Appendix A: Better Regret with Known Price Coefficient

In the main paper, all the regret bounds are on the order of $O(\sqrt{T})$. Here we consider a setting of known price coefficient, i.e., β^* in (1) is known and show that a $O(\log T)$ regret is achievable under such a setting. This means that the knowledge of the price coefficient is critical in determining the regret order. The algorithm and analysis presented in the following are mainly for two purposes: (i) to identify a difference between the dynamic pricing problem and the bandits problem; (ii) to explain the achievability of $O(\log T)$ regret dependency in the literature. In the rest of this subsection, we denote $p^*((\alpha,\beta),x) = \arg\max_{p\in[\underline{p},\bar{p}]} r(p; x^\top\alpha, x^\top\beta)$ as the optimal price under the covariates $x\in\mathcal{X}$ and the parameter (α,β).

Assumption 7 **(Known β^* and smoothness)**. *Assume β^* in (1) is known. In addition, assume there exists a constant C such that the optimal expected revenue function satisfies*

$$\left|r^*(x^\top\alpha^*, x^\top\beta^*) - r\left(p^*(\theta,x); x^\top\alpha^*, x^\top\beta^*\right)\right| \le C(x^\top\alpha^* - x^\top\alpha)^2,$$

for all $x\in\mathcal{X}$, $\theta,\theta^*\in\Theta$ with $\theta=(\alpha,\beta^*)$ and $\theta^*=(\alpha^*,\beta^*)$.

To interpret the condition in the assumption, the left-hand-side represents the revenue loss caused by using a wrong parameter θ for pricing, while the right-hand-side is quadratic in terms of the linear estimation error. One sufficient condition for the assumption is that $r(p; x^\top\alpha^*, x^\top\beta^*)$ is continuously twice differentiable with respect to p for all possible θ^* and x, and $p^*(\theta,x)$ is Lipschitz in $x^\top\alpha$. Essentially, this condition does not impose extra restriction upon Assumptions 1, 2, and 3; in other words, almost all the demand models that satisfy the previous assumptions also meet this condition under the knowledge of β^*. For example, this condition can be met by the binary demand model (18) with a log-concave unknown noise (Javanmard and Nazerzadeh 2019) and by the linear demand model (19). It is also analogous to the "well separation" condition in the covariate-free case (Broder and Rusmevichientong 2012).

To proceed with the algorithm description, we first slightly revise the MLE estimator in Section 3.1 for known β^*. Specifically, we redefine the misspecified likelihood function for the case of known β^* as

$$\tilde{l}_t(\alpha) := -\int_{D_t}^{g(x_t^\top\alpha + x_t^\top\beta^*\cdot p_t)} \frac{1}{h(u)}(u - D_t)\mathrm{d}u,$$

where the function $h(u)$ is the same as in Section 3.1. Then the estimator becomes

$$\hat{\alpha}_t := \arg\max_{\alpha\in\Theta_\alpha} -\frac{\lambda\underline{g}\|\alpha\|_2^2}{2} + \sum_{\tau=1}^{t}\tilde{l}_\tau(\alpha), \tag{14}$$

where Θ_α denotes the subspace $\{\alpha : (\alpha,\beta^*)\in\Theta\}$. Compared to the previous case of unknown β^*, the only change made here is to plug in the known value of β^* and to restrict the attention to

estimating the unknown α^*. Accordingly, we revise the definition of the (cumulative) design matrix as

$$\tilde{M}_t := \lambda I_d + \sum_{\tau=1}^{t} x_\tau x_\tau^\top$$

with I_d as an identity matrix of dimension d.

The following result is parallel to Corollary 1. We omit the proof as it is the same as the previous case of unknown β^* except for some minor notation changes.

COROLLARY 2. *For all* $\lambda > 0$,

$$\mathbb{P}\left(\exists t \in \{1,...,T\} : \|\hat{\alpha}_t - \alpha^*\|_{\tilde{M}_t} \geq 2\sqrt{\lambda}\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log(T) + d\log\left(\frac{d\lambda+T}{d\lambda}\right)}\right) \leq \frac{1}{T}.$$

Algorithm 4 describes a certainty-equivalent pricing policy. At each time step, the algorithm performs a regularized quasi-MLE to obtain the estimator $\hat{\alpha}_{t-1}$. Then it assumes $\hat{\alpha}_{t-1}$ to be the true parameter and finds the corresponding optimal price.

Algorithm 4 Certainty-Equivalent Pricing

Input: Regularization parameter λ.

for $t = 1,...,T$ **do**

Compute the estimator $\hat{\alpha}_{t-1}$ by (14), observe feature x_t and set the price by

$$p_t = p^*\left(\hat{\theta}_{t-1}, x_t\right)$$

where $\hat{\theta}_{t-1} = (\hat{\alpha}_{t-1}, \beta^*)$.

end for

THEOREM 6. *Under Assumptions 1, 2, 3, 7 and with any sequence* $\{x_t\}_{t=1,...,T}$, *if we choose the regularization parameter* $\lambda = 1$, *the regret of Algorithm 4 is upper bounded by*

$$2C\bar{\gamma}'^2 d\log\left(\frac{d+T}{d}\right) + \bar{p}\bar{D} = \tilde{O}(d^2\log^2 T)$$

where $\bar{\gamma}' = 2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log T + d\log\left(\frac{d+T}{d}\right)}$ *and* C *is defined in Assumption 7.*

Theorem 6 provides a regret upper bound for Algorithm 4. We remark that it is unnecessary for Algorithm 4 to compute the estimator at every time step. Javanmard and Nazerzadeh (2019) solve an L_1 regularized linear regression on geometric time intervals, and the scheme can also be applied to Algorithm 4 with the same order of regret bound. The frequent or infrequent estimation scheme makes no analytical difference and the choice mainly accounts for computation consideration. Xu

and Wang (2021) study the special case of the binary demand model with unit price coefficient (i.e., known β^*), and they derive the same order of regret bound as Theorem 6 under arbitrary covariates using online Newton's method. The intuition is that the convergence rate of Newton's method is on the same order with the MLE estimator, so the corresponding output can be viewed as an approximate MLE estimator at each time step, and the approximation will not deteriorate the regret performance. This is aligned with Theorem 9 and Theorem 10 where an approximate quasi-MLE estimator is used.

In general, many existing $o(\sqrt{T})$ regret bounds (Broder and Rusmevichientong 2012, Javanmard 2017, Javanmard and Nazerzadeh 2019, Xu and Wang 2021) fall into this paradigm of

$$\text{known price coefficient} \;+\; \text{certainty-equivalent policy}.$$

Intuitively, when the price coefficient is known, the price p_t will not interfere the learning of α^*. Thus there is no need to do price exploration like UCB or TS, and the regret purely reflects the cumulative learning rate of α^*. This disentanglement of pricing decisions from parameter estimation makes the setting of known β^* analogous to the "full information" setting in online learning literature. In contrast, when the price coefficient is unknown, the pricing decisions will affect the learning rate of β^*, thus the setting of unknown β^* is more aligned with the "partial information" setting such as the bandits problem.

Interpreting the result under a linear demand model.

We use a linear demand model to further illustrate the contrast between $O(\sqrt{T})$ and $O(\log T)$ regret dependency. Consider the demand follows

$$D_t = a^* + b^* p_t + \epsilon_t$$

for some $a^* > 0$ and $b^* < 0$. At time t, the seller sets the price by $p_t = -\frac{\hat{a}_t}{2\hat{b}_t}$ for some estimators $\hat{a}_t$ and $\hat{b}_t$ which could be from either Algorithm 1 (optimistic estimators) or Algorithm 4 (CE estimators). Then the single step regret can be expressed by a function of the true parameters and the estimators,

$$\text{Reg}_t = r^*(a^*, b^*) - r(p_t; a^*, b^*) = -2b^* \left(\frac{\hat{a}_t}{2\hat{b}_t} - \frac{a^*}{2b^*} \right)^2 = -\frac{(a^*\hat{b}_t - \hat{a}_t b^*)^2}{2b^*\hat{b}_t^2} \tag{15}$$

- When the price coefficient is known, $\hat{b}_t = b^*$. The equality becomes

$$\text{Reg}_t = -\frac{1}{2b^*}(\hat{a}_t - a^*)^2.$$

- When the price coefficient is unknown, we cannot do more than a first-order Talyor expansion when we want to upper bound Reg_t by the estimation error, i.e.,

$$\text{Reg}_t \le c(|\hat{a}_t - a^*| + |\hat{b}_t - b^*|) \tag{16}$$

for some $c > 0$.

For this example, whether the price coefficient b^* is known determines the space in which we view the right-hand-side of (15) as a function of $\hat{a}_t$ and $\hat{b}_t$. Intuitively, suppose that the estimation error is on the order of $\sqrt{1/t}$ (the intuition is precise when the covariates are i.i.d.). Then the right-hand-side will recover two different regret bounds under the two settings.

Generally, we remark that the first-order bound like (16) under a proper norm is always the first step for the analysis of UCB and TS algorithms, including our analysis for the dynamic pricing problem. For linear bandits problem, the LinUCB algorithm (Chu et al. 2011, Abbasi-Yadkori et al. 2011) can directly obtain this first-order bound and under TS algorithms, it can be obtained by some Bayesian arguments (Russo and Van Roy 2014) or by maintaining a constant probability of choosing an optimistic action with anti-concentration sampling (Abeille and Lazaric 2017).

Proof of Theorem 6.

Proof. Denote $\bar{\gamma}' = 2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log T + d\log\left(\frac{d+T}{d}\right)}$. Revise the definition of the "good event" as

$$\tilde{\mathcal{E}} := \left\{ \|\hat{\alpha}_t - \alpha^*\|_{\tilde{M}_t} \le \bar{\gamma}' \text{ for } t = 0, ..., T-1 \right\},$$

From Corollary 2, we know

$$\mathbb{P}(\tilde{\mathcal{E}}) \ge 1 - 1/T.$$

Under the event $\tilde{\mathcal{E}}$, the single period regret can be bounded by

$$\begin{aligned}
\text{Reg}_t &= r^*(x_t^\top \alpha^*, x_t^\top \beta^*) - r(p_t; x_t^\top \alpha^*, x_t^\top \beta^*) \\
&\le C \left| x_t^\top \alpha^* - x_t^\top \hat{\alpha}_{t-1} \right|^2 \\
&\le C \|x_t\|^2_{\tilde{M}_{t-1}^{-1}} \|\hat{\alpha}_{t-1} - \alpha^*\|^2_{\tilde{M}_{t-1}} \\
&\le C\bar{\gamma}'^2 \|x_t\|^2_{\tilde{M}_{t-1}^{-1}}.
\end{aligned}$$

Here the second line is from Assumption 7, the third line is from Holder's inequality, and the last inequality is by the event $\tilde{\mathcal{E}}$.

Thus, the total expected regret (the expectation is with respect to the randomness of demand shocks) can be bounded by

$$\begin{aligned}
\text{Reg}_T^{\pi_{\text{CE}}}(x_1, ..., x_T) &= \sum_{t=1}^{T} \mathbb{E}\left[\text{Reg}_t \cdot \mathbb{1}_{\tilde{\mathcal{E}}}\right] + \mathbb{E}\left[\text{Reg}_t \cdot \mathbb{1}_{\tilde{\mathcal{E}}^c}\right] \\
&\le \sum_{t=1}^{T} C\bar{\gamma}'^2 \|x_t\|^2_{\tilde{M}_{t-1}^{-1}} + \mathbb{P}\left(\tilde{\mathcal{E}}^c\right) \cdot \bar{p}T\bar{D} \\
&\le 2C\bar{\gamma}'^2 d\log\left(\frac{d+T}{d}\right) + \bar{p}\bar{D}
\end{aligned}$$

where π_{CE} denotes Algorithm 4 and $\tilde{\mathcal{E}}^c$ denotes the complement of the event $\tilde{\mathcal{E}}$. The last inequality is because of Lemma 4. □

Appendix B: More Discussions on the Existing Literature

Here we discuss some modeling issues and assumptions in the existing dynamic pricing literature and also show how the general demand model (1) captures the binary demand model and the linear demand model.

B.1. Coefficient of Price and the Distribution of the Random Shock

We first point out an issue with the assumptions on the stochastic quasi-linear demand model or the so-called *binary demand* model. The binary demand model is stated as follows. At time t, the demand is

$$D_t = \begin{cases} 1, & \text{if } x_t^\top \alpha^* + \zeta_t \geq p_t \\ 0, & \text{if } x_t^\top \alpha^* + \zeta_t < p_t \end{cases} \tag{17}$$

where $x_t \in \mathbb{R}^d$ denotes the covariate vector, $\alpha^* \in \mathbb{R}^d$ is the linear coefficient vector, p_t is the price, and ζ_t models an unobserved utility shock. Under this model, $x_t^\top \alpha^* + \zeta_t$ represents the customer's utility where the term $x_t^\top \alpha^*$ captures the part of the utility explained by the covariates x_t.

An assumption often made is (i) the coefficient of p_t is fixed at 1 and (ii) ζ_t's are i.i.d. and follow a distribution with known parameters. The vector α^* then is assumed unknown that we wish to learn from observations over time. This is roughly the setting used in Javanmard (2017), Cohen et al. (2020), Xu and Wang (2021) and part of Javanmard and Nazerzadeh (2019).

There are two issues in general with this setting:

1. Estimation of the parameters of the distribution generating ζ_t cannot be separated from the estimation of the α^*. So it raises the question of how one arrives at this knowledge of the parameters of the distribution of ζ_t's.

2. Somewhat related to the above point – say the seller starts collecting new data covariates $y_t \in \mathbb{R}^m$ in addition to x_t. The additional covariates y_t can potentially improve the prediction of customer utility and thus reduce the variance of the demand shock ζ_t. Assuming the variance is known will then be problematic.

B.1.1. Parametric Distribution of the Error Term Assume the distribution of ζ_t belongs to a parametric family of distributions with some unknown parameter(s). Say ζ_t has zero mean and is scale-invariant with variance σ^{-2} for some unknown $\sigma > 0$.

Scaling appropriately, model (17) is then equivalent to the following binary demand model where the distribution of ζ_t is known but the price coefficient σ is unknown,

$$D_t = \begin{cases} 1, & \text{if } x_t^\top \alpha^* + \tilde{\zeta}_t \geq \sigma p_t \\ 0, & \text{if } x_t^\top \alpha^* + \tilde{\zeta}_t < \sigma p_t \end{cases} \tag{18}$$

where $\tilde{\zeta}_t$ follows some known distribution with mean zero and variance one.

So in (18) we can assume either (i) known price coefficient (say fixed to 1) and unknown variance, or (ii) unknown price coefficient and known variance (say 1). We note that it is unnecessary to assume both parts unknown because the thresholding conditions in (17) and (18) are scale-invariant.

REMARK 1. In the literature, Javanmard and Nazerzadeh (2019) and Luo et al. (2021) consider the model (18). Specifically, Javanmard and Nazerzadeh (2019) focus on the high-dimensional setting and impose the i.i.d. assumption on the covariates; Luo et al. (2021) allow a non-parametric structure for the shock distribution (more general than (18)) and achieve an $\tilde{O}(dT^{2/3})$ regret. Xu and Wang (2021) also mention the model (18) and leave the achievability of $\sqrt{T}$ regret as an open question. The generalized linear model replaces the price coefficient σ in (18) with a linear function of the covariates x_t. Thus it can be viewed as a further generalization of the model (18). The rationale is that the extent to which the covariates x_t explain the customer utility can be dependent not only on some unknown constant σ but also on x_t, and the dependency is unknown. Our result resolves the open question in Xu and Wang (2021).

