Meta Dynamic Pricing: Learning Across Experiments

Hamsa Bastani
Operations, Information and Decisions, Wharton School, hamsab@wharton.upenn.edu
David Simchi-Levi
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, dslevi@mit.edu
Ruihao Zhu
Statistics and Data Science Center, Massachusetts Institute of Technology, rzhu@mit.edu

We study the problem of learning shared structure *across* a sequence of dynamic pricing experiments for related products. We consider a practical formulation where the unknown demand parameters for each product come from an unknown distribution (prior) that is shared across products. We then propose a meta dynamic pricing algorithm that learns this prior online while solving a sequence of Thompson sampling pricing experiments (each with horizon T) for N different products. Our algorithm addresses two challenges: (i) balancing the need to learn the prior (*meta-exploration*) with the need to leverage the estimated prior to achieve good performance (*meta-exploitation*), and (ii) accounting for uncertainty in the estimated prior by appropriately "widening" the prior as a function of its estimation error, thereby ensuring convergence of each price experiment. Unlike prior-independent approaches, our algorithm's meta regret grows sublinearly in N; an immediate consequence of our analysis is that the price of an unknown prior in Thompson sampling is negligible in experiment-rich environments with shared structure (large N). Numerical experiments on synthetic and real auto loan data demonstrate that our algorithm significantly speeds up learning compared to prior-independent algorithms or a naive approach of greedily using the updated prior across products.

Key words: Thompson sampling, transfer learning, dynamic pricing, meta learning

1. Introduction

Experimentation is popular on online platforms to optimize a wide variety of elements such as search engine design, homepage promotions, and product pricing. This has led firms to perform an increasing number of experiments, and several platforms have emerged to provide the infrastructure for these firms to perform experiments at scale (see, *e.g.*, Optimizely 2019). State-of-the-art techniques in these settings employ bandit algorithms (*e.g.*, Thompson sampling), which seek to adaptively learn treatment effects while optimizing performance *within* each experiment (Thompson 1933, Scott 2015). However, the large number of related experiments begs the question: can we transfer knowledge *across* experiments?

We study this question for Thompson sampling algorithms in dynamic pricing applications that involve a large number of related products. Dynamic pricing algorithms enable retailers to optimize profits by sequentially experimenting with product prices, and learning the resulting customer

demand (Kleinberg and Leighton 2003, Besbes and Zeevi 2009). Such algorithms have been shown to be especially useful for products that exhibit relatively short life cycles (Ferreira et al. 2015), stringent inventory constraints (Xu et al. 2019), strong competitive effects (Fisher et al. 2017), or the ability to offer personalized coupons/pricing (Zhang et al. 2017, Ban and Keskin 2017). In all these cases, the demand of a product is estimated as a function of the product's price (chosen by the decision-maker) and a combination of exogenous features as well as product-specific and customer-specific features. Through carefully chosen price experimentation, the decision-maker can learn the price-dependent demand function for a given product, and choose an optimal price to maximize profits (Qiang and Bayati 2016, Cohen et al. 2016, Javanmard and Nazerzadeh 2019). Dynamic pricing algorithms based on Thompson sampling have been shown to be particularly successful in striking the right balance between exploring (learning the demand) and exploiting (offering the estimated optimal price), and are widely considered to be state-of-the-art (Thompson 1933, Agrawal and Goyal 2013, Russo and Van Roy 2014, Ferreira et al. 2018).

The decision-maker typically runs a separate pricing experiment (*i.e.*, dynamic pricing algorithm) for each product. However, this approach can waste valuable samples re-discovering information shared across different products. For example, students may be more price-sensitive than general customers; as a result, many firms such as restaurants, retailers and movie theaters offer student discounts. This implies that the coefficient of student-specific price elasticity in the demand function is positive for many products (although the specific value of the coefficient likely varies across products). Similarly, winter clothing may have higher demand in the fall and lower demand at the end of winter. This implies that the demand functions of winter clothing may have similar coefficients for the features indicating time of year. In general, there may even be complex correlations between coefficients of the demand functions of products that are shared. For example, the price-elasticities of products are often negatively correlated with their demands, *i.e.*, customers are willing to pay higher prices when the demand for a product is high. Thus, one may expect that the demand functions for related products may share some (a priori unknown) common structure, which can be learned *across* products. Note that the demand functions are unlikely to be exactly the same, so a decision-maker would still need to conduct separate pricing experiments for each product. However, accounting for shared structure during these experiments may significantly speed up learning per product, thereby improving profits.

