal policy is unique. If not, all the optimal policies lie on the line segment with the given ends.

Under the worst-case approach, the optimal policies are various in different cases. For $Q_1 = Q_2 = 0.8$, the optimal policy is deterministic,

	θ_1	θ_2
$\hat{H}_1$	$R(\pi_1 \mid \theta_1, \hat{H}_1) = 3$	$R(\pi_1 \mid \theta_2, \hat{H}_1) = 2$
	$R(\pi_2 \mid \theta_1, \hat{H}_1) = 4$	$R(\pi_2 \mid \theta_2, \hat{H}_1) = 1$
	$\mathbb{P}(\hat{H}_1 \mid \theta_1) = Q_1$	$\mathbb{P}(\hat{H}_1 \mid \theta_2) = Q_2$
$\hat{H}_2$	$R(\pi_1 \mid \theta_1, \hat{H}_2) = 1$	$R(\pi_1 \mid \theta_2, \hat{H}_2) = 4$
	$R(\pi_2 \mid \theta_1, \hat{H}_2) = 2$	$R(\pi_2 \mid \theta_2, \hat{H}_2) = 3$
	$\mathbb{P}(\hat{H}_2 \mid \theta_1) = 1 - Q_1$	$\mathbb{P}(\hat{H}_2 \mid \theta_2) = 1 - Q_2$

Table 1: Details of Example 4

Conditions		Uniqueness	Optimal (P_1, P_2)
$Q_1 + Q_2 \geq 1.5$		Unique	$(1, 1)$
$1 < Q_1 + Q_2 < 1.5$	$Q_1 > Q_2$	Unique	$(1, \frac{Q_1+Q_2-1}{2-Q_1-Q_2})$
	$Q_1 < Q_2$	Unique	$(\frac{3Q_1+3Q_2-3}{Q_1+Q_2}, 1)$
	$Q_1 = Q_2$	Multiple	$(1, \frac{Q_1+Q_2-1}{2-Q_1-Q_2})$ to $(\frac{3Q_1+3Q_2-3}{Q_1+Q_2}, 1)$
$Q_1 + Q_2 = 1$	$Q_1 > Q_2$	Unique	$(1, 0)$
	$Q_1 < Q_2$	Unique	$(0, 1)$
	$Q_1 = Q_2$	Multiple	$(1, 0)$ to $(0, 1)$
$0.5 < Q_1 + Q_2 < 1$	$Q_1 > Q_2$	Unique	$(\frac{2Q_1+2Q_2-1}{Q_1+Q_2}, 0)$
	$Q_1 < Q_2$	Unique	$(0, \frac{2Q_1+2Q_2-1}{2-Q_1-Q_2})$
	$Q_1 = Q_2$	Multiple	$(\frac{2Q_1+2Q_2-1}{Q_1+Q_2}, 0)$ to $(0, \frac{2Q_1+2Q_2-1}{2-Q_1-Q_2})$
$Q_1 + Q_2 \leq 0.5$		Unique	$(0, 0)$

Table 2: The Optimal Policy of Example 4 for the Worst-case Approach

with $P_1 = 1, P_2 = 1$. For $Q_1 = 0.8, Q_2 = 0.4$, the optimal policy is stochastic, with $P_1 = 1, P_2 = 0.25$. For $Q_1 = Q_2 = 0.6$, there are infinite many optimal policies, all stochastic, satisfying $0.6P_1 + 0.4P_2 = 0.7$.

We have introduced a policy: applying π_1 if $\hat{H}_1$ is observed, and applying π_2 if $\hat{H}_2$ is observed, which corresponds to $P_1 = 1, P_2 = 0$. Although this is the combination of the optimal policy for each $\hat{H}$, it is optimal only when $Q_1 + Q_2 = 1$, $Q_1 \geq Q_2$.

For the Bayesian approach, the optimal policy is very simple. Assume the prior distribution on Θ is $\mathbb{P}(\theta_1) = \mathbb{P}(\theta_2) = 0.5$, then the measurement of regret is

$$2.5 + Q_1 - Q_2 + (-0.5Q_1 + 0.5Q_2)P_1 \\ +(0.5Q_1 - 0.5Q_2)P_2$$

When $Q_1 > Q_2$, the optimal policy is $P_1 = 1, P_2 = 0$. When $Q_1 < Q_2$, the optimal policy is $P_1 = 0, P_2 = 1$. When $Q_1 = Q_2$, all policies have the same total regret, thus are all optimal. We can see that there is always a deterministic optimal policy, and any stochastic optimal policy (if exists) is a combination of deterministic optimal policies.