Linear demand model. Another prevalent demand model (den Boer and Zwart (2014), Keskin and Zeevi (2014, 2018), Qiang and Bayati (2016), Javanmard and Nazerzadeh (2019), Ban and Keskin (2021), Bastani et al. (2021)) is the *linear demand model*, where the demand is

$$D_t = x_t^\top \alpha^* + x_t^\top \beta^* \cdot p_t + \epsilon_t \tag{19}$$

and α^* and β^* are unknown parameter vectors. The demand shock ϵ_t's are mean-zero random variables, and it makes no essential difference to adapt the distribution of ϵ_t to the history up to time t.

Under the linear demand model, the distribution of ϵ_t makes no difference to the optimal pricing strategy as long as it is mean-zero and sub-Gaussian. This explains why papers on the linear demand model, unlike the case of the binary demand model's unobserved utility shock ζ, do not assume knowledge of the distribution of ϵ_t.

B.2. Generalized Linear Demand Model Class

Now we show how the general model (1) recovers the binary demand model (18) and the linear demand model (19) as special cases.

EXAMPLE 1 (BINARY DEMAND MODEL). The binary demand model (18) (also (17)) is considered by a number of papers as in Table 1 (denoted by "binary model" in the "Demand Model" column). We first restrict the first dimension of x_t to be always 1. To recover (18), we can then set the function $g(\cdot)$ in (1) to be the cumulative distribution function of $-\tilde{\zeta}_t$, $\beta^* = (-\sigma, 0, \ldots, 0)^\top$, and

$$\epsilon_t = \begin{cases} 1 - g\left(x_t^\top \alpha^* - \sigma p_t\right) & \text{w.p. } g\left(x_t^\top \alpha^* - \sigma p_t\right), \\ -g\left(x_t^\top \alpha^* - \sigma p_t\right) & \text{w.p. } 1 - g\left(x_t^\top \alpha^* - \sigma p_t\right). \end{cases}$$

Thus it becomes the binary demand model (18). The parameter σ is an unknown parameter that represents the price coefficient or describes the variance of the utility shock in (18); the sub-Gaussian parameter $\bar{\sigma}^2$ can be easily chosen as $1/4$ from the boundedness of ϵ_t.

EXAMPLE 2 (LINEAR DEMAND MODEL). The linear demand model (19) can be easily recovered from (1) by letting the function $g(\cdot)$ be an identity function. As in Example 1, we restrict the first dimension of x_t to be always 1. Then, the model in Qiang and Bayati (2016) can be recovered by setting $\beta^* = (b, 0, \ldots, 0)$; the covariate-free linear demand model in den Boer and Zwart (2014), Keskin and Zeevi (2014) can be recovered by setting $\alpha^* = (a, 0, \ldots, 0)$ and $\beta^* = (b, 0, \ldots, 0)$.

B.3. Practical Motivation for No Assumptions on the Covariate Distributions

The covariates are typically customer information or features of the environment and are almost always exogenous and not in the control of the firm. Making assumptions on their distribution is restrictive for the following reasons:

- Network/peer effects: A wide range of products are influenced by network or peer effects (Seiler et al. 2017, Goolsbee and Klenow 2002, Bailey et al. 2019, Nasr and Elshar 2018). Baardman et al. (2020) show that demand prediction can be improved by incorporating such effects. Thus, the customers' features are more likely to exhibit short-term dependencies.
- Seasonality and life-cycle of product: Seasonality, day-of-week and time-of-day patterns create serial correlation in the covariates (Neale and Willems 2009). Product life-cycle effects (for tech or fashion items) also come into play, so the distribution of the covariates changes over the life-cycle of the product as the customer segment mix may be different at different stages of the product life-cycle.
- Competitors: Competing products and the action of the competitors influence demand (Armstrong et al. 2005). However, the effect from competitors' actions is complicated and the covariates of the environment cannot be identically distributed.

Appendix C: Appendix for Proofs

Throughout the following proofs, we denote

$$\bar{\gamma} := 2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}} \sqrt{2\log T + 2d\log\left(\frac{2d+T}{2d}\right)}.$$

C.1. Proofs for Section 3

C.1.1. Proofs for Subsection 3.1 We first introduce some notations and ancillary lemmas. First, the gradient and Hessian of l_t are

$$\nabla l_t(\theta) = \frac{g'(z_t^\top \theta)\cdot z_t}{h(g(z_t^\top \theta))}\left(D_t - g(z_t^\top \theta)\right) = \xi_t(\theta) z_t \tag{20}$$

$$\nabla^2 l_t(\theta) = -\eta_t(\theta) z_t z_t^\top, \tag{21}$$

where $\xi_t(\theta) \coloneqq D_t - g\left(z_t^\top \theta\right)$ and $\eta_t(\theta) \coloneqq g'\left(z_t^\top \theta\right)$. The concise form of the gradient and Hessian justifies the choice of $h(u)$ in (2).

The following lemma states that under a non-anticipatory pricing policy/algorithm, the sequence of $\{\xi_t(\theta^*)\}_{t=1}^T$ is a martingale difference sequence adapted to history observations with zero-mean $\bar{\sigma}^2$-sub-Gaussian increments.

LEMMA 3. *For $t = 1, \ldots, T$, we have*

$$\mathbb{E}\left[\xi_t(\theta^*) | \mathcal{H}_{t-1}\right] = 0.$$

In addition, $\xi_t(\theta^*)|\mathcal{H}_{t-1}$ is $\bar{\sigma}^2$-sub-Gaussian.

Proof. Note that

$$\mathbb{E}\left[\xi_t(\theta^*) | \mathcal{H}_{t-1}\right] = \mathbb{E}\left[\epsilon_t | \mathcal{H}_{t-1}\right] = 0.$$

Both the last part and the sub-Gaussianity come from Assumption 3. □

Let the (cumulative) score function

$$S_t \coloneqq \sum_{t'=1}^{t} \frac{\xi_{t'}(\theta^*)}{\bar{\sigma}} z_{t'}.$$

The following theorem measures S_t's deviation in terms of the metric induced by M_t. It can be easily proved by an application of the martingale maximal inequality on the sequence of $\xi_t(\theta^*)$.

THEOREM 7 **(Theorem 20.4, Lattimore and Szepesvári (2020))**. *For any regularization parameter $\lambda > 0$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists t \in \{1 \ldots, T\} : \|S_t\|^2_{M_t^{-1}} \geq 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det M_t}{\lambda^{2d}}\right)\right) \leq \delta.$$

Now we are ready to prove Proposition 1:

Proof of Proposition 1.

Proof. We perform a second-order Taylor's expansion for the objective function of regularized quasi-MLE (3) around the true parameter θ^*. Let

$$Q_t(\theta) \coloneqq \sum_{t'=1}^{t} l_{t'}(\theta).$$

We have

$$\begin{aligned} & Q_t(\theta^*) - \frac{\lambda \underline{g} \|\theta^*\|_2^2}{2} - Q_t(\theta) + \frac{\lambda \underline{g} \|\theta\|_2^2}{2} \\ = & -\left\langle \nabla Q_t(\theta^*) - \lambda \underline{g} \theta^*, \theta - \theta^* \right\rangle - \frac{1}{2}\left\langle \theta - \theta^*, \left(\nabla^2 Q_t(\theta') - \lambda \underline{g} I_{2d}\right)(\theta - \theta^*) \right\rangle \end{aligned} \tag{22}$$

for some θ' on the line segment between θ and θ^*.

By the optimality of $\hat{\theta}_t$,

$$Q_t(\theta^*) - \frac{\lambda \underline{g} \|\theta^*\|_2^2}{2} \le Q_t(\hat{\theta}_t) - \frac{\lambda \underline{g} \|\hat{\theta}_t\|_2^2}{2}.$$

Then from (22), we have

$$\left\langle \nabla Q_t(\theta^*) - \lambda \underline{g}\theta^*, \hat{\theta}_t - \theta^* \right\rangle + \frac{1}{2}\left\langle \hat{\theta}_t - \theta^*, \left(\nabla^2 Q_t(\tilde{\theta}') - \lambda \underline{g} I_{2d}\right)(\hat{\theta}_t - \theta^*) \right\rangle \ge 0, \tag{23}$$

for some $\tilde{\theta}'$ on the line segment between θ and θ^*.

From Assumption 2,

$$-\nabla^2 Q_t(\tilde{\theta}) = \sum_{t'=1}^{t} g'\left(z_t^\top \tilde{\theta}\right) z_{t'} z_{t'}^\top \ge \underline{g} \cdot \sum_{t'=1}^{t} z_{t'} z_{t'}^\top \quad \forall \tilde{\theta} \in \Theta.$$

Further, with Holder's inequality and (23),

$$\begin{aligned}
\left\|\nabla Q_t(\theta^*) - \lambda \underline{g}\theta^*\right\|_{M_t^{-1}} \left\|\hat{\theta}_t - \theta^*\right\|_{M_t} &\ge \left\langle \nabla Q_t(\theta^*) - \lambda \underline{g}\theta^*, \hat{\theta}_t - \theta^* \right\rangle \\
&\ge \frac{1}{2}\left\langle \hat{\theta}_t - \theta^*, \left(-\nabla^2 Q_t(\theta') + \lambda \underline{g} I_{2d}\right)(\hat{\theta}_t - \theta^*) \right\rangle \\
&\ge \frac{1}{2}\underline{g}\left\langle \hat{\theta}_t - \theta^*, M_t(\hat{\theta}_t - \theta^*) \right\rangle \\
&= \frac{1}{2}\underline{g}\left\|\hat{\theta}_t - \theta^*\right\|_{M_t}^2
\end{aligned}$$

almost surely. Consequently,

$$\left\|\nabla Q_t(\theta^*) - \lambda \underline{g}\theta^*\right\|_{M_t^{-1}} \ge \frac{1}{2}\underline{g}\left\|\hat{\theta}_t - \theta^*\right\|_{M_t} \quad a.s. \tag{24}$$

Recall that

$$S_t = \sum_{t'=1}^{t} \frac{\xi_{t'}(\theta^*)}{\bar{\sigma}} z_{t'} = \frac{1}{\bar{\sigma}} \nabla Q_t(\theta^*),$$

which implies

$$\begin{aligned}
\frac{1}{2\bar{\sigma}}\underline{g}\left\|\hat{\theta}_t - \theta^*\right\|_{M_t} &\le \frac{1}{\bar{\sigma}}\left\|-\nabla Q_t(\theta^*) + \lambda \underline{g}\theta^*\right\|_{M_t^{-1}} \\
&= \frac{1}{\bar{\sigma}}\left\|-\bar{\sigma} S_t + \lambda \underline{g}\theta^*\right\|_{M_t^{-1}} \\
&\le \|S_t\|_{M_t^{-1}} + \frac{\sqrt{\lambda}\underline{g}}{\bar{\sigma}}\sqrt{(\theta^*)^\top (\lambda M_t^{-1})\theta^*} \\
&\le \|S_t\|_{M_t^{-1}} + \frac{\sqrt{\lambda}\underline{g}}{\bar{\sigma}}\|\theta^*\|_2.
\end{aligned}$$

Here the first line comes from (24), the second line comes from the definition of S_t, the third lines comes from the norm inequality, and the last line is from the fact that $\lambda M_t^{-1} \le I_{2d}$. Thus, we complete the proof from combining Theorem 7 with $\|\theta^*\|_2 \le \bar{\theta}$.

□

Proof of Corollary 1.

Proof. Given that $\|z_t\|_2^2 \le 1$ by assumption, we can apply Lemma 19.4 of Lattimore and Szepesvári (2020) (purely algebraic analysis with no stochasticity) with $\delta = \frac{1}{T}$ in Proposition 1 to obtain the result. □

C.1.2. Proofs for Subsection 3.2 For Theorem 1, the proof idea is very intuitive. As the algorithm represents the confidence set based on the matrix M_{t-1}, the current observation x_t will either induce a small single step regret or reduce the confidence set significantly. Then we upper bound the regret of the algorithm to a summation sequence involving the covariates x_t's and matrices M_t's and then employ the elliptical potential lemma (as follows) to conclude the proof.

LEMMA 4 **(Elliptical Potential Lemma, (Lai and Wei 1982))**. *For any constant $\lambda \ge 1$ and sequence of $\{x_t\}_{t\ge 1}$ with $\|x_t\|_2 \le 1$ for all $t \ge 1$ and $x_t \in \mathbb{R}^d$, define the sequence of covariance matrices:*

$$\Sigma_0 := \lambda I_d, \quad \Sigma_t := \lambda I_d + \sum_{t'=1}^{t} x_{t'} x_{t'}^\top \quad \forall t \ge 1,$$

where I_d is the identity matrix with dimension d. Then for any $T \ge 1$, the following inequality holds

$$\sum_{t=1}^{T} \|x_t\|^2_{\Sigma_{t-1}^{-1}} \le 2d \log\left(\frac{\lambda d + T}{\lambda d}\right).$$

Proof of Theorem 1.

Proof. We define the "good event" as $\mathcal{E} = \{\theta^* \in \Theta_t \text{ for } t = 1, \dots, T\}$. From Corollary 1, we know

$$\mathbb{P}(\mathcal{E}) \ge 1 - 1/T. \tag{25}$$

At time t, under the event $\mathcal{E}$, the choice of $\theta_t = (\alpha_t, \beta_t)$ in Algorithm 1 ensures

$$r^*(x_t^\top \alpha_t, x_t^\top \beta_t) \ge r^*(x_t^\top \alpha^*, x_t^\top \beta^*). \tag{26}$$

Thus, under the event $\mathcal{E}$, the single period regret can be bounded by

$$\begin{aligned}
\text{Reg}_t := r^*(x_t^\top \alpha^*, x_t^\top \beta^*) - r(p_t; x_t^\top \alpha^*, x_t^\top \beta^*) &\le r^*(x_t^\top \alpha_t, x_t^\top \beta_t) - r(p_t; x_t^\top \alpha^*, x_t^\top \beta^*) \\
&= p_t \cdot \left(g\left(x_t^\top \alpha_t + x_t^\top \beta_t \cdot p_t\right) - g\left(x_t^\top \alpha^* + x_t^\top \beta^* \cdot p_t\right)\right) \\
&\le \bar{g}\bar{p}\left|z_t^\top (\theta_t - \theta^*)\right| \\
&\le \bar{g}\bar{p}\|z_t\|_{M_{t-1}^{-1}} \|\theta_t - \theta^*\|_{M_{t-1}} \\
&\le \bar{g}\bar{p}\|z_t\|_{M_{t-1}^{-1}} \left(\|\theta_t - \hat{\theta}_{t-1}\|_{M_{t-1}} + \|\hat{\theta}_{t-1} - \theta^*\|_{M_{t-1}}\right) \\
&\le 2\bar{g}\bar{p}\bar{\gamma}\|z_t\|_{M_{t-1}^{-1}}
\end{aligned}$$

where the functions r^* and r are introduced in the Section 2. Here the first line is from (26), the third line is from Assumption 1, the fourth line is from Holder's inequality, and the last inequality is by Corollary 1 under the event $\mathcal{E}$.