In this paper, we propose an approach to learning shared structure across pricing experiments. We begin by noting that the key (and only) design decision in Thompson sampling methods is the Bayesian prior over the unknown parameters. This prior captures shared structure of the kind we described above — *e.g.*, the mean of the prior on the student-specific price-elasticity coefficient may be positive with a small standard deviation. It is well known that choosing a good (bad) prior

significantly improves (hurts) the empirical performance of the algorithm (Chapelle and Li 2011, Honda and Takemura 2014, Liu and Li 2015, Russo et al. 2018). However, the prior is typically unknown in practice, particularly when the decision-maker faces a cold start. While the decision-maker can use a *prior-independent* algorithm (Agrawal and Goyal 2013), such an approach achieves poor empirical performance due to over-exploration; we demonstrate a substantial gap between the prior-independent and prior-dependent approaches in our experiments on synthetic and real data. In particular, knowledge of the correct prior enables Thompson sampling to appropriately balance exploration and exploitation (Russo and Van Roy 2014). Thus, the decision-maker needs to learn the true prior (*i.e.*, shared structure) *across* products to achieve good performance. We propose a meta dynamic pricing algorithm that efficiently achieves this goal.

We first formulate the problem of learning the true prior online while solving a sequence of pricing experiments for different products. Our meta dynamic pricing algorithm requires two key ingredients. First, for each product, we must balance the need to learn about the prior ("meta-exploration") with the need to leverage the prior to achieve strong performance for the current product ("meta-exploitation"). In other words, our algorithm balances an additional exploration-exploitation tradeoff across price experiments. Second, a key technical challenge is that finite-sample estimation errors of the prior may significantly impact the performance of Thompson sampling for any given product. In particular, vanilla Thompson sampling may fail to converge with an incorrect prior; as a result, directly using the estimated prior across products can result in poor performance. In order to maintain strong performance guarantees for every product, we increase the variance of the estimated prior by a term that is a function of the prior's estimated finite-sample error. Thus, we use a more conservative approach (a wide prior) for earlier products when the prior is uncertain; over time, we gain a better estimate of the prior, and can leverage this knowledge for better empirical performance. Our algorithm provides an exact prior correction path over time to achieve strong performance guarantees across all pricing problems. We prove that, when using our algorithm, the price of an unknown prior for Thompson sampling is negligible in experiment-rich environments (*i.e.*, as the number of products grows large).

1.1. Related Literature

Experimentation is widely used to optimize decisions in a data-driven manner. This has led to a rich literature on bandits and A/B testing (Lai and Robbins 1985, Auer 2002, Dani et al. 2008, Rusmevichientong and Tsitsiklis 2010, Besbes et al. 2014, Johari et al. 2015, Bhat et al. 2019). This literature primarily proposes learning algorithms for a single experiment, while our focus is on meta-learning across experiments. There has been some work on meta-learning algorithms in the bandit setting (Hartland et al. 2006, Maes et al. 2012, Wang et al. 2018, Sharaf and Daumé III

2019) as well as the more general reinforcement learning setting (Finn et al. 2017, 2018, Yoon et al. 2018). Relatedly, Raina et al. (2006) propose constructing an informative prior based on data from similar learning problems. These papers provide heuristics for learning exploration strategies given a fixed set of past problem instances. However, they do not prove any theoretical guarantees on the performance or regret of the meta-learning algorithm. To the best of our knowledge, our paper is the first to propose a meta-learning algorithm in a bandit setting with provable regret guarantees.