Acknowledgement. The authors would like to thank the anonymous referees for providing helpful comments that improve the quality of this paper.

References

Ban G-Y, Keskin N B (2021). Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Manag. Sci.*, 67(9):5549–5568.

Bastani H, Simchi-Levi D, Zhu R (2022). Meta dynamic pricing: Transfer learning across experiments. *Manag. Sci.*, 68(3):1865–1881.

Billingsley P (2013). *Convergence of Probability Measures*. John Wiley & Sons.

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

den Boer A V (2015). Dynamic pricing and learning: historical origins, current research, and new directions. *Surv. Oper. Res. Manag. Sci.*, 20(1):1–18.

den Boer A V, Zwart B (2015). Dynamic pricing and learning with finite inventories. *Oper. Res.*, 63(4):965–978.

Durrett R (2019). *Probability: Theory and Examples*. Cambridge University Press.

Eysenbach B, Salakhutdinov R R, Levine S (2019). Search on the replay buffer: Bridging planning and reinforcement learning. *Adv. Neural Inf. Process Syst.*, 32.

Fujimoto S, Meger D, Precup D (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR.

Gallego G, Topaloglu H (2019). *Revenue Management and Pricing Analytics*. Springer, New York.

Harrison J M, Keskin N B, Zeevi A (2012). Bayesian dynamic pricing policies: Learning

and earning under a binary prior distribution. *Manag. Sci.*, 58(3):570–586.

Keskin N B, Zeevi A (2014). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.*, 62(5):1142–1167.

Kirschner J, Krause A (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference on Learning Theory*, pages 358–384. PMLR.

Munos R, Stepleton T, Harutyunyan A, Bellemare M (2016). Safe and efficient off-policy reinforcement learning. *Adv. Neural Inf. Process Syst.*, 29.

Prokhorov Y V (1956). Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.*, 1(2):157–214.

Rakelly K, Zhou A, Finn C, Levine S, Quillen D (2019). Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International Conference on Machine Learning*, pages 5331–5340. PMLR.

Rolnick D, Ahuja A, Schwarz J, Lillicrap T, Wayne G (2019). Experience replay for continual learning. *Adv. Neural Inf. Process Syst.*, 32.

Russo D, Van Roy B (2014). Learning to optimize via posterior sampling. *Math. Oper. Res.*, 39(4):1221–1243.

Russo D, Van Roy B (2018). Learning to optimize via information-directed sampling. *Oper. Res.*, 66(1):230–252.

Russo D J, Van Roy B, Kazerouni A, Osband I, Wen Z (2018). A tutorial on Thompson sampling. *Found. Trends Mach. Learn.*, 11(1):1–96.

Srinivas N, Krause A, Kakade S M, Seeger M W (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory*, 58(5):3250–3265.

Thomas P, Brunskill E (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR.

Wang Y, Wang L (2020). Causal inference in degenerate systems: An impossibility result. In *International Conference on Artificial Intelligence and Statistics*, pages 3383–3392. PMLR.

Wang Y, Zheng Z, Shen Z-J M (2023). Online pricing with polluted offline data. *Available at SSRN 4320324*.

Zanette A, Brandfonbrener D, Brunskill E, Pirotta M, Lazaric A (2020). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR.

Dr. Yue Wang is a postdoctoral fellow at the Department of Computational Medicine, University of California, Los Angeles since 2021. During 2018-2021, Dr. Wang was a postdoctoral researcher at Institut des Hautes Études Scientifiques in France. Dr. Wang received Ph.D. in applied mathematics from the University of Washington in 2018, and B.Sc. in mathematics

from Peking University in 2013. Dr. Wang applies different mathematical tools, such as modeling, simulation, algorithm, statistical analysis, theoretical analysis with discrete mathematics, differential equation, and stochastic process, to biology, e.g., population dynamics, gene regulation, and developmental biology. Dr. Wang also applies probability, stochastic process, and discrete mathematics to different subjects, such as reinforcement learning, causal inference, statistical physics, biochemistry, dynamical system, and law.

Dr. Zeyu Zheng is an assistant professor at the Department of Industrial Engineering and Operations Research, University of California, Berkeley since 2018. Dr. Zheng received a PhD degree in Operations Research from Stanford University in 2018, an MS degree in Economics from Stanford University in 2016 and a Bachelor degree in Mathematics from Peking University in 2012. He has done research in Monte Carlo simulation theory and simulation optimization. He is also interested in non-stationary stochastic modeling.