Thus, the total expected regret (the expectation is with respect to the randomness of demand shocks) can be bounded by

$$\begin{aligned}
\mathrm{Reg}_T^{\mathrm{UCB}}(x_1,\ldots,x_T) &= \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{Reg}_t \cdot \mathbb{1}_{\mathcal{E}}\right] + \mathbb{E}\left[\mathrm{Reg}_t \cdot \mathbb{1}_{\mathcal{E}^c}\right] \\
&\le 2\sum_{t=1}^{T} \bar{g}\bar{p}\bar{\gamma}\|z_t\|_{M_{t-1}^{-1}} + \mathbb{P}\left(\mathcal{E}^c\right)\cdot \bar{p}\bar{D}T \\
&\le 2\bar{g}\bar{p}\bar{\gamma}\sqrt{T\sum_{t=1}^{T}\|z_t\|^2_{M_{t-1}^{-1}}} + \mathbb{P}\left(\mathcal{E}^c\right)\cdot \bar{p}\bar{D}T \\
&\le 4\bar{g}\bar{p}\bar{\gamma}\sqrt{Td\log\left(\frac{2d\lambda+T}{2d\lambda}\right)} + \bar{p}\bar{D}
\end{aligned}$$

where UCB denotes the pricing policy specified by Algorithm 1 and $\mathcal{E}^c$ denotes the complement of the event $\mathcal{E}$. The second inequality is by Holder's inequality and the last inequality is because (25) and Lemma 4. □

Proof of Theorem 2.

Proof. Note by definitions, the single period regret at t can be bounded by:

$$\begin{aligned}
\mathrm{Reg}_t &:= r^*(x_t^\top\alpha^*, x_t^\top\beta^*) - r(p_t; x_t^\top\alpha^*, x_t^\top\beta^*) \\
&= r^*(x_t^\top\alpha^*, x_t^\top\beta^*) - r^*(x_t^\top\alpha_t, x_t^\top\beta_t) + r^*(x_t^\top\alpha_t, x_t^\top\beta_t) - r(p_t; x_t^\top\alpha^*, x_t^\top\beta^*) \\
&= r^*(x_t^\top\alpha_t^{\mathrm{MC}}, x_t^\top\beta_t^{\mathrm{MC}}) - r(p_t; x_t^\top\alpha^*, x_t^\top\beta^*) + r^*(x_t^\top\alpha^*, x_t^\top\beta^*) - r^*(x_t^\top\alpha_t, x_t^\top\beta_t) \\
&\quad + r^*(x_t^\top\alpha_t, x_t^\top\beta_t) - r^*(x_t^\top\alpha_t^{\mathrm{MC}}, x_t^\top\beta_t^{\mathrm{MC}}).
\end{aligned}$$

Then following the identical proof idea of Theorem 1 (noting $r^*(x_t^\top\alpha^*, x_t^\top\beta^*) - r^*(x_t^\top\alpha_t, x_t^\top\beta_t) < 0$ under the good event $\mathcal{E}$), we can conclude the result. □

C.2. Proof for Section 4

For the original Thompson Sampling, its analysis in Abeille and Lazaric (2017) needs its key Lemma 3. However, Algorithm 2 can not directly apply this lemma: (1) the sampling step in (7) is based on $(\hat{a}_t, \hat{b}_t)$ and $\tilde{M}_{t-1}$ but we want it to be still conditional on the event $\theta^* = (\alpha^*, \beta^*) \in \Theta_t$ due to Corollary 1; (2) the reward function is non-convex, and not Lipschitz in the estimators even in the linear demand model (so our reward function can not satisfy the setting of convex optimization

problems with linear observation as discussed in Section 6 or the Assumption 4 in Abeille and Lazaric (2017)). In summary, we want to lower bound the probability

$$\mathbb{P}\left(r^*(\tilde{a}_t,\tilde{b}_t)\geq r^*(a_t^*,b_t^*)\middle|\mathcal{H}_{t-1},\theta^*\in\Theta_t\right),$$

where $r^*(a,b)$ can be non-convex and non-Lipschitz in (a,b).

To solve the above challenge, we will first show $(\tilde{a}_t,\tilde{b}_t)$ has a distribution matched with the sample $(x_t^\top\tilde{\alpha}_{t-1},x_t^\top\tilde{\beta}_{t-1})$ conditional on $\mathcal{H}_{t-1}$ in Lemma 5, where

$$(\tilde{\alpha}_{t-1},\tilde{\beta}_{t-1}):=\hat{\theta}_{t-1}+\bar{\gamma}M_{t-1}^{-1/2}\eta_t'$$

is defined as the original Thompson sampling parameters (here $\eta'\sim\mathcal{N}(0,I_{2d})$, see Appendix D.2 for details), to connect with the Lemma 3 in Abeille and Lazaric (2017) and get Lemma 1. Then we utilize the special structure of pricing problem shown in Lemma 2 to finally conclude Lemma 6:

LEMMA 5. $(x_t^\top\tilde{\alpha}_{t-1},x_t^\top\tilde{\beta}_{t-1})|\mathcal{H}_{t-1}\overset{d}{=}(\tilde{a}_t,\tilde{b}_t)|\mathcal{H}_{t-1}$, *where* $\overset{d}{=}$ *means equal in distribution.*

Proof. Let $\eta'\sim\mathcal{N}(0,I_{2d})$, we have

$$\begin{aligned}(x_t^\top\tilde{\alpha}_{t-1},x_t^\top\tilde{\beta}_{t-1})|\mathcal{H}_{t-1}&=\begin{pmatrix}x_t & \mathbf{0}\\ \mathbf{0} & x_t\end{pmatrix}^\top\hat{\theta}_{t-1}+\bar{\gamma}\begin{pmatrix}x_t & \mathbf{0}\\ \mathbf{0} & x_t\end{pmatrix}^\top M_{t-1}^{-1/2}\eta'\\&\overset{d}{=}(\hat{a}_t,\hat{b}_t)+\bar{\gamma}\mathcal{N}\left(0,\tilde{M}_{t-1}^{-1}\right)\\&\overset{d}{=}(\hat{a}_t,\hat{b}_t)+\bar{\gamma}\tilde{M}_{t-1}^{-1/2}\eta\\&=(\tilde{a}_t,\tilde{b}_t)|\mathcal{H}_{t-1}.\end{aligned}$$

□

Proof for Lemma 1.

Proof. By Lemma 5 and Lemma 3 in Abeille and Lazaric (2017), we can get the result. □

Proof for Lemma 2.

Proof. Assume $\tilde{a}_t+\tilde{b}_t\cdot p_t^*\geq a_t^*+b_t^*\cdot p_t^*$, then

$$\begin{aligned}r^*(\tilde{a}_t,\tilde{b}_t)\geq&p_t^*\cdot g(\tilde{a}_t+\tilde{b}_t\cdot p_t^*)\\\geq&p_t^*\cdot g(a_t^*+b_t^*\cdot p_t^*)\\=&r^*(a_t^*,b_t^*),\end{aligned}$$

where the first line is by the the definition of r^*, the second line is by g is increasing from Assumption 4 and the last line is again by the definition of r^*. □

LEMMA 6.

$$\mathbb{P}\left(r^*(\tilde{a}_t,\tilde{b}_t) \geq r^*(a_t^*,b_t^*)\middle|\mathcal{H}_{t-1},\theta^*\in\Theta_t\right) \geq \frac{1}{4\sqrt{e\pi}}.$$

Proof. Denote the optimal pricing at t by $p_t^* := \arg\max_{p\in[\underline{p},\bar{p}]} r(p;a_t^*,b_t^*)$. Then since $p_t^*\in\mathcal{H}_{t-1}$, by Lemma 1 and the fact that linear function is convex,

$$\mathbb{P}\left(\tilde{a}_t+\tilde{b}_t\cdot p_t^* \geq a_t^*+b_t^*\cdot p_t^*\middle|\mathcal{H}_{t-1},\theta^*\in\Theta_t\right) \geq \frac{1}{4\sqrt{e\pi}}.$$

With Lemma 2 we can conclude the result. □

Remark. Note the above result can be extended to the reward function with cost $r^*(\mu,a,b) := \max_{p\in[\mu,\bar{p}]}(p-\mu)\cdot g(a+b\cdot p)$ for any cost $\mu\in[\underline{p},\bar{p}]$, which will be applied in the proof of Theorem 5.

C.2.1. Proof of Theorem 3

Proof. Let $\bar{\kappa} := 2\bar{\gamma}\sqrt{\log(4T^2)}$. Define

$$\tilde{\Theta}_t := \left\{(a,b)\in\mathbb{R}^2 : \|(a,b)-(\hat{a}_t,\hat{b}_t)\|_{\tilde{M}_{t-1}} \leq \bar{\kappa}\right\}.$$

Now, we define the good event as

$$\mathcal{E} := \{\theta^*\in\Theta_t,\ (\tilde{a}_t,\tilde{b}_t)\in\tilde{\Theta}_t \text{ for } t=1,\dots,T\}$$

where recall

$$\Theta_t = \left\{\theta\in\Theta : \left\|\theta-\hat{\theta}_{t-1}\right\|_{M_{t-1}} \leq \bar{\gamma}\right\}.$$

Then by Definition 1 and Corollary 1, we have

$$\mathbb{P}(\mathcal{E}) \geq 1-\frac{2}{T}. \tag{27}$$

Now under event $\mathcal{E}$, the regret can be decomposed into

$$\begin{aligned}\mathrm{Reg}_T^{\mathrm{TS}}\cdot\mathbb{1}_{\mathcal{E}} &= \sum_{t=1}^{T}\left(r^*(a_t^*,b_t^*)-r(p_t;a_t^*,b_t^*)\right)\cdot\mathbb{1}_{\mathcal{E}}\\ &= \sum_{t=1}^{T}\left(r^*(a_t^*,b_t^*)-r^*(\tilde{a}_t,\tilde{b}_t)+r^*(\tilde{a}_t,\tilde{b}_t)-r(p_t;a_t^*,b_t^*)\right)\cdot\mathbb{1}_{\mathcal{E}},\end{aligned}$$

where TS denotes the pricing policy specified by Algorithm 2.

Let

$$\mathrm{Reg}_t^{(1)} := \left(r^*(a_t^*,b_t^*)-r^*(\tilde{a}_t,\tilde{b}_t)\right)\cdot\mathbb{1}_{\mathcal{E}},$$

$$\mathrm{Reg}_t^{(2)} := \left(r^*(\tilde{a}_t,\tilde{b}_t)-r(p_t;a_t^*,b_t^*)\right)\cdot\mathbb{1}_{\mathcal{E}}.$$

We first focus on analyzing $\text{Reg}_t^{(1)}$. Denote

$$\Theta_t^{\text{OPT}} := \left\{(a,b) \in \tilde{\Theta}_t : r^*(a,b) \geq r^*(a_t^*, b_t^*)\right\}.$$

Then by Lemma 6,

$$\mathbb{P}\left(r^*(\tilde{a}_t, \tilde{b}_t) \geq r^*(a_t^*, b_t^*) \middle| \mathcal{H}_{t-1}, \theta^* \in \Theta_t\right) \geq \frac{1}{4\sqrt{e\pi}}.$$

Further, by Definition 1, when $T \geq 6$,

$$\mathbb{P}\left((\tilde{a}_t, \tilde{b}_t) \notin \tilde{\Theta}_t \middle| \mathcal{H}_{t-1}, \theta^* \in \Theta_t\right) \leq \frac{1}{T^2} \leq \frac{1}{8\sqrt{e\pi}}.$$

Then by applying union bound to the above two events,

$$\mathbb{P}\left((\tilde{a}_t, \tilde{b}_t) \in \Theta_t^{\text{OPT}} \middle| \mathcal{H}_{t-1}, \theta^* \in \Theta_t\right) \geq \frac{1}{4\sqrt{e\pi}} - \frac{1}{8\sqrt{e\pi}} = \frac{1}{8\sqrt{e\pi}}. \tag{28}$$

For any $(\tilde{a}, \tilde{b}) \in \Theta_t^{\text{OPT}}$, we have

$$\begin{aligned}
\mathbb{E}\left[\text{Reg}_t^{(1)} | \mathcal{H}_{t-1}\right] &\leq \mathbb{E}\left[\left(r^*(\tilde{a}, \tilde{b}) - r^*(\tilde{a}_t, \tilde{b}_t)\right) \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[\left(r^*(\tilde{a}, \tilde{b}) - p_t \cdot g(\tilde{a}_t + \tilde{b}_t \cdot p_t)\right) \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&\leq \mathbb{E}\left[p^*(\tilde{a}, \tilde{b}) \cdot \left(g(\tilde{a} + \tilde{b} \cdot p^*(\tilde{a}, \tilde{b})) - g(\tilde{a}_t + \tilde{b}_t \cdot p^*(\tilde{a}, \tilde{b}))\right) \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&\leq \bar{g}\bar{p}\mathbb{E}\left[\left|\tilde{a} + \tilde{b} \cdot p^*(\tilde{a}, \tilde{b}) - (\tilde{a}_t + \tilde{b}_t \cdot p^*(\tilde{a}, \tilde{b}))\right| \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&= \bar{g}\bar{p}\mathbb{E}\left[\left|(1, p^*(\tilde{a}, \tilde{b}))^\top (\tilde{a} - \tilde{a}_t, \tilde{b} - \tilde{b}_t)\right| \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&\leq \bar{g}\bar{p}\mathbb{E}\left[\|(1, p^*(\tilde{a}, \tilde{b}))\|_{\tilde{M}_{t-1}^{-1}} \|(\tilde{a} - \tilde{a}_t, \tilde{b} - \tilde{b}_t)\|_{\tilde{M}_{t-1}} \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&= \bar{g}\bar{p}\mathbb{E}\left[\|\tilde{z}_t(\tilde{a}, \tilde{b}))\|_{M_{t-1}^{-1}} \|(\tilde{a} - \tilde{a}_t, \tilde{b} - \tilde{b}_t)\|_{\tilde{M}_{t-1}} \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&\leq 4\bar{g}\bar{p}\bar{\gamma}\sqrt{\log(4T^2)}\mathbb{E}\left[\|\tilde{z}_t(\tilde{a}, \tilde{b}))\|_{M_{t-1}^{-1}} \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right],
\end{aligned} \tag{29}$$

where $p^*(a,b) = \arg\max_{p \in [\underline{p}, \bar{p}]} r(p; a, b)$ and $\tilde{z}_t(a,b) := (x_t, p^*(a,b)x_t)$. Here the first line is by the definition of Θ_t^{OPT}, the third line is from the optimality of p_t, the fourth line is from Assumptions 1 and 4, the sixth line is by Holder's inequality, the seventh line is by

$$\|(1, p^*(\tilde{a}, \tilde{b}))\|_{\tilde{M}_{t-1}^{-1}} = \sqrt{(1, p^*(\tilde{a}, \tilde{b}))^\top \begin{pmatrix} x_t & \mathbf{0} \\ \mathbf{0} & x_t \end{pmatrix}^\top M_{t-1}^{-1} \begin{pmatrix} x_t & \mathbf{0} \\ \mathbf{0} & x_t \end{pmatrix} (1, p^*(\tilde{a}, \tilde{b}))} = \|\tilde{z}_t(\tilde{a}, \tilde{b}))\|_{M_{t-1}^{-1}},$$

and the last line comes from the event $\mathcal{E}$ and $(\tilde{a}, \tilde{b}) \in \Theta_t^{\text{OPT}}$ with triangle inequality.