We study the specific case of dynamic pricing, which aims to learn an unknown demand curve in order to optimize profits. We focus on dynamic pricing because meta-learning is particularly important in this application, *e.g.*, online retailers such as Rue La La may run numerous pricing experiments for related fashion products. We believe that a similar approach could be applied to multi-armed or contextual bandit problems, in order to inform the prior for Thompson sampling across a sequence of related bandit problems.

Dynamic pricing has been found to be especially useful in settings with short life cycles or limited inventory (*e.g.*, fast fashion or concert tickets, see Ferreira et al. 2015, Xu et al. 2019), among online retailers that constantly monitor competitor prices and adjust their own prices in response (Fisher et al. 2017), or when prices can be personalized based on customer-specific price elasticities (*e.g.*, through personalized coupons, see Zhang et al. 2017). Several papers have designed near-optimal dynamic pricing algorithms for pricing a product by balancing the resulting exploration-exploitation tradeoff (Kleinberg and Leighton 2003, Besbes and Zeevi 2009, Araman and Caldentey 2009, Farias and Van Roy 2010, Harrison et al. 2012, Broder and Rusmevichientong 2012, den Boer and Zwart 2013, Keskin and Zeevi 2014). Recently, this literature has shifted focus to pricing policies that dynamically optimize the offered price with respect to exogenous features (Qiang and Bayati 2016, Cohen et al. 2016, Javanmard and Nazerzadeh 2019) as well as customer-specific features (Ban and Keskin 2017). We adopt the linear demand model proposed by Ban and Keskin (2017), which allows for feature-dependent heterogeneous price elasticities.

We note that the existing dynamic pricing literature largely focuses on the single-product setting. A few papers consider performing price experiments jointly on a set of products with overlapping inventory constraints, or with substitutable demand (Keskin and Zeevi 2014, Agrawal and Devanur 2014, Ferreira et al. 2018). However, in these papers, price experimentation is still performed independently per product, and any learned parameter knowledge is not shared across products to inform future learning. In contrast, we propose a meta dynamic pricing algorithm that learns the distribution of unknown parameters of the demand function across products.

Our learning strategy is based on Thompson sampling, which is widely considered to be state-of-the-art for balancing the exploration-exploitation tradeoff (Thompson 1933). Several papers have studied the sensitivity of Thompson sampling to prior misspecification. For example, Honda and

Takemura (2014) show that Thompson sampling still achieves the optimal theoretical guarantee with an incorrect but uninformative prior, but can fail to converge if the prior is not sufficiently conservative. Liu and Li (2015) provide further support for this finding by showing that the performance of Thompson sampling for any given problem instance depends on the probability mass (under the provided prior) placed on the underlying parameter; thus, one may expect that Thompson sampling with a more conservative prior (*i.e.*, one that places nontrivial probability mass on a wider range of parameters) is more likely to converge when the true prior is unknown. It is worth noting that Agrawal and Goyal (2013) and Bubeck and Liu (2013) propose a *prior-independent* form of Thompson sampling, which is guaranteed to converge to the optimal policy even when the prior is unknown by conservatively increasing the variance of the posterior over time. However, the use of a more conservative prior creates a significant cost in empirical performance (Chapelle and Li 2011). For instance, Bastani et al. (2017) empirically find through simulations that the conservative prior-independent Thompson sampling is significantly outperformed by vanilla Thompson sampling even when the prior is misspecified. We empirically find, through experiments on synthetic and real datasets, that learning and leveraging the prior can yield much better performance compared to a prior-independent approach. As such, the choice of prior remains an important design choice in the implementation of Thompson sampling (Russo et al. 2018). We propose a meta-learning algorithm that learns the prior across pricing experiments on related products to attain better performance. We also empirically demonstrate that a naive approach of greedily using the updated prior performs poorly, since it may cause Thompson sampling to fail to converge to the optimal policy for some products. Instead, our algorithm gracefully tunes the width of the estimated prior as a function of the uncertainty in the estimate over time.

1.2. Main Contributions

We highlight our main contributions below:

1. *Model:* We formulate our problem as a sequence of N different dynamic pricing problems, each with horizon T. Importantly, the unknown parameters of the demand function for each product are drawn i.i.d. from a shared (unknown) multivariate gaussian prior.

2. *Algorithm:* We propose two meta-learning pricing policies, `Meta-DP` and `Meta-DP++`. The former learns only the mean of the prior, while the latter learns both the mean and the covariance of the prior across products. Both algorithms address two challenges: (i) balancing the need to learn the prior (*meta-exploration*) with the need to leverage the current estimate of the prior to achieve good performance (*meta-exploitation*), and (ii) accounting for uncertainty in the estimated prior by conservatively widening the prior as a function of its estimation error (as opposed to directly using the estimated prior, which may cause Thompson sampling to fail on some products).

3. *Theory:* Unlike standard approaches, our algorithm can leverage shared structure across products to achieve regret that scales sublinearly in the number of products N. In particular, we prove upper bounds $\widetilde{O}(\sqrt{NT})$ and $\widetilde{O}(N^{\frac{3}{4}}\sqrt{T})$ on the meta regret of `Meta-DP` and `Meta-DP++` respectively.

4. *Numerical Experiments:* We demonstrate on both synthetic and real auto loan data that our approach significantly speeds up learning compared to ignoring shared structure (*i.e.*, using prior-independent Thompson sampling) or greedily using the updated prior across products.

2. Problem Formulation

Notation: Throughout the paper, all vectors are column vectors by default. We define $[n]$ to be the set $\{1, 2, \ldots, n\}$ for any positive integer n. We use $\|\mathbf{x}\|_u$ to denote the ℓ_u norm of a vector $\mathbf{x}$, but we often omit the subscript when we refer to the ℓ_2 norm. For a positive definite matrix $A \in \Re^{d\times d}$ and vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $\|\mathbf{x}\|_A$ denote the matrix norm $\sqrt{\mathbf{x}^\top A\mathbf{x}}$ and $\langle \mathbf{x}, \mathbf{y}\rangle$ denote the inner product $\mathbf{x}^\top\mathbf{y}$. We also denote $x \vee y$ and $x \wedge y$ as the maximum and minimum between $x, y \in \mathbb{R}$, respectively. When logarithmic factors are omitted, we use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to denote function growth.

2.1. Model

We first describe the classical dynamic pricing formulation for a single product; we then formalize our meta-learning formulation over a sequence of N products.

Classical Formulation: Consider a seller who offers a single product over a selling horizon of T periods. The seller can dynamically adjust the offered price in each period. At the beginning of each period $t \in [T]$, the seller observes a random feature vector (capturing exogenous and/or customer-specific features) that is independently and identically distributed from an unknown distribution. Upon observing the feature vector, the seller chooses a price for that period. The seller then observes the resulting demand, which is a noisy function of both the observed feature vector and the chosen price. The seller's revenue in each period is given by the chosen price multiplied by the corresponding realized demand. The goal in this setting is to develop a policy π that maximizes the seller's cumulative revenue by balancing exploration (learning the demand function) with exploitation (offering the estimated revenue-maximizing price).

Meta-learning Formulation: We consider a seller who sequentially offers N related products, each with a selling horizon of T periods. For simplicity, a new product is not introduced until the life cycle of the previous product ends[1]. We call each product's life cycle an *epoch*, *i.e.*, there are N epochs that last T periods each. Each product (and corresponding epoch) is associated with a different (unknown) demand function, and constitutes a different instance of the classical dynamic pricing problem described above. We now formalize the problem.

[1] We model epochs as fully sequential for simplicity; if epochs overlap, we would need to additionally model a customer arrival process for each epoch. Our algorithms straightforwardly generalize for overlapping epochs; see remark in §4.4.