Let $(\tilde{a}_t', \tilde{b}_t')$ be an independent copy of $(\tilde{a}_t, \tilde{b}_t)$ following the same distribution. Then we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_t^{(1)} \middle| \mathcal{H}_{t-1}\right] &\le 4\bar{g}\bar{p}\bar{\gamma}\sqrt{\log(4T^2)}\mathbb{E}\left[\|\tilde{z}_t(\tilde{a}_t', \tilde{b}_t')\|_{M_{t-1}^{-1}} \cdot \mathbb{1}_{\mathcal{E}} \middle| (\tilde{a}_t', \tilde{b}_t') \in \Theta_t^{\mathrm{OPT}}, \mathcal{H}_{t-1}\right] \\
&\le 4\bar{g}\bar{p}\bar{\gamma}\sqrt{\log(4T^2)}\mathbb{E}\left[\|\tilde{z}_t(\tilde{a}_t', \tilde{b}_t')\|_{M_{t-1}^{-1}} \cdot \mathbb{1}_{\mathcal{E}} \cdot \mathbb{1}_{(\tilde{a}_t', \tilde{b}_t') \in \Theta_t^{\mathrm{OPT}}} \middle| \mathcal{H}_{t-1}\right] \mathbb{P}^{-1}\left((\tilde{a}_t', \tilde{b}_t') \in \Theta_t^{\mathrm{OPT}} | \mathcal{H}_{t-1}, \theta^* \in \Theta_{t-1}\right) \\
&\le 4\bar{g}\bar{p}\bar{\gamma}\sqrt{\log(4T^2)}\mathbb{E}\left[\|\tilde{z}_t(\tilde{a}_t', \tilde{b}_t')\|_{M_{t-1}^{-1}} \middle| \mathcal{H}_{t-1}\right] \cdot \mathbb{P}^{-1}\left((\tilde{a}_t', \tilde{b}_t') \in \Theta_t^{\mathrm{OPT}} | \mathcal{H}_{t-1}, \theta^* \in \Theta_{t-1}\right) \\
&\le 4\bar{g}\bar{p}\bar{\gamma}\sqrt{\log(4T^2)}\mathbb{E}\left[\|\tilde{z}_t(\tilde{a}_t', \tilde{b}_t')\|_{M_{t-1}^{-1}} \middle| \mathcal{H}_{t-1}\right] \cdot 8\sqrt{e\pi} \\
&= 32\bar{g}\bar{p}\bar{\gamma}\sqrt{e\pi\log(4T^2)}\mathbb{E}\left[\|z_t\|_{M_{t-1}^{-1}} \middle| \mathcal{H}_{t-1}\right]. \qquad (30)
\end{aligned}$$

where $z_t = (x_t, p_t x_t)$ and p_t is the price used in the algorithm. Here the first line comes by replacing $(\tilde{a}, \tilde{b})$ in (29) with a randomized parameter of $(\tilde{a}_t', \tilde{b}_t')$ restricted to the set Θ_t^{OPT}. The second line comes from the property of conditional expectation. The third line removes the indicator functions. The fourth line applies (28). The last line comes from the definition of $\tilde{z}_t(\tilde{a}_t', \tilde{b}_t')$ and z_t.

Next, for $\mathrm{Reg}_t^{(2)}$,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_t^{(2)} \middle| \mathcal{H}_{t-1}\right] &= \mathbb{E}\left[p_t \cdot (g(\tilde{a}_t + \tilde{b}_t \cdot p_t) - g(a_t^* + b_t^* \cdot p_t)) \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&\le \bar{g}\bar{p}\mathbb{E}\left[|\tilde{a}_t + \tilde{b}_t \cdot p_t - (a_t^* + b_t^* \cdot p_t)| \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&= \bar{g}\bar{p}\mathbb{E}\left[|\tilde{a}_t + \tilde{b}_t \cdot p_t - (\hat{a}_t + \hat{b}_t \cdot p_t) + (\hat{a}_t + \hat{b}_t \cdot p_t) - (a_t^* + b_t^* \cdot p_t)| \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&\le \bar{g}\bar{p}\mathbb{E}\left[\left(\|(1, p^*(\tilde{a}_t, \tilde{b}_t))\|_{\tilde{M}_{t-1}^{-1}} \|(\tilde{a}_t - \hat{a}_t, \tilde{b}_t - \hat{b}_t)\|_{\tilde{M}_{t-1}} + \|z_t\|_{M_{t-1}^{-1}} \|\hat{\theta}_{t-1} - \theta^*\|_{M_{t-1}}\right) \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&= \bar{g}\bar{p}\mathbb{E}\left[\left(\|z_t\|_{M_{t-1}^{-1}} \|(\tilde{a}_t - \hat{a}_t, \tilde{b}_t - \hat{b}_t)\|_{\tilde{M}_{t-1}} + \|z_t\|_{M_{t-1}^{-1}} \|\hat{\theta}_{t-1} - \theta^*\|_{M_{t-1}}\right) \cdot \mathbb{1}_{\mathcal{E}} \middle| \mathcal{H}_{t-1}\right] \\
&\le (2\sqrt{\log(4T^2)} + 1)\bar{g}\bar{p}\bar{\gamma}\mathbb{E}\left[\|z_t\|_{M_{t-1}^{-1}} \middle| \mathcal{H}_{t-1}\right] \\
&< 4\bar{g}\bar{p}\bar{\gamma}\sqrt{e\pi\log(4T^2)}\mathbb{E}\left[\|z_t\|_{M_{t-1}^{-1}} \middle| \mathcal{H}_{t-1}\right] \qquad (31)
\end{aligned}$$

where $z_t = (x_t, p_t x_t)$ and p_t is the price used in the algorithm. The second line is from Assumptions 1 and 4, the fourth line is by Holder's inequality, the fifth line is due to the same computation used for $\mathrm{Reg}_t^{(1)}$, and the sixth line is from the definition of event $\mathcal{E}$. Now we can combine them and conclude the final result:

$$\mathrm{Reg}_T^{\mathrm{TS}}(x_1, \ldots, x_T) \le \sum_{t=1}^{T} \mathbb{E}\left[\left(\mathrm{Reg}_t^{(1)} + \mathrm{Reg}_t^{(2)}\right) \cdot \mathbb{1}_{\mathcal{E}}\right] + \mathbb{E}\left[\bar{p}\bar{D}T \cdot \mathbb{1}_{\mathcal{E}^c}\right]$$

$$\leq 36\bar{g}\bar{p}\bar{\gamma}\sqrt{e\pi\log(4T^2)}\mathbb{E}\left[\sum_{t=1}^{T}\|z_t\|_{M_{t-1}^{-1}}\right]+\mathbb{P}\left(\mathcal{E}^c\right)\cdot\bar{p}\bar{D}T$$

$$\leq 36\bar{g}\bar{p}\bar{\gamma}\sqrt{e\pi\log(4T^2)}\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\|z_t\|^2_{M_{t-1}^{-1}}}\right]+\mathbb{P}\left(\mathcal{E}^c\right)\cdot\bar{p}\bar{D}T$$

$$\leq 36\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)}+2\bar{D}\bar{p}$$

where $\mathcal{E}^c$ denotes the complement of the event $\mathcal{E}$. The first line is by the single-period regret will be bounded by $\bar{p}\bar{D}$, the third line is by Holder's inequality and the last inequality is because (27) and Elliptical Potential Lemma, i.e. Lemma 4. □

C.3. Proofs for Section 5

C.3.1. Proof of Theorem 4

Proof. **For** $\Omega(\sqrt{T})$, it follows directly from the existing lower regret bound for unconstrained dynamic pricing problems (Broder and Rusmevichientong 2012, Keskin and Zeevi 2014). This is because unconstrained dynamic pricing problems can be seen as a special set of instances of the constrained ones where the initial inventory $C\to\infty$. Thus, a lower bound of the unconstrained case is also a lower of the constrained case. Here, we provide a proof sketch from Keskin and Zeevi (2014) for completeness. We assume a linear demand model:

$$D_t=a+b\cdot p_t+\epsilon_t,\quad \forall t=1,2,\ldots,T,$$

which can be seen as a special case of (1) by taking the link function $g(\cdot)$ as the identity function and $x_t\equiv 1,\ \forall t=1,\ldots,T$. We assume there exists $\Theta\subset\mathbb{R}^2$ such that $(a,b)\in\Theta$. Then by some algebraic manipulation, it can be proved that the single period regret by using a price p_t is lower bounded by $\lambda_1(p^*(a,b)-p_t)^2$, where $p^*(a,b)=-\frac{a}{2b}$ is the optimal price (where we little overload the notation) and $\lambda_1>0$ is a constant related to the feasible range of b. Thus, it is sufficient to lower bound $\sum_{t=1}^{T}(p^*(a,b)-p_t)^2$ with some constructed (a,b).

By van Trees inequality (Gill and Levit 1995) and treating the price p_t (from a policy mapping from past observations to p_t) as an estimator of the optimal price $p^*(a,b)$ at time t, it can be shown that

$$\sup_{(a,b)\in\Theta}\sum_{t=2}^{T}\mathbb{E}[(p^*(a,b)-p_t)^2]\geq\sum_{t=2}^{T}\frac{\lambda_2}{\lambda_3+\sup_{(a,b)\in\Theta}\sum_{t'=1}^{t-1}\mathbb{E}[(p^*(a,b)-p_{t'})^2]},$$

where the expectation is with respect to demand shocks ϵ_t's, and λ_2,λ_3 are two positive constants dependent on the boundedness assumption and the variance of demand shocks. Then with some algebraic manipulation and observing $\sup_{(a,b)\in\Theta}\sum_{t'=1}^{t-1}\mathbb{E}[(p^*(a,b)-p_{t'})^2]$ is non-decreasing in t, we can show $\sup_{(a,b)\in\Theta}\sum_{t=2}^{T}\mathbb{E}[(p^*(a,b)-p_t)^2]=\Omega(\sqrt{T})$ for any policy's prices p_t's and conclude the result.

For $\Omega(\mathcal{W}_T(\mathcal{P}_1,\ldots,\mathcal{P}_T))$, we assume $\alpha^*=(2,0)$ and $\beta^*=(0,-1)$ and there exists some $K>1$, $c=\frac{K}{2}$ and $x_t=(K,K)$ for all $t=1,\ldots,\frac{T}{2}$. For the second half of the horizon, the nature randomly chooses between the following two cases: (1) $x_t=(K,K-\delta)$ or (2) $x_t=(K,K+\delta)$ with some $\delta\in(1,K)$ for all $t=\frac{T}{2}+1,\ldots,T$. And further we assume there exists no error on demands, $\mathbb{P}(X_t=x_t)=1$ for all $t=1,\ldots,T$ and $g(x)=x$, i.e. with price p, $D_t(p)=2K-K\cdot p$ for all $t=1,\ldots,\frac{T}{2}$ and $D_t(p)=2K-(K-\delta)\cdot p$ or $D_t(p)=2K-(K+\delta)\cdot p$ for all $t=\frac{T}{2}+1,\ldots,T$ based on the nature's choice.

Now we consider the optimal pricing policy under the first case. Denote the proportion of inventory used in the second half horizon as q. Note $p\cdot D_t(p)$ is concave in p for all t and thus by this concavity, it is easy to see for a fixed q the optimal pricing policy is $p_t=1+q$ for $t=1,\ldots,\frac{T}{2}$ and $p_t=\frac{K(2-q)}{K-\delta}$ for $t=\frac{T}{2}+1,\ldots,T$. So it is sufficient to find the optimal $q^*(\delta)$ for the optimal policy. Indeed, the revenue under optimal pricing of a given q is:

$$T\cdot\left(\frac{K(1+q)(1-q)}{2}+\frac{K^2q(2-q)}{2(K-\delta)}\right),$$

and thus the optimal q can be computed as $q^*(\delta)=\frac{K}{2K-\delta}$ with optimal revenue as $\frac{TK^2}{2(2K-\delta)(K-\delta)}+\frac{TK}{2}$. Intuitively, when δ is larger, since customers are less sensitive to the price in the second half, the seller should save more inventory for the second half. Similarly, for the second case the optimal the optimal q can be computed as $q^*(\delta)=\frac{K}{2K+\delta}$ with optimal value as $\frac{TK^2}{2(2K+\delta)(K+\delta)}+\frac{TK}{2}$.

Then for any online policy π, if the proportion of inventory used in the second half horizon of it $q^\pi\le\frac{1}{2}$, then with probability $\frac{1}{2}$ the environment will become the first case and will cause the regret at least:

$$\frac{TK^2}{2(2K-\delta)(K-\delta)}+\frac{TK}{2}-T\cdot\left(\frac{3K}{8}+\frac{3K^2}{8(K-\delta)}\right)=\frac{KT\delta^2}{8(2K-\delta)(K-\delta)}.$$

Similarly if the proportion of inventory used in the second half horizon of it $q^\pi>\frac{1}{2}$, then with probability $\frac{1}{2}$ the environment will become the second case and will cause the regret at least:

$$\frac{KT\delta^2}{8(2K+\delta)(K+\delta)}.$$

Note $\mathcal{W}_T(\mathcal{P}_1,\ldots,\mathcal{P}_T)=\sum_{t=1}^T\frac{\delta}{\sqrt{2}}=\frac{T\delta}{\sqrt{2}}$ for both two cases, we finish the proof. □

C.3.2. Proof of Theorem 5 In this subsection, we denote $\tilde{r}(p;\mu,a,b)=(p-\mu)\cdot g(a+b\cdot p)$ as the revised revenue function with virtual cost μ, and $\tilde{r}^*(\mu,a,b)=\max_{p\in[\mu,\bar{p}]}\tilde{r}(p;\mu,a,b)$ as the corresponding optimal revised revenue function (assume $\mu\le\bar{p}$).

Proof sketch. The proof idea of Theorem 5 is to decompose the regret as follows:

$$\mathbb{E}\left[\sum_{t=1}^{\tau+1}\left(\tilde{r}^*(\mu_t,A_t^*,B_t^*)-\tilde{r}(p_t;\mu_t,A_t^*,B_t^*)+\mu_t(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\right)+\bar{p}c\cdot(T-\tau-1)\right]+O\left(\mathcal{W}_T(\mathcal{P}_1,\ldots,\mathcal{P}_T)\right),$$

where τ is the termination time of the algorithm:

$$\tau := \max\left\{t = 1, \ldots, T-1 : \sum_{t'=1}^{t} D_{t'} \le cT\right\},$$

which is the last period before the inventory is exhausted or equal to $T-1$ when still has inventory at $T-1$. The last term $O\left(\mathcal{W}_T(\mathcal{P}_1, \ldots, \mathcal{P}_T)\right)$ measures the effect of the non-stationarity, and we can then focus on bounding the terms in the expectation:

- $\mathbb{E}\left[\sum_{t=1}^{\tau+1} \tilde{r}^*(\mu_t, A_t^*, B_t^*) - \tilde{r}(p_t; \mu_t, A_t^*, B_t^*)\right]$. This term captures the revised regret before algorithm stops based on the revised revenue function $\tilde{r}(p; \mu_t, A_t^*, B_t^*)$, which considers the cost of each unit as μ_t. Since the price p_t is the optimal price from sampled variables $(\tilde{a}_t, \tilde{b}_t)$, this term can be bounded by a similar analysis as in Theorem 3.
- $\mathbb{E}\left[\sum_{t=1}^{\tau+1} \mu_t(c - g(\tilde{a}_t + \tilde{b}_t \cdot p_t))\right]$. This term captures the "extra revenues/negative regrets" from using inventory with the "price" μ_t, which can be bounded by $c\bar{p}\mathbb{E}[\tau + 1 - T] + O(\sqrt{T})$. Intuitively, the algorithm will consume all inventory cT at $\tau+1$ while the target consumption is only $c \cdot (\tau+1)$. Thus, these extra consumed inventories should earn some extra revenues or negative regrets. Thanks to the online gradient descent, such extra revenues can be bounded by $c\bar{p}\mathbb{E}[\tau + 1 - T] + O(\sqrt{T})$. The analysis is similar to Agrawal and Devanur (2016), which, however, focuses on the UCB algorithm under a stationary environment.
- $\mathbb{E}\left[\bar{p}\bar{c} \cdot (T - \tau - 1)\right]$. This term captures the regret caused by the early stop of the algorithm: the algorithm stops at $\tau + 1 \le T$. Such an early stop means the algorithm may consume the inventory too fast and can at most make $\mathbb{E}\left[\bar{p}\bar{c} \cdot (T - \tau - 1)\right]$ loss. However, such loss can be covered by the extra revenues $\mathbb{E}\left[\sum_{t=1}^{\tau+1} \mu_t(c - g(\tilde{a}_t + \tilde{b}_t \cdot p_t))\right]$ as discussed above.

Proof of Theorem 5.

Proof. Here we rewrite the deterministic optimization problem (9) as the following in a more detailed way:

$$\begin{aligned} r_T^*(\{\mathcal{P}_t\}) = \max & \sum_{t=1}^{T} \mathbb{E}_t\left[p_t(X_t) g(A_t^* + B_t^* \cdot p_t(X_t))\right] \\ \text{s.t.} & \sum_{t=1}^{T} \mathbb{E}_t\left[g(A_t^* + B_t^* \cdot p_t(X_t))\right] \le cT, \\ & p_t(x) : \mathcal{X} \to [\underline{p}, \bar{p}] \text{ is a measurable function for } t = 1, \ldots, T, \end{aligned}$$

where $\mathbb{E}_t$ is the expectation taken with respect to $X_t \sim \mathcal{P}_t$. Then one standard result in revenue management literature (Gallego and Van Ryzin 1997, Talluri et al. 2004, Jiang et al. 2020) is the benchmark $\mathbb{E}[\mathrm{OPT}(X_1, \ldots, X_T, cT)]$ can be upper bounded by $r_T^*(\{\mathcal{P}_t\})$:

LEMMA 7 **(Lemma 1 in (Jiang et al. 2020))**. *Under Assumption 1, 3, 4, 5, 6,*

$$\mathbb{E}[\mathrm{OPT}(X_1,\ldots,X_T,cT)] \le r_T^*(\{\mathcal{P}_t\}),$$

where the expectation is with respect to $X_t \sim \mathcal{P}_t$ *for* $t=1,\ldots,T$.

Although $r_T^*(\{\mathcal{P}_t\})$ is more tractable than $\mathbb{E}[\mathrm{OPT}(X_1,\ldots,X_T,cT)]$, we need to further upper bound it with a pricing policy which can utilize the dual variables μ_t's solved by the algorithm. We define the termination time

$$\tau = \max\left\{t=1,\ldots,T-1 : \sum_{t'=1}^{t} D_{t'} \le cT\right\},$$

which is the last period before the inventory is exhausted or equals to $T-1$ when still has inventory at $T-1$. Further, recall $\tilde{r}^*(\mu,a,b) = \max_{p\in[\underline{p},\bar{p}]} \tilde{r}(p;\mu,a,b)$ and $(A_t^*, B_t^*) = (X_t^\top \alpha^*, X_t^\top \beta^*)$, then

LEMMA 8.

$$r_T^*(\{\mathcal{P}_t\}) \le \mathbb{E}\left[\sum_{t=1}^{\tau+1}(c\mu_t + \tilde{r}^*(\mu_t, A_t^*, B_t^*)) + \bar{p}c\cdot(T-\tau-1)\right] + \sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}_T(\mathcal{P}_1,\ldots,\mathcal{P}_T).$$

The above lemma says the benchmark performance can be bounded by three parts: $\sum_{t=1}^{\tau+1}(c\mu_t + \tilde{r}^*(\mu_t; A_t^*, B_t^*))$ captures the Lagrangian formulation (with the dual variables μ_t's) of revenues gotten when inventory is exhausted, $\bar{p}c\cdot(T-\tau-1)$ captures the maximum revenue gotten after inventory is exhausted, and $\mathcal{W}_T(\mathcal{P}_1,\ldots,\mathcal{P}_T)$ captures the effect of non-stationarity of $\mathcal{P}_t$'s. The proof of the above lemma is mainly based on the dual formulations of both $r_T^*(\{\mathcal{P}_t\})$ and Wasserstein distance function. We postpone the proof to Appendix C.3.3.

Now we introduce some similar notations as in the proof of Theorem 3. Let $\bar{\gamma} = 2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log T + 2d\log\left(\frac{2d+T}{2d}\right)}$, $\bar{\kappa} = 2\bar{\gamma}\sqrt{\log(4T^2)}$ and

$$\tilde{\Theta}_t = \left\{(a,b)\in\mathbb{R}^2 : \|(a,b)-(\hat{a}_t,\hat{b}_t)\|_{\tilde{M}_{t-1}} \le \bar{\kappa}\right\}.$$

We define the good event as

$$\mathcal{E} = \{\theta^*\in\Theta_t,\ (\tilde{a}_t,\tilde{b}_t)\in\tilde{\Theta}_t \text{ for } t=1,\ldots,T\}$$

where recall

$$\Theta_t = \left\{\theta\in\Theta : \left\|\theta-\hat{\theta}_{t-1}\right\|_{M_{t-1}} \le \bar{\gamma}\right\}.$$

Then by Definition 1 and Corollary 1, we have

$$\mathbb{P}(\mathcal{E}) \ge 1-\frac{2}{T}. \tag{32}$$

For $t=1,\ldots,\tau+1$, we denote

$$\mathrm{Reg}_t^{(1)} := \left(c\mu_t + \tilde{r}^*(\mu_t, A_t^*, B_t^*) - r(p_t;\tilde{a}_t,\tilde{b}_t)\right)\cdot\mathbb{1}_{\mathcal{E}},$$

$$\mathrm{Reg}_t^{(2)} := \left(r(p_t;\tilde{a}_t,\tilde{b}_t) - r(p_t;A_t^*,B_t^*)\right)\cdot \mathbb{1}_{\mathcal{E}},$$

where μ_t, p_t are from Algorithm 3.

We first focus on analyzing $\mathrm{Reg}_t^{(1)}$. Denote

$$\Theta_t^{\mathrm{OPT}} := \left\{(a,b)\in\tilde{\Theta}_t : \tilde{r}^*(\mu_t,a,b) \geq \tilde{r}^*(\mu_t,A_t^*,B_t^*)\right\}.$$

Then by (the remark below) Lemma 6,

$$\mathbb{P}\left(\tilde{r}^*(\mu_t,\tilde{a}_t,\tilde{b}_t) \geq \tilde{r}^*(\mu_t,A_t^*,B_t^*)\middle|\mathcal{H}_{t-1},\theta^*\in\Theta_t\right) \geq \frac{1}{4\sqrt{e\pi}},$$

by noting $\mu_t\in\mathcal{H}_{t-1}$. With similar computation as in the proof of Theorem 3, we have

$$\mathbb{P}\left((\tilde{a}_t,\tilde{b}_t)\in\Theta_t^{\mathrm{OPT}}\middle|\mathcal{H}_{t-1},\theta^*\in\Theta_t\right) \geq \frac{1}{8\sqrt{e\pi}}. \tag{33}$$

For any $(\tilde{a},\tilde{b})\in\Theta_t^{\mathrm{OPT}}$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_t^{(1)}\middle|\mathcal{H}_{t-1}\right] &\leq \mathbb{E}\left[\left(c\mu_t + \max_{p\in[\mu_t,\bar{p}]}\tilde{r}(p;\mu_t,\tilde{a},\tilde{b}) - r(p_t;\tilde{a}_t,\tilde{b}_t)\right)\cdot\mathbb{1}_{\mathcal{E}}\middle|\mathcal{H}_{t-1}\right]\\
&= \mathbb{E}\left[\left(\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t)) + \max_{p\in[\mu_t,\bar{p}]}\tilde{r}(p;\mu_t,\tilde{a},\tilde{b}) - \tilde{r}(p_t;\mu_t,\tilde{a}_t,\tilde{b}_t)\right)\cdot\mathbb{1}_{\mathcal{E}}\middle|\mathcal{H}_{t-1}\right]\\
&\leq \mathbb{E}\left[\left(\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t)) + \tilde{r}(\tilde{p}_t^*;\mu_t,\tilde{a},\tilde{b}) - \tilde{r}(\tilde{p}_t^*;\mu_t,\tilde{a}_t,\tilde{b}_t)\right)\cdot\mathbb{1}_{\mathcal{E}}\middle|\mathcal{H}_{t-1}\right]\\
&\leq \mathbb{E}\left[\left(4\bar{g}\bar{p}\bar{\gamma}\sqrt{\log(4T^2)}\|\tilde{z}_t(\tilde{a},\tilde{b})\|_{M_{t-1}^{-1}} + \mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t)\right)\cdot\mathbb{1}_{\mathcal{E}}\middle|\mathcal{H}_{t-1}\right]
\end{aligned}$$

where $\tilde{z}_t(a,b) := (x_t, \tilde{p}_t^* x_t)$ and $\tilde{p}_t^* := \arg\max_{p\in[\mu_t,\bar{p}]}\tilde{r}(p;\mu_t,\tilde{a},\tilde{b})$. Here the first line is by definition of Θ_t^{OPT}, the second line is by definition of $\tilde{r}(p;\mu,a,b)$, the third line is by the optimality of p_t, and the last line is from similar computation steps as in (29). Then by the same computation steps in (30), we get

$$\mathbb{E}\left[\mathrm{Reg}_t^{(1)}\middle|\mathcal{H}_{t-1}\right] \leq 32\bar{g}\bar{p}\bar{\gamma}\sqrt{e\pi\log(4T^2)}\mathbb{E}\left[\|z_t\|_{M_{t-1}^{-1}}\middle|\mathcal{H}_{t-1}\right] + \mathbb{E}\left[\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\cdot\mathbb{1}_{\mathcal{E}}\middle|\mathcal{H}_{t-1}\right], \tag{34}$$

where $z_t = (x_t, p_t x_t)$ and p_t is the price used in the algorithm. Thus, $\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathrm{Reg}_t^{(1)}\right]$ can be bounded by

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathrm{Reg}_t^{(1)}\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{t\leq\tau+1\}}\mathbb{E}\left[\mathrm{Reg}_t^{(1)}\middle|\mathcal{H}_{t-1}\right]\right]\\
&\leq 32\bar{g}\bar{p}\bar{\gamma}\sqrt{e\pi\log(4T^2)}\sum_{t=1}^{T}\mathbb{E}\left[\|z_t\|_{M_{t-1}^{-1}}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{t\leq\tau+1\}}\mathbb{E}\left[\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\cdot\mathbb{1}_{\mathcal{E}}\middle|\mathcal{H}_{t-1}\right]\right]\\
&\leq 32\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)} + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{t\leq\tau+1\}}\mathbb{E}\left[\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\cdot\mathbb{1}_{\mathcal{E}}\middle|\mathcal{H}_{t-1}\right]\right]
\end{aligned}$$

$$= 32\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)} + \mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\cdot\mathbb{1}_{\mathcal{E}}\right]$$
$$\le 36\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)} + c\bar{p}\mathbb{E}[\tau+1-T] + \bar{p}\sqrt{(\bar{D}^2+\bar{\sigma}^2)T} + 2\bar{p}\bar{D} \tag{35}$$

where the first line is by the tower rule and $\mathbb{1}_{\{t\le\tau+1\}}\in\mathcal{H}_{t-1}$, the second line is by (34), the third line is by Lemma 4, the fourth line is again by the tower rule and $\mathbb{1}_{\{t\le\tau+1\}}\in\mathcal{H}_{t-1}$, and the last line is by the lemma below whose proof can be found in Appendix C.3.3:

Lemma 9.

$$\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\cdot\mathbb{1}_{\mathcal{E}}\right] \le c\bar{p}\mathbb{E}[\tau+1-T]+\bar{p}\sqrt{(\bar{D}^2+\bar{\sigma}^2)T}+2\bar{p}\bar{D}+4\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)}.$$

And with the same argument used in the proof of Theorem 3, $\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathrm{Reg}_t^{(2)}\right]$ can be bounded by

$$\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathrm{Reg}_t^{(2)}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{t\le\tau+1\}}\mathbb{E}\left[\mathrm{Reg}_t^{(2)}\middle|\mathcal{H}_{t-1}\right]\right] \le 4\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)}. \tag{36}$$

We use TSC(Thompson Sampling with Constraint) to denote the Algorithm 3, then

$$\begin{aligned}
\mathrm{Reg}_T^{\mathrm{TSC}} &\le \mathbb{E}\left[\sum_{t=1}^{\tau+1}(c\mu_t+\tilde{r}^*(\mu_t,A_t^*,B_t^*))+\bar{p}c\cdot(T-\tau-1)\right]+\sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}_T(\mathcal{P}_1,\dots,\mathcal{P}_T)-\mathbb{E}\left[\sum_{t=1}^{\tau}r(p_t;A_t^*,B_t^*)\right]\\
&\le \mathbb{E}\left[\sum_{t=1}^{\tau+1}\left(\mathrm{Reg}_t^{(1)}+\mathrm{Reg}_t^{(2)}+(c\mu_t+\tilde{r}^*(X_t,\mu_t))\cdot\mathbb{1}_{\mathcal{E}^c}\right)+\bar{p}c\cdot(T-\tau-1)+r(p_{\tau+1};A_{\tau+1}^*,B_{\tau+1}^*)\right]\\
&\quad+\sqrt{2}(\bar{p}\vee\bar{p})^2\bar{g}\bar{\theta}\mathcal{W}_T(\mathcal{P}_1,\dots,\mathcal{P}_T)\\
&\le \mathbb{E}\left[\sum_{t=1}^{\tau+1}(\mathrm{Reg}_t^{(1)}+\mathrm{Reg}_t^{(2)})\right]+\mathbb{P}(\mathcal{E}^c)\cdot(c+\bar{D})\bar{p}T+\bar{p}c\mathbb{E}\left[T-\tau-1\right]+\bar{D}\bar{p}+\sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}_T(\mathcal{P}_1,\dots,\mathcal{P}_T)\\
&\le \mathbb{E}\left[\sum_{t=1}^{\tau+1}(\mathrm{Reg}_t^{(1)}+\mathrm{Reg}_t^{(2)})\right]+2(c+\bar{D})\bar{p}+\bar{p}c\mathbb{E}\left[T-\tau-1\right]+\bar{D}\bar{p}+\sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}_T(\mathcal{P}_1,\dots,\mathcal{P}_T)\\
&\le 40\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)}+\bar{p}\sqrt{(\bar{D}^2+\bar{\sigma}^2)T}+7\bar{D}\bar{p}+\sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}_T(\mathcal{P}_1,\dots,\mathcal{P}_T),
\end{aligned}$$

where the first inequality is by Lemma 8, the second inequality is by definitions, the third inequality is by $\mu_t\le\bar{p}$ and $r(p_{\tau+1};A_{\tau+1}^*,B_{\tau+1}^*),\tilde{r}^*(X_t,\mu_t)\le\bar{p}\bar{D}$, the fourth inequality is by (32) and the last inequality is by (35), (36) and Assumption 6. □

C.3.3. Proofs for Ancillary Lemmas

Proof for Lemma 8.