In epoch $i \in [N]$ at time $t \in [T]$, the seller observes a random feature vector $\mathbf{x}_{i,t} \in \mathbb{R}^d$, which is independently and identically distributed from an unknown distribution $\mathcal{P}_i$ (note that the distribution may vary across products/epochs). She then chooses a price $p_{i,t}$ for that period. Based on practical constraints, we will assume that the allowable price range is bounded across periods and products, *i.e.*, $p_{i,t} \in [p_{\min}, p_{\max}]$ and $0 < p_{\min} < p_{\max} < \infty$. The seller then observes the resulting induced demand

$$D_{i,t}(p_{i,t}, \mathbf{x}_{i,t}) = \langle \alpha_i, \mathbf{x}_{i,t} \rangle + p_{i,t} \langle \beta_i, \mathbf{x}_{i,t} \rangle + \varepsilon_{i,t}\,,$$

where $\alpha_i \in \mathbb{R}^d$ and $\beta_i \in \mathbb{R}^d$ are unknown fixed constants throughout epoch i, and $\varepsilon_{i,t}$ is zero-mean σ-subgaussian noise (see Definition 1 below). This demand model was recently proposed by Ban and Keskin (2017), and captures several salient aspects. In particular, the observed feature vector $\mathbf{x}_{i,t}$ in period t determines both the baseline demand (through the parameter α_i) and the price-elasticity of the demand (through the parameter β_i) of product i.

Definition 1. A random variable $z \in \mathbb{R}$ is σ-subgaussian if $\mathbb{E}[e^{tz}] \leq e^{\sigma^2 t^2/2}$ for every $t \in \mathbb{R}$.

This definition implies $\mathrm{Var}[z] \leq \sigma^2$. Many classical distributions are subgaussian; typical examples include any bounded, centered distribution, or the normal distribution. Note that the errors need not be identically distributed.

Shared Structure: For ease of notation, we define $\theta_i = [\alpha_i;\ \beta_i] \in \mathbb{R}^{2d}$; following the classical formulation of dynamic pricing, θ_i is the unknown parameter vector that must be learned within a given epoch in order for the seller to maximize her revenues over T periods. When there is no shared structure between the $\{\theta_i\}_{i=1}^N$, our problem reduces to N independent dynamic pricing problems.

However, we may expect that related products share a similar potential market, and thus may have some shared structure that can be learned across products. We model this relationship by positing that the product demand parameter vectors $\{\theta_i\}_{i=1}^N$ are independent and identically distributed draws from a common unknown distribution, *i.e.*, $\theta_i \sim \mathcal{N}(\theta_*, \Sigma_*)$ for each $i \in [N]$. As discussed earlier, knowledge of the distribution over the unknown demand parameters can inform the prior for Thompson sampling, thereby avoiding the need to use a conservative prior that can result in poor empirical performance (Honda and Takemura 2014, Liu and Li 2015). The mean of the shared distribution θ_* is unknown; we will consider settings where the covariance of this distribution Σ_* is known and unknown. We propose using meta-learning to learn this distribution from past epochs to inform and improve the current product's pricing strategy.

Remark 1. Following the literature on Thompson sampling, we consider a multivariate gaussian distribution since the posterior has a simple closed form, thereby admitting a tractable theoretical analysis. When implementing such an algorithm in practice, more complex distributions can be considered (*e.g.*, see discussion in Russo et al. 2018).

2.2. Background on Thompson Sampling with Known Prior

In this subsection, we consider the setting where the true prior $\mathcal{N}(\mu_*, \Sigma_*)$ over the unknown product demand parameters is *known*. This setting will inform our definition of the meta oracle and meta regret in the next subsection. When the prior is known, a natural candidate policy for minimizing Bayes regret is the Thompson sampling algorithm (Thompson 1933). The Thompson sampling algorithm adapted to our dynamic pricing setting for a single epoch $i \in [N]$ is formally given in Algorithm 1 below. Since the prior is known, there is no additional shared structure to exploit across products, so we can treat each epoch independently.