Proof. Denote $\tilde{\mathcal{H}}_t \coloneqq \sigma(p_1,\dots,p_t,\epsilon_1,\dots,\epsilon_t,X_1,\dots,X_t)$ and $\tilde{\mathcal{H}}_0 = \sigma(\emptyset,\Omega)$, then

$$\begin{aligned}
r_T^*(\{\mathcal{P}_t\}) &\le \min_{\mu\ge 0} c\mu T + \sum_{t=1}^{T} \mathbb{E}_t\left[\tilde{r}^*(\mu, X_t^\top\alpha^*, X_t^\top\beta^*)\right] \\
&= \min_{\mu\ge 0} c\mu T + \sum_{t=1}^{T} \mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu, X^\top\alpha^*, X^\top\beta^*)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{\tau+1}\left(\min_{\mu\ge 0} c\mu + \mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu, X^\top\alpha^*, X^\top\beta^*)\right]\right)\right] + \mathbb{E}\left[\sum_{t=\tau+2}^{T}\left(\min_{\mu\ge 0} c\mu + \mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu, X^\top\alpha^*, X^\top\beta^*)\right]\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{t\le\tau+1\}}\cdot\left(c\mu_t + \mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu_t, X^\top\alpha^*, X^\top\beta^*)\middle|\tilde{\mathcal{H}}_{t-1}\right]\right)\right] \\
&+ \mathbb{E}\left[\sum_{t=\tau+2}^{T}\left(\min_{\mu\ge 0} c\mu + \mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu, X^\top\alpha^*, X^\top\beta^*)\right]\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{t\le\tau+1\}}\cdot\left(c\mu_t + \mathbb{E}_t\left[\tilde{r}^*(\mu_t, X_t^\top\alpha^*, X_t^\top\beta^*)\middle|\tilde{\mathcal{H}}_{t-1}\right]\right)\right] + \sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}_T(\mathcal{P}_1,\dots,\mathcal{P}_T) \\
&+ c\bar{p}\cdot\mathbb{E}[T-\tau-1],
\end{aligned}$$

where the first line is from weak duality and the definition of $\tilde{r}^*$, the second line is because $\bar{\mathcal{P}}_T$ is the uniform mixture of $\mathcal{P}_t$'s, and the third line is by splitting the summation into two parts and use $\sum_{T+1}^{T}(\cdot) = 0$ for simplifying the notation, the fourth line is by the suboptimal μ_t and $\mu_t, \mathbb{1}_{\{t\le\tau+1\}} \in \tilde{\mathcal{H}}_{t-1}$ with the tower rule of conditional expectation, and the last line is from the following two lemmas and their proofs can be found in this subsection.

Lemma 10. *For $t = 1,\dots,\tau+1$,*

$$\mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu_t, X^\top\alpha^*, X^\top\beta^*)\middle|\tilde{\mathcal{H}}_{t-1}\right] \le \mathbb{E}_t\left[\tilde{r}^*(\mu_t, X_t^\top\alpha^*, X_t^\top\beta^*)\middle|\tilde{\mathcal{H}}_{t-1}\right] + \sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}(\mathcal{P}_t,\bar{\mathcal{P}}_T) \quad \text{almost surely.}$$

Lemma 11.

$$\mathbb{E}\left[\sum_{t=\tau+2}^{T}\left(\min_{\mu\ge 0} c\mu + \mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu, X^\top\alpha^*, X^\top\beta^*)\right]\right)\right] \le c\bar{p}\cdot\mathbb{E}[T-\tau-1].$$

Further, since $\mu_t, \mathbb{1}_{\{t\le\tau+1\}} \in \tilde{\mathcal{H}}_{t-1}$, with the tower rule of conditional expectation, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{t\le\tau+1\}}\cdot\left(c\mu_t + \mathbb{E}_t\left[\tilde{r}^*(\mu_t, X_t^\top\alpha^*, X_t^\top\beta^*)\middle|\tilde{\mathcal{H}}_{t-1}\right]\right)\right] = \mathbb{E}\left[\sum_{t=1}^{\tau+1}\left(c\mu_t + \tilde{r}^*(\mu_t, X_t^\top\alpha^*, X_t^\top\beta^*)\right)\right],$$

and can conclude the result. □

Proof for Lemma 9.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\cdot\mathbb{1}_{\mathcal{E}}\right] &= \mathbb{E}\left[\sum_{t=1}^{\tau+1}\Big(\mu_t\cdot\Big((c-D_t)+g(A_t^*+B_t^*\cdot p_t)-g(\tilde{a}_t+\tilde{b}_t\cdot p_t)+\epsilon_t\Big)\Big)\cdot\mathbb{1}_{\mathcal{E}}\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{\tau+1}(\mu_t\cdot(c-D_t+\epsilon_t))\cdot\mathbb{1}_{\mathcal{E}}\right]\\
&\quad+\mathbb{E}\left[\sum_{t=1}^{\tau+1}\Big(\mu_t\cdot(g(A_t^*+B_t^*\cdot p_t)-g(\tilde{a}_t+\tilde{b}_t\cdot p_t))\Big)\cdot\mathbb{1}_{\mathcal{E}}\right]\\
&\le \mathbb{E}\left[\sum_{t=1}^{\tau+1}(\mu_t\cdot(c-D_t+\epsilon_t))\cdot\mathbb{1}_{\mathcal{E}}\right]+4\bar{g}\bar{p}\bar{\gamma}\sqrt{2e\pi dT\log\left(\frac{2d\lambda+T}{2d\lambda}\right)\log(4T^2)},
\end{aligned}$$

where the first line is by definition of D_t and the inequality is by similar computation steps used in proof of Theorem 3 with $0\le\mu_t\le\bar{p}$.

Denote the random variable $\mu^*:=\arg\min_{\mu\in[0,\bar{p}]}\sum_{t=1}^{\tau+1}\mu\cdot(c-D_t)$. By a standard analysis of online gradient descent on Online Convex Optimization (e.g., Theorem 3.1. in Hazan et al. (2016)), we have

$$\sum_{t=1}^{\tau+1}\mu_t\cdot(c-D_t)\le\sum_{t=1}^{\tau+1}\mu^*\cdot(c-D_t)+\frac{\bar{p}^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{\tau+1}(c-D_t)^2\quad\text{almost surely,}$$

where $\eta=\frac{\bar{p}}{\sqrt{(\bar{D}^2+\bar{\sigma}^2)T}}$. We claim $\sum_{t=1}^{\tau+1}\mu^*\cdot(c-D_t)\le c\bar{p}(\tau+1-T)$ almost surely: if $\tau=T-1$, then we can pick $\mu=0$ and thus $\sum_{t=1}^{\tau+1}\mu^*\cdot(c-D_t)\le 0$. If $\tau<T-1$, then by its definition we have $\sum_{t=1}^{\tau+1}D_t>cT$, and thus we can pick $\mu=\bar{p}$ and

$$\sum_{t=1}^{\tau+1}\mu^*\cdot(c-D_t)\le c\bar{p}(\tau+1-T).$$

Further,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau+1}(c-D_t)^2\right] &= \mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathbb{E}\left[(c-D_t)^2\middle|\mathcal{H}_{t-1},p_t\right]\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathbb{E}\left[(c-g(A_t^*+B_t^*\cdot p_t))^2+\epsilon_t^2-2\epsilon_t\cdot(c-g(A_t^*+B_t^*\cdot p_t))\middle|\mathcal{H}_{t-1},p_t\right]\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathbb{E}\left[(c-g(A_t^*+B_t^*\cdot p_t))^2+\epsilon_t^2\middle|\mathcal{H}_{t-1},p_t\right]\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{\tau+1}(c-g(A_t^*+B_t^*\cdot p_t))^2\right]+\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mathbb{E}\left[\epsilon_t^2\middle|\mathcal{H}_{t-1}\right]\right]\\
&\le \mathbb{E}\left[(\tau+1)\bar{D}^2+(\tau+1)\bar{\sigma}^2\right]\\
&\le T(\bar{D}^2+\bar{\sigma}^2),
\end{aligned}$$

where the first line is by the tower rule of conditional expectation and $\mathbb{1}_{t\le\tau+1}\in\mathcal{H}_{t-1}$, the third line is because ϵ_t is independent with p_t conditional on $\mathcal{H}_{t-1}$ and has zero mean, while $A_t^*, B_t^*\in\mathcal{H}_{t-1}$, the fifth line is by Assumption 4 and the property of sub-Gaussian variables (e.g., Proposition 2.5.2 in Vershynin (2018)), and the last line is by $\tau\le T-1$ by definition.

Combine above all with $\eta=\frac{\bar{p}}{\sqrt{(\bar{D}^2+\bar{\sigma}^2)T}}$, we can conclude

$$\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-D_t)\right]\le c\bar{p}\mathbb{E}\left[(\tau+1-T)\right]+\bar{p}\sqrt{(\bar{D}^2+\bar{\sigma}^2)T}. \tag{37}$$

Further,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\tau+1}(\mu_t\cdot(c-D_t))\cdot\mathbb{1}_{\mathcal{E}}\right]&=\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-D_t)\right]-\mathbb{E}\left[\sum_{t=1}^{\tau+1}(\mu_t\cdot(c-D_t))\cdot\mathbb{1}_{\mathcal{E}^c}\right]\\&\le\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-D_t)\right]+\mathbb{E}\left[\sum_{t=1}^{\tau+1}\bar{p}\bar{D}\cdot\mathbb{1}_{\mathcal{E}^c}\right]+\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\epsilon_t\cdot\mathbb{1}_{\mathcal{E}^c}\right]\\&\le\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-D_t)\right]+2\bar{p}\bar{D}+\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\epsilon_t\cdot\mathbb{1}_{\mathcal{E}^c}\right],\end{aligned}$$

where the second line is by $D_t\le\bar{D}+\epsilon_t$ and $\mu_t\in[0,\bar{p}]$, the third line is by (32). And thus,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\tau+1}(\mu_t\cdot(c-D_t+\epsilon_t))\cdot\mathbb{1}_{\mathcal{E}}\right]&\le\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-D_t)\right]+2\bar{p}\bar{D}+\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\epsilon_t\right]\\&=\mathbb{E}\left[\sum_{t=1}^{\tau+1}\mu_t\cdot(c-D_t)\right]+2\bar{p}\bar{D}+\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{t\le\tau+1\}}\mu_t\mathbb{E}[\epsilon_t|\mathcal{H}_{t-1}]\right]\\&\le c\bar{p}\mathbb{E}[\tau+1-T]+\bar{p}\sqrt{(\bar{D}^2+\bar{\sigma}^2)T}+2\bar{p}\bar{D},\end{aligned}$$

where the second line is by the tower rule and $\mu_t,\mathbb{1}_{\{t\le\tau+1\}}\in\mathcal{H}_{t-1}$ and last line is by (37). And combine all above we can conclude the lemma. □

Proof for Lemma 10.

Proof. We first note given any $\mu_t\in[0,\bar{p}]$, $\tilde{r}^*(\mu_t,x^\top\alpha^*,x^\top\beta^*)$ is Lipschitz in x: for $x,x'\in\mathcal{X}$, without loss of generality, we assume $\tilde{r}^*(\mu_t,x^\top\alpha^*,x^\top\beta^*)\ge\tilde{r}^*(\mu_t,(x')^\top\alpha^*,(x')^\top\beta^*)$ and denote $\tilde{p}_t^*(x):=\arg\max_{p\in[\mu_t,\bar{p}]}\tilde{r}(p;\mu_t,x^\top\alpha^*,x^\top\beta^*)$, then

$$\begin{aligned}&\tilde{r}^*(\mu_t,x^\top\alpha^*,x^\top\beta^*)-\tilde{r}^*(\mu_t,(x')^\top\alpha^*,(x')^\top\beta^*)\\=&(\tilde{p}_t^*(x)-\mu_t)g(x^\top\alpha^*+x^\top\beta^*\cdot\tilde{p}_t^*(x))-(\tilde{p}_t^*(x')-\mu_t)g((x')^\top\alpha^*+(x')^\top\beta^*\cdot\tilde{p}_t^*(x'))\\\le&(\tilde{p}_t^*(x')-\mu_t)\cdot\left|g(x^\top\alpha^*+x^\top\beta^*\cdot\tilde{p}_t^*(x'))-g((x')^\top\alpha^*+(x')^\top\beta^*\cdot\tilde{p}_t^*(x')))\right|\\\le&\bar{g}\bar{p}|(x-x',x-x')^\top(\alpha^*,\tilde{p}_t^*(x')\beta^*)|\\\le&\sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\|x-x'\|_2,\end{aligned}$$

where the first line is by definitions, the second line is by the optimality of $\tilde{p}_t^*(x)$, and the last two lines are by Assumption 1 and 4 and Holder's inequality. Thus, by the dual formulation of Wasserstein distance (Kantorovich and Rubinshtein 1958), we have for $t=1,\ldots,\tau+1$,

$$\mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu_t; X^\top\alpha^*, X^\top\beta^*)\middle|\tilde{\mathcal{H}}_{t-1}\right] - \mathbb{E}_t\left[\tilde{r}^*(\mu_t; X_t^\top\alpha^*, X_t^\top\beta^*)\middle|\tilde{\mathcal{H}}_{t-1}\right] \le \sqrt{2}(\bar{p}\vee\bar{p}^2)\bar{g}\bar{\theta}\mathcal{W}(\mathcal{P}_t, \bar{\mathcal{P}}_T).$$

□

Proof for Lemma 11.

Proof. Define the demand optimization problem:

$$\begin{aligned} r^*_{\bar{\mathcal{P}}_T} := \max\ & \mathbb{E}_{\bar{\mathcal{P}}_T}\left[D(X)\cdot\frac{g^{-1}(D(X)) - X^\top\alpha^*}{X^\top\beta^*}\right] \\ \text{s.t. } & \mathbb{E}_{\bar{\mathcal{P}}_T}\left[D(X)\right] \le c, \\ & D(x) \in [g(x^\top\alpha^* + x^\top\beta^*\cdot\bar{p}), g(x^\top\alpha^* + x^\top\beta^*\cdot\underline{p})] \quad \forall x\in\mathcal{X}, \\ & D(x): \mathcal{X}\to[0,\bar{D}] \text{ is a measurable function.} \end{aligned}$$

By Assumption 6, $r^*_{\bar{\mathcal{P}}_T}$ is a concave maximization problem and Slater's condition holds. Thus, by strong duality (Section 5.2.3 in Boyd et al. (2004)) and replacing the optimal demand function with its corresponding pricing function, we get the dual problem $\min_{\mu\ge 0} c\mu + \mathbb{E}_{\bar{\mathcal{P}}_T}\left[\tilde{r}^*(\mu; X^\top\alpha^*, X^\top\beta^*)\right] = r^*_{\bar{\mathcal{P}}_T} \le c\bar{p}$. By summing this inequality on both sides over $t=\tau+2,\ldots,T$ when $\tau\le T-2$ and noting $\tau = T-1$ the result still holds, and the proof is finished. □

Appendix D: More Algorithm Discussions and Extensions

D.1. Discussion on the UCB Pricing Algorithm

D.1.1. Discussion on the Original (GLM-)UCB Algorithm The main difference between Algorithm 1 and GLM-UCB (Filippi et al. 2010) is that the estimation of $\hat{\theta}_t$ comes from the quasi-MLE on the observed demands as in (3), instead of the observed rewards (revenues) as in GLM-UCB. This difference is based on the special structure of the pricing problem, i.e., the misalignment of the direct observations (demands) and the rewards (revenues), which is crucial to reducing the estimation error. Specifically, if the price $p_t > 1$, the error term on revenue r_t will become $p_t\bar{\sigma}^2$-sub-Gaussian, which results in larger variances and estimation errors. Indeed, when g is the identical mapping and $\epsilon_{t'}$'s are i.i.d., it is well known that the weighted least square estimation on revenues $r_{t'}$'s with weights $w_{t'} = \frac{1}{p_{t'}^2}$ will lead to the best linear unbiased estimated of θ^*, which just recovers the original linear regression on demands $D_{t'}$'s. This difference will be further exemplified in the case of Thompson sampling. As we noted in the analysis of Algorithm 2, this special structure is the key to reducing the dependence on d and relaxing the convexity assumption.