The algorithm begins with the true prior, and performs a single initialization period ($t = 1$). For each time $t \geq 2$, the Thompson sampling algorithm (1) samples the unknown product demand parameters $\mathring{\theta}_{i,t} = \left[\mathring{\alpha}_{i,t}; \mathring{\beta}_{i,t}\right]$ from the posterior $\mathcal{N}\left(\theta^{\mathrm{TS}}_{i,t}, \Sigma^{\mathrm{TS}}_{i,t}\right)$, and (2) solves and offers the resulting optimal price based on the demand function given by the sampled parameters

$$p^{\mathrm{TS}}_{i,t} = \underset{p \in [p_{\min}, p_{\max}]}{\arg\max} \; p \cdot \langle \mathring{\alpha}_{i,t}, \mathbf{x}_{i,t} \rangle + p^2 \cdot \left\langle \mathring{\beta}_{i,t}, \mathbf{x}_{i,t} \right\rangle . \tag{1}$$

Upon observing the actual realized demand $D_{i,t}\left(p^{\mathrm{TS}}_{i,t}, \mathbf{x}_{i,t}\right)$, the algorithm computes the posterior $\mathcal{N}\left(\theta^{\mathrm{TS}}_{i,t+1}, \Sigma^{\mathrm{TS}}_{i,t+1}\right)$ for round $t+1$. The same algorithm is applied independently to each epoch $i \in [N]$.

Algorithm 1 Thompson Sampling Algorithm

Input: The prior mean vector μ_* and covariance matrix Σ_*, the index i of epoch, the length of each epoch T, the subgaussian parameter σ.
Initialization: $\left(\theta^{\mathrm{TS}}_{i,1}, \Sigma^{\mathrm{TS}}_{i,1}\right) \leftarrow (\theta_*, \Sigma_*)$.
Observe feature vector $\mathbf{x}_{i,1}$, and set initial price $p_{i,1} \leftarrow \begin{cases} p_{\min} & \text{if } i \text{ is even,} \\ p_{\max} & \text{otherwise.} \end{cases}$
Observe demand $D_{i,1}(p_{i,1}, \mathbf{x}_{i,1})$, and compute the posterior $\mathcal{N}\left(\theta^{\mathrm{TS}}_{i,2}, \Sigma^{\mathrm{TS}}_{i,2}\right)$.
for $t = 2, \ldots, T$ **do**
    Observe feature vector $\mathbf{x}_{i,t}$.
    Sample parameter $\mathring{\theta}_{i,t} \leftarrow \left[\mathring{\alpha}_{i,t}; \mathring{\beta}_{i,t}\right] \sim \mathcal{N}\left(\theta^{\mathrm{TS}}_{i,t}, \Sigma^{\mathrm{TS}}_{i,t}\right)$.
    $p^{\mathrm{TS}}_{i,t} \leftarrow \arg\max_{p \in [p_{\min}, p_{\max}]} \; p \cdot \langle \mathring{\alpha}_{i,t}, \mathbf{x}_{i,t} \rangle + p^2 \cdot \left\langle \mathring{\beta}_{i,t}, \mathbf{x}_{i,t} \right\rangle$.
    Observe demand $D_{i,t}\left(p^{\mathrm{TS}}_{i,t}, \mathbf{x}_i\right)$, and compute the posterior $\mathcal{N}\left(\theta^{\mathrm{TS}}_{i,t+1}, \Sigma^{\mathrm{TS}}_{i,t+1}\right)$.
end for

As evidenced by the large literature on the practical success of Thompson sampling (Chapelle and Li 2011, Russo and Van Roy 2014, Ferreira et al. 2018), Algorithm 1 is a very attractive choice for implementation in practice.