D.1.2. Sample Complexity on the Monte Carlo Approximation of UCB Optimization

Given a target approximation error ϵ such that $\|(\alpha_t,\beta_t)-(\alpha_t^{\text{MC}},\beta_t^{\text{MC}})\|_2\le\epsilon$, noting that Θ_t's volume is $\text{Vol}(\Theta_t)\propto\det(M_t)$, and a ball with radius ϵ has volume $\propto\epsilon^d$, by standard covering number computation (Proposition 4.2.12 in Vershynin (2018)), the covering number of Θ_t is $\propto\frac{\det(M_t)}{\epsilon^d}$. We assume the covering number is achieved by the set of $\epsilon-$balls denoted as $\mathcal{B}_\epsilon$, then by the definition of the covering number, there needs at least one Monte Carlo sample in every ball in $\mathcal{B}_\epsilon$ if $\|(\alpha_t,\beta_t)-(\alpha_t^{\text{MC}},\beta_t^{\text{MC}})\|_2\le\epsilon$. Thus, the expected number of samples needed for the Monte Carlo to achieve $\|(\alpha_t,\beta_t)-(\alpha_t^{\text{MC}},\beta_t^{\text{MC}})\|_2\le\epsilon$ is $\propto\frac{\det(M_t)}{\epsilon^d}$, which follows the standard analysis of coupon collector's problem (Motwani and Raghavan 1995) and suffers from the curse of dimensionality.

D.2. Discussion on Original Thompson Sampling Algorithm

For completeness, here we introduce the original Thompson sampling in Abeille and Lazaric (2017) with some ancillary lemmas needed in the proof of Theorem 3. We first state some requirements for the sampling distribution.

Definition 1 (Sampling distribution (Abeille and Lazaric 2017)). A distribution $\mathcal{D}^{\text{TS}}$ is *suitable* for Thompson sampling if it is a multivariate distribution on $\mathbb{R}^{2d}$ absolutely continuous with respect to the Lebesgues measure which satisfies the following properties:

- (anti-concentration) there exists a positive probability q such that for any $u\in\mathbb{R}^{2d}$ with $\|u\|_2=1$,

$$\mathbb{P}_{\eta\sim\mathcal{D}^{\text{TS}}}(u^\top\eta\ge 1)\ge q,$$

- (concentration) there exist positive constants c,c' such that $\forall\delta\in(0,1)$,

$$\mathbb{P}_{\eta\sim\mathcal{D}^{\text{TS}}}\left(\|\eta\|_2\le\sqrt{2cd\log\frac{2c'd}{\delta}}\right)\ge 1-\delta.$$

As shown in Abeille and Lazaric (2017), the Gaussian distribution $\eta\sim\mathcal{N}(0,I_2)$ that we use in Algorithm 2 satisfies the above definition with $d=1$, $\delta=1/T^2$, $c=c'=2$ and $q=\frac{1}{4\sqrt{e\pi}}$.

As a comparison, the direct application of the original Thompson Sampling in Abeille and Lazaric (2017) on the pricing problem is shown as below:

Algorithm 5 Original Thompson Sampling (Abeille and Lazaric 2017) for Dynamic Pricing

Input: Regularization parameter λ.

for $t = 1, \ldots, T$ **do**

Compute the estimator $\hat{\theta}_{t-1}$ by (3) and observe the covariates x_t.

Sample $\eta_t \sim \mathcal{N}(0, I_{2d})$ and compute the parameter

$$\tilde{\theta}_{t-1} = (\tilde{\alpha}_{t-1}, \tilde{\beta}_{t-1}) := \hat{\theta}_{t-1} + \left(2\sqrt{\lambda}\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log T + 2d\log\left(\frac{2d\lambda + T}{2d\lambda}\right)}\right) M_{t-1}^{-1/2}\eta_t.$$

Set the price by

$$p_t = \arg\max_{p \in [\underline{p}, \bar{p}]} r(p; x_t^\top \tilde{\alpha}_{t-1}, x_t^\top \tilde{\beta}_{t-1}),$$

and observe the demand D_t.

end for

With a modification of Assumption 2 which enlarges the domain width of g from $O(d)$ to $O(d^{3/2})$, we can get the original Thompson Sampling's regret bound as follows:

ASSUMPTION 8 **(Properties of $g(\cdot)$).** *Let*

$$\tilde{\Theta} := \left\{\theta \in \mathbb{R}^{2d} : \|\theta - \tilde{\theta}\|_2 \le 2\sqrt{d\log(4dT^2)}\left(2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log T + 2d\log\left(\frac{2d+T}{2d}\right)}\right) \text{ for some } \tilde{\theta} \in \Theta\right\}$$

where Θ is defined in Assumption 1. We assume $g(z)$ is strictly increasing, differentiable, and there exist constants $\underline{g}, \bar{g} \in \mathbb{R}$ such that $0 < \underline{g} \le g'(z) \le \bar{g} < \infty$ for all $z = x^\top\alpha + x^\top\beta \cdot p$ where $x \in \mathcal{X}$, $\theta = (\alpha, \beta) \in \tilde{\Theta}$, and $p \in [\underline{p}, \bar{p}]$.

THEOREM 8. *Under Assumption 1, 3, 8, with any $T \ge 6$, if we choose the regularization parameter $\lambda = 1$, the regret of Algorithm 5 can be bounded by*

$$36d\bar{g}\bar{p}\bar{\gamma}\sqrt{2T\log(4dT^2)\log\left(\frac{2d+T}{2d}\right)} + 2\bar{p}\bar{D} = \tilde{O}\left(d^{\frac{3}{2}}\sqrt{T}\right),$$

where $\bar{\gamma} = 2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log T + 2d\log\left(\frac{2d+T}{2d}\right)}$.

We omit the proof here since it is largely following Abeille and Lazaric (2017). Compared to Algorithm 2, there is an extra factor of $\sqrt{d}$ in both the Assumption 8 (for domain enlargement) and regret bound. To the best of our knowledge, this extra factor is also inevitable for the existing analyses of Thompson sampling algorithms on the linear bandits problem (Agrawal and Goyal 2013, Abeille and Lazaric 2017).

D.3. Pricing with Approximate Quasi-MLE Estimator

As discussed in Section 3.1, the quasi-MLE problem (3) generally cannot be solved in closed-form. In this subsection, we first show how to compute its approximate (optimal) solution $\check{\theta}_t$ through standard projected gradient descent (Algorithm 6) with bounded approximation gap (Proposition 2). We further show that both the UCB and the Thompson sampling pricing algorithms can also be adapted to the case of only accessing to the approximate solution $\check{\theta}_t$ (Algorithm 7 and 8 respectively) with regret upper bounds (Theorem 9 and Theorem 10 respectively). Finally, we provide a numerical experiment (Figure 3) to showcase the influence on the regret by using such approximate solutions.

D.3.1. (Approximate) Optimization Algorithm of Quasi-MLE problem.

In general, the quasi-MLE problem (3) does not have a closed-form solution. However, by introducing the regularization term, the (minus of) objective function $\lambda \underline{g}\|\theta\|_2^2/2 - \sum_{t'=1}^{t} l_{t'}(\theta)$ is strongly convex and thus (3) can be solved by the standard projected gradient descent as shown in Algorithm 6 (we refer readers to Section 3.4.1 in Bubeck et al. (2015) for more details).

Algorithm 6 Projected Gradient Descent for (3)

Input: Total update steps K, initial point $\theta'_1 \in \Theta$.

for $k = 1, \ldots, K-1$ **do**

Update

$$\theta'_{k+1} = \mathrm{Proj}_{\Theta}\left(\theta'_k - \eta_k \left(\sum_{t'=1}^{t} (D_{t'} - g(z_{t'}^{\top}\theta'_k)\cdot z_{t'} - \lambda \underline{g}\theta'_k\right)\right),$$

where $\mathrm{Proj}_{\Theta}(\theta)$ is the projection function to project θ in Θ and $\eta_k = \frac{2}{\underline{g}\lambda_{\min}(M_t)(k+1)}$ (here $\lambda_{\min}(M_t) \geq \lambda$ is the minimum eigenvalue of M_t).

end for

Output: Approximate solution $\check{\theta}_t := \frac{\sum_{k=1}^{K} 2k\theta'_k}{K(K+1)}$.

The following Proposition 2 shows that the approximation gap between the output approximate solution $\check{\theta}_t$ of Algorithm 6 and the optimal solution $\hat{\theta}_t$ of the quasi-MLE problem (3), which measures the gap of their values of (3), can be bounded as a function of the total update steps K:

Proposition 2. *For any $\lambda > 0$ and $t = 1, \ldots, T-1$, function $\frac{\lambda \underline{g}\|\theta\|_2^2}{2} - \sum_{t'=1}^{t} l_{t'}(\theta)$ is $\underline{g}\lambda_{min}(M_t)$-strongly convex, where $\lambda_{\min}(M_t) \geq \lambda$ is the minimum eigenvalue of M_t. By applying Algorithm 6 with K steps, its output $\check{\theta}_t$ is an approximate solution of (3) such that*

$$\sum_{t'=1}^{t} l_{t'}(\hat{\theta}_t) - \frac{\lambda \underline{g}\|\hat{\theta}_t\|_2^2}{2} - \sum_{t'=1}^{t} l_{t'}(\check{\theta}_t) + \frac{\lambda \underline{g}\|\check{\theta}_t\|_2^2}{2} \leq \frac{2L_t^2}{\underline{g}\lambda_{\min}(M_t)(K+1)},$$

where $L_t := \max_{\theta\in\Theta} \|\sum_{t'=1}^{t}(D_{t'} - g(z_{t'}^{\top}\theta))\cdot z_{t'} - \lambda \underline{g}\theta\|_2$ and $\mathbb{E}[L_t^2] \leq 12\bar{g}^2\bar{\theta}^2 t^2 + 3\bar{\sigma}^2 t + 3\lambda^2 \underline{g}^2\bar{\theta}^2$.

Although $\mathbb{E}[L_t^2]$ increases squarely in t, the curvature of $\sum_{t'=1}^t l_{t'}(\theta)$ and thus $\lambda_{\min}(M_t)$ will also change based on the collected data. As an example, if x_t's are i.i.d. from some distribution with a positive definite covariance matrix, $-\frac{\lambda \underline{g}\|\theta\|_2^2}{2} + \sum_{t'=1}^t l_{t'}(\theta)$ will be strongly convex with parameter $\lambda_{\min}(M_t) = \Omega(t)$ in expectation and thus the above convergence rate is roughly $O(t/K)$ for all $t = 1, \ldots, T$.

Proof of Proposition 2.

Proof. We first prove the strong convexity in Θ of function $\lambda \underline{g}\|\theta\|_2^2/2 - \sum_{t'=1}^t l_{t'}(\theta)$. By the second-order Taylor's expansion, for any $\theta, \theta' \in \Theta$, recall $Q_t(\theta) = \sum_{t'=1}^t l_{t'}(\theta)$,

$$\begin{aligned}
&\frac{\lambda \underline{g}\|\theta\|_2^2}{2} - Q_t(\theta) - \frac{\lambda \underline{g}\|\theta'\|_2^2}{2} + Q_t(\theta') \\
=&\left\langle -\nabla Q_t(\theta) + \lambda \underline{g}\theta, \theta - \theta' \right\rangle - \frac{1}{2}\left\langle \theta - \theta', \left(-\nabla^2 Q_t(\tilde{\theta}) + \lambda \underline{g} I_{2d}\right)(\theta - \theta') \right\rangle \\
\le&\left\langle -\nabla Q_t(\theta) + \underline{g}\theta, \theta - \theta' \right\rangle - \frac{\underline{g}\lambda_{\min}(M_t)}{2}\|\theta - \theta'\|_2^2,
\end{aligned}$$

where $\tilde{\theta} \in \Theta$ is some point on the line segment between θ and θ^* by noting Θ is convex and the second line is from Assumption 2:

$$-\nabla^2 Q_t(\tilde{\theta}) + \lambda \underline{g} I_{2d} = \sum_{t'=1}^t g'\left(z_t^\top \tilde{\theta}\right) z_{t'} z_{t'}^\top + \lambda \underline{g} I_{2d} \ge \underline{g} \cdot \sum_{t'=1}^t z_{t'} z_{t'}^\top + \lambda \underline{g} I_{2d} \ge \underline{g} M_t.$$

Thus, we can conclude the strong convexity. Further, noting $\lambda \underline{g}\|\theta\|_2^2/2 - Q_t(\theta)$ is L_t-Lipschitz with respect to Euclidean norm since L_t upper bounds the Euclidean norm of gradient $\lambda \underline{g}\theta - \nabla Q_t(\theta)$ by definition, we can directly apply the standard convergence analysis of projected gradient descent on strongly convex and Lipschitz objective function (e.g., Theorem 3.9 in Bubeck et al. (2015)), and get the second part of the proposition.

Now for bounding $\mathbb{E}[L_t^2]$, note

$$\begin{aligned}
L_t =& \max_{\theta \in \Theta} \left\| \sum_{t'=1}^t (D_{t'} - g(z_{t'}^\top \theta)) \cdot z_{t'} - \lambda \underline{g}\theta \right\|_2 \\
\le& \max_{\theta \in \Theta} \sum_{t'=1}^t \|(g(z_{t'}^\top \theta^*) - g(z_{t'}^\top \theta)) \cdot z_{t'}\|_2 + \left\| \sum_{t'=1}^t \epsilon_{t'} z_{t'} \right\|_2 + \lambda \underline{g}\|\theta\|_2 \\
\le& \max_{\theta \in \Theta} \bar{g}\|\theta^* - \theta\|_2 \sum_{t'=1}^t \|z_{t'}\|_2^2 + \left\| \sum_{t'=1}^t \epsilon_{t'} z_{t'} \right\|_2 + \lambda \underline{g}\|\theta\|_2 \\
\le& 2\bar{g}\bar{\theta} t + \left\| \sum_{t'=1}^t \epsilon_{t'} z_{t'} \right\|_2 + \lambda \underline{g}\bar{\theta}
\end{aligned}$$

where the second line is by triangle inequality and definition of $D_{t'}$, the third line is by Holder's inequality with Assumption 1, and the last line is still from Assumption 1. Further, since

$$\mathbb{E}\left[\left\| \sum_{t'=1}^t \epsilon_{t'} z_{t'} \right\|_2^2 \right] = \mathbb{E}\left[\sum_{t'=1}^t \epsilon_{t'}^2 \|z_{t'}\|_2^2 \right] \le \mathbb{E}\left[\sum_{t'=1}^t \epsilon_{t'}^2 \right] \le t\bar{\sigma}^2,$$

by Assumption 3 and properties of sub-Gaussian variables (Proposition 2.5.2 in Vershynin (2018)), with elementary inequality $(x+y+z)^2 \le 3(x^2+y^2+z^2)$ we can conclude the result. □

D.3.2. Pricing Algorithms with Approximate Quasi-MLE Estimator We first denote $\Delta_t, t = 1, ..., T-1$ as the upper bounds of approximation gap between the optimal solution $\hat{\theta}_t$ of (3) and its approximate solution $\check{\theta}_t$ (from Algorithm 6 with some K) as the following:

$$\sum_{t'=1}^{t} l_{t'}(\hat{\theta}_t) - \frac{\lambda \underline{g} \|\hat{\theta}_t\|_2^2}{2} - \sum_{t'=1}^{t} l_{t'}(\check{\theta}_t) + \frac{\lambda \underline{g} \|\check{\theta}_t\|_2^2}{2} \le \Delta_t \quad \forall t = 1, ..., T-1. \tag{38}$$

By Proposition 2, we can choose $\Delta_t = \frac{2L_t^2}{\underline{g}\lambda_{\min}(M_t)(K+1)}$ and use the total update steps K to control Δ_t. For a general link function $g(\cdot)$, $L_t = \max_{\theta\in\Theta} \| \sum_{t'=1}^{t} (D_{t'} - g(z_{t'}^\top \theta)) \cdot z_{t'} - \lambda \underline{g}\theta \|_2$ can be hard to compute. In practice, one option is to use Monte Carlo method, where we randomly generate M θ's (denoted by $\theta_m,\ m = 1, \ldots, M$) from a uniform distribution over Θ, and get an approximated value of L_t by $L_t^{\mathrm{MC}} = \max_{m=1,\ldots,M} \| \sum_{t'=1}^{t} (D_{t'} - g(z_{t'}^\top \theta_m)) \cdot z_{t'} - \lambda \underline{g}\theta_m \|_2$.