It is worth noting that Algorithm 1 attains a strong performance guarantee under the classical formulation compared to a the classical *oracle* that knows all N product demand parameters $\{\theta_i\}_{i=1}^N$

in advance. In particular, this oracle would offer the expected optimal price in each period $t \in [T]$ in epoch $i \in [N]$, *i.e.*,

$$\begin{aligned} p_{i,t}^* &= \arg\max_{p \in [p_{\min}, p_{\max}]} \; p \cdot \mathbb{E}_\varepsilon[D_{i,t}(p, \mathbf{x}_{i,t})] \\ &= \arg\max_{p \in [p_{\min}, p_{\max}]} \; p\langle \alpha_i, \mathbf{x}_{i,t} \rangle + p^2 \langle \beta_i, \mathbf{x}_{i,t} \rangle \,. \end{aligned}$$

The resulting *Bayes regret* (Russo and Van Roy 2014) of a given policy π relative to the oracle is defined as:

$$\text{Bayes Regret}_{N,T}(\pi) = \mathbb{E}_{\theta, \mathbf{x}, \varepsilon} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} p_{i,t}^* D(p_{i,t}^*, \mathbf{x}_{i,t}) - \sum_{i=1}^{N} \sum_{t=1}^{T} p_{i,t}^\pi D(p_{i,t}^\pi, \mathbf{x}_{i,t}) \,, \right] \tag{2}$$

where the expectation is taken with respect to the unknown product demand parameters, the observed random feature vectors, and the noise in the realized demand. The following theorem bounds the Bayes regret of the Thompson sampling dynamic pricing algorithm:

THEOREM 1. *The Bayes regret of Algorithm 1 satisfies*

$$\text{Bayes Regret}_{N,T}(\pi) = \widetilde{O}\left(dN\sqrt{T} \right) ,$$

when the prior over the product demand parameters is known.

Theorem 1 follows from a similar argument used for the linear bandit setting presented in Russo and Van Roy (2014), coupled with standard concentration bounds for multivariate normal distributions. The proof is given in Appendix A for completeness. Note that the regret scales linearly in N, since each epoch is an independent learning problem.

REMARK 2. Prior-independent Thompson sampling (Agrawal and Goyal 2013) achieves the same overall Bayes regret as Algorithm 1. However, we document a substantial gap in empirical performance between the two approaches in §5, motivating our study of learning the prior.

2.3. Meta Oracle and Meta Regret

We cannot directly implement Algorithm 1 in our setting, since the prior over the product demand parameters $\mathcal{N}(\theta_*, \Sigma_*)$ is unknown. In this paper, we seek to learn the prior (shared structure) *across* products in order to leverage the superior performance of Thompson sampling with a known prior. Thus, a natural question to ask is:

What is the price of not knowing the prior in advance?

To answer this question, we first define our performance metric. Since our goal is to converge to the policy given in Algorithm 1 (which knows the true prior), we define this policy as our *meta*

oracle[2]. Comparing the revenue of our policy relative to the meta oracle leads naturally to the definition of *meta regret* $\mathcal{R}_{N,T}$ for a policy π, *i.e.*,

$$\mathcal{R}_{N,T}(\pi) = \mathbb{E}_{\theta,\mathbf{x},\varepsilon}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} p_{i,t}^{\mathrm{TS}} D(p_{i,t}^{\mathrm{TS}}, \mathbf{x}_{i,t}) - \sum_{i=1}^{N}\sum_{t=1}^{T} p_{i,t}^{\pi} D(p_{i,t}^{\pi}, \mathbf{x}_{i,t})\right],$$

where the expectation is taken with respect to the unknown product demand parameters, the observed random feature vectors, and the noise in the realized demand.

Note that prior-independent Thompson sampling and UCB treat each epoch independently, and would thus achieve meta regret that grows linearly in N. Our goal is to design a policy with meta regret that grows sublinearly in N and at most linearly in $\sqrt{T}$. Recall that Theorem 1 bounds the Bayes regret of Thompson sampling with a known prior as $\widetilde{O}\left(N\sqrt{T}\right)$. Thus, if our meta regret (*i.e.*, the performance of our meta-learning policy relative to Algorithm 1) grows sublinearly in N (and no faster than $\sqrt{T}$), it would imply that the price of not knowing the prior $\mathcal{N}(\theta_*, \Sigma_*)$ in advance is negligible in experiment-rich environments (*i.e.*, as N grows large