The following Algorithm 7 and Algorithm 8 show how to adapt the UCB (Algorithm 1) and the Thompson sampling (Algorithm 2) pricing algorithms to the case with the approximate solution $\check{\theta}_t$, with provable regret upper bounds as demonstrated by Theorem 9 and Theorem 10 respectively.

To account for this approximation gap, the confidence bound needs to be enlarged in the UCB pricing (Algorithm 7), and the sampling distribution should also be inflated in the Thompson sampling pricing (Algorithm 8). We also remark that the Assumption 4 for Algorithm 8 should be slightly changed in that the set $\tilde{\Theta}$ should be redefined by

$$\tilde{\Theta} := \left\{ \theta \in \mathbb{R}^{2d} : \exists \tilde{\theta} \in \Theta, \|\theta - \tilde{\theta}\|_2 \le 4\bar{\theta}\sqrt{\log(4T^2)} + \frac{4\bar{\sigma}}{\underline{g}} \sqrt{\left(2\log T + 2d\log\left(\frac{2d+T}{2d}\right)\right)\log(4T^2)} + 2\sqrt{\log(4T^2)}\bar{\Delta} \right\},$$

where $\bar{\Delta} = \max_{t=1,\ldots,T-1} \Delta_t$. This is because the approximate quasi-MLE solution may further enlarge the sampling region.

THEOREM 9. *Under Assumptions 1, 2 and 3, with any sequence $\{x_t\}_{t=1,\ldots,T}$, if we choose the regularization parameter $\lambda = 1$, the regret of Algorithm 7 is upper bounded by*

$$4\bar{g}\bar{p}\left(\bar{\gamma} + \sqrt{\frac{2\bar{\Delta}}{\underline{g}}}\right)\sqrt{dT\log\left(\frac{2d+T}{2d}\right)} + 2\bar{p}\bar{D} = \tilde{O}\left(d\sqrt{T} + \sqrt{dT\bar{\Delta}}\right),$$

where $\bar{\Delta} = \max_{t=1,\ldots,T-1} \Delta_t$ and $\bar{\gamma} = 2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}}\sqrt{2\log T + 2d\log\left(\frac{2d+T}{2d}\right)}$ represents an upper bound for the confidence volume.

THEOREM 10. *Under Assumption 1, 3, 4, with any sequence $\{x_t\}_{t=1,\ldots,T}$ and any $T \ge 6$, if we choose the regularization parameter $\lambda = 1$, the regret of Algorithm 8 can be bounded by*

$$36\bar{g}\bar{p}\left(\bar{\gamma} + \sqrt{\frac{2\bar{\Delta}}{\underline{g}}}\right)\sqrt{2e\pi dT\log\left(\frac{2d+T}{2d}\right)\log(4T^2)} + 2\bar{p}\bar{D} = \tilde{O}\left(d\sqrt{T} + \sqrt{dT\bar{\Delta}}\right),$$

Algorithm 7 UCB Pricing with Approximation

Input: Regularization parameter λ.

for $t = 1, ..., T$ **do**

Compute the estimators $\check{\theta}_{t-1}$ of (3) with approximation gap upper bounded by Δ_{t-1} and its confidence set

$$\Theta_t := \left\{ \theta \in \Theta : \left\| \theta - \check{\theta}_{t-1} \right\|_{M_{t-1}} \leq 2\sqrt{\lambda}\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}} \sqrt{2\log T + 2d\log\left(\frac{2d\lambda + T}{2d\lambda}\right)} + \sqrt{\frac{2}{\underline{g}}\Delta_{t-1}} \right\}.$$

Observe covariates x_t and choose the UCB parameter which maximizes the expected revenue:

$$(\alpha_t, \beta_t) = \arg\max_{(\alpha,\beta)\in\Theta_t} \; r^*\left(x_t^\top \alpha, x_t^\top \beta\right). \tag{39}$$

Set the price by

$$p_t = \arg\max_{p\in[\underline{p},\bar{p}]} r(p; x_t^\top \alpha_t, x_t^\top \beta_t).$$

end for

Algorithm 8 Thompson Sampling Pricing with Approximation

Input: Regularization parameter λ.

for $t = 1, \ldots, T$ **do**

Compute the estimator $\check{\theta}_{t-1}$ of (3) with approximation gap upper bounded by Δ_{t-1}, observe feature x_t.

Compute the estimator $(\hat{a}_t, \hat{b}_t) := (x_t^\top \check{\alpha}_{t-1}, x_t^\top \check{\beta}_{t-1})$.

Sample $\eta_t \sim \mathcal{N}(0, I_2)$ and compute the parameter

$$(\tilde{a}_t, \tilde{b}_t) := (\hat{a}_t, \hat{b}_t) + \left(2\sqrt{\lambda}\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}} \sqrt{2\log T + 2d\log\left(\frac{2d\lambda + T}{2d\lambda}\right)} + \sqrt{\frac{2}{\underline{g}}\Delta_{t-1}} \right) \tilde{M}_{t-1}^{-1/2} \eta_t. \tag{40}$$

Set the price by

$$p_t = \arg\max_{p\in[\underline{p},\bar{p}]} r(p; \tilde{a}_t, \tilde{b}_t),$$

and observe the demand D_t.

end for

where $\bar{\Delta} = \max_{t=1,\ldots,T-1} \Delta_t$ *and* $\bar{\gamma} = 2\bar{\theta} + \frac{2\bar{\sigma}}{\underline{g}} \sqrt{2\log T + 2d\log\left(\frac{2d+T}{2d}\right)}$.

Theorem 9 and Theorem 10 provide regret upper bounds for Algorithm 7 and Algorithm 8 respectively. Compared to Theorem 1 and Theorem 3, the additional term $\tilde{O}\left(\sqrt{dT\bar{\Delta}}\right)$ captures

the regret by applying approximate quasi-MLE estimators. The proof ideas of them are almost identical to those of Theorem 1 and Theorem 3 by using the following lemma that captures the distance between the approximate solution $\check{\theta}_t$ and the optimal solution $\hat{\theta}_t$. The lemma justifies the choice of the confidence set in Algorithm 7 and the sampling step in Algorithm 8, and can be implied from the optimality condition. The exact proof can be founded in the end of this subsection.

LEMMA 12. *Recall that $\hat{\theta}_t$ is the optimal solution to the optimization problem (3). For any $\theta \in \Theta$, we have*

$$\sum_{t'=1}^{t} l_{t'}(\hat{\theta}_t) - \frac{\lambda \underline{g} \|\hat{\theta}_t\|_2^2}{2} - \sum_{t'=1}^{t} l_{t'}(\theta) + \frac{\lambda \underline{g} \|\theta\|_2^2}{2} \geq \frac{1}{2} \underline{g} \|\hat{\theta}_t - \theta\|_{M_t}^2.$$

Numerical experiment.

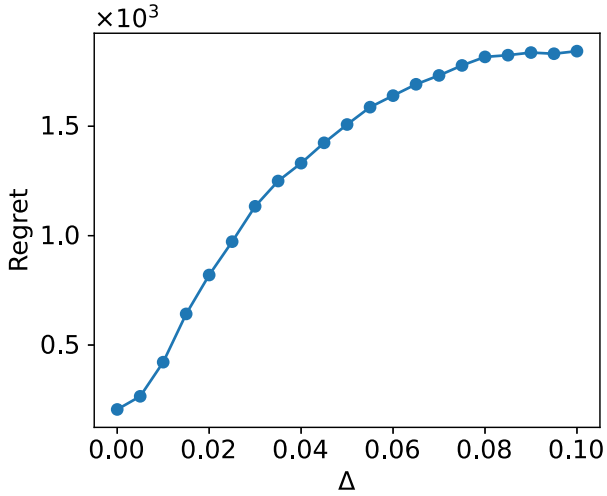

Figure 3 Regret when varying Δ in Algorithm 8.

Figure 3 demonstrates the performances of Algorithm 8 on the same demand setting as §6 with different approximation gaps $\bar{\Delta}$. We set

$$\check{\theta}_t = \hat{\theta}_t + \Delta \times [1, 1, \ldots, 1]$$

as the approximate estimators used in Algorithm 8 and thus by simple computation $\bar{\Delta} \propto \Delta$. We plot the regrets at $T = 1500$ and $d = 6$ over $\Delta = 0, 0.005, 0.01, \ldots, 0.1$. From Figure 3, the regret is roughly $\propto \sqrt{\Delta}$, which is consistent with the analysis of Algorithm 8 in Theorem 10. Here the sampling step of Algorithm 8 is by

$$(\hat{a}_t, \hat{b}_t) + \left(\frac{\sqrt{d}}{10} + \frac{T\Delta}{10} \right) \tilde{M}_{t-1}^{-1/2} \eta_t.$$

Proof of Lemma 12.

Proof. Recall that the regularized quasi-MLE at time t is defined as

$$Q_t(\theta) - \frac{\lambda \underline{g} \|\theta\|_2^2}{2},$$

where $Q_t(\theta) = \sum_{t'=1}^{t} l_t(\theta)$, with Hessian matrix

$$\sum_{t'=1}^{t} -g'(z_{t'}^\top \theta) z_{t'} z_{t'}^\top - \lambda \underline{g} I_{2d},$$

which is negative definite by the assumption that $g'(z_t^\top \theta) > 0$ (in the feasible domain of related parameters). Thus, the regularized quasi-MLE is concave in θ. We can then perform a second-order Taylor's expansion around the optimal solution $\hat{\theta}_t$ in Θ with any point $\theta \in \Theta$,

$$\begin{aligned}
&Q_t(\hat{\theta}_t) - \frac{\lambda \underline{g} \|\hat{\theta}_t\|_2^2}{2} - Q_t(\theta) + \frac{\lambda \underline{g} \|\theta\|_2^2}{2} \\
=& -\left\langle \nabla Q_t(\hat{\theta}_t) - \lambda \underline{g} \hat{\theta}_t, \theta - \hat{\theta}_t \right\rangle - \frac{1}{2} \left\langle \hat{\theta}_t - \theta, \left(\nabla^2 Q_t(\theta') - \lambda \underline{g} I_{2d}\right)(\hat{\theta}_t - \theta) \right\rangle \\
\geq& -\frac{1}{2} \left\langle \hat{\theta}_t - \theta, \left(\nabla^2 Q_t(\theta') - \lambda \underline{g} I_{2d}\right)(\hat{\theta}_t - \theta) \right\rangle \\
\geq& \frac{1}{2} \underline{g} \|\hat{\theta}_t - \theta\|_{M_t}^2,
\end{aligned}$$

where $\theta' \in \Theta$ is a point between θ and $\hat{\theta}_t$, the first inequality is by the concavity and the optimality of $\hat{\theta}_t$ in the compact set Θ, and the second inequality is by $-\nabla^2 Q_t(\theta') \geq \underline{g} \sum_{t'=1}^{t} z_{t'} z_{t'}^\top$. □

Appendix E: Appendix for Numerical Experiments

Benchmarks.

- **CILS.** By covariate-free constrained iterated least square (CILS) algorithm (Keskin and Zeevi 2014), the price p_t is set by

$$p_t = \begin{cases} \bar{p}_{t-1} + \mathrm{sgn}(\delta_t) \kappa t^{-\frac{1}{4}}, & \text{if } |\delta_t| < \kappa t^{-\frac{1}{4}}, \\ p^*(\hat{\theta}_t), & \text{otherwise}, \end{cases}$$

where $\hat{\theta}_t$ is the least square estimator for the unknown parameters, $\bar{p}_{t-1}$ is the average of the prices over the period 1 to $t-1$, and $\delta_t = p^*(\hat{\theta}_t) - \bar{p}_{t-1}$. The intuition is that if the tentative price $p^*(\hat{\theta}_t)$ stays too close to the history average, we will introduce a small perturbation as price experimentation to encourage the parameter learning. The parameter κ is a hyper-parameter.

- **Greedy_Single.** The price p_t is set as the solution of the single-period pricing problem with inventory constraint based on $(\hat{a}_t, \hat{b}_t)$:

$$\begin{aligned}
\max_{p_t} \quad & p_t \cdot g(\hat{a}_t + \hat{b}_t \cdot p_t) \\
\text{s.t.} \quad & g(\hat{a}_t + \hat{b}_t \cdot p_t) \leq c, \\
& p_t \in [0.1, 5].
\end{aligned}$$

Further, when $\hat{a}_t < 0$ or $\hat{b}_t > 0$, we choose $p_t = 5$ for saving the inventory under large uncertainty.

- **Greedy_Dual.** Greedy_Dual is same as TS_Dual expect replacing $(\tilde{a}_t, \tilde{b}_t)$ by $(\hat{a}_t, \hat{b}_t)$. Specifically, we replace equation (12) by

$$p_t = \arg\max_{p\in[\mu_t,5]} \tilde{r}(p;\mu_t,\hat{a}_t,\hat{b}_t).$$

Hyper-parameter Tuning.

For all UCB and TS algorithms, we choose the regularization parameter $\lambda = 1$. And after moderate tuning, we choose the hyper-parameter (if any) of each algorithm as follows:

- CILS: We choose $\kappa = \frac{d}{10}$ in our experiments.
- UCB: We set the confidence set for UCB algorithm (with any number of Monte Carlo samples) by

$$\Theta_t = \left\{\theta\in\Theta : \left\|\hat{\theta}_{t-1}-\theta\right\|^2_{M_{t-1}} \le \frac{d}{10}\right\}.$$

- TS and TS_Ori: We sample the TS parameter by $(\hat{a}_t,\hat{b}_t) + \frac{\sqrt{d}}{10}\tilde{M}_{t-1}^{-1/2}\eta_t$ for TS and $\hat{\theta}_{t-1} + \frac{\sqrt{d}}{25}M_{t-1}^{-1/2}\eta_t$ for TS_Ori.
- TS_Dual: We sample the TS parameter by $(\hat{a}_t,\hat{b}_t) + \frac{\sqrt{d}}{10}\tilde{M}_{t-1}^{-1/2}\eta_t$, and the update step for the dual variable is set by:

$$\mu_{t+1} = \mathrm{Proj}_{[0,5]}\left(\mu_t + 0.05\cdot(D_t - c)\right).$$

- Greedy_Dual: The update step for the dual variable is set by:

$$\mu_{t+1} = \mathrm{Proj}_{[0,5]}\left(\mu_t + 0.05\cdot(D_t - c)\right).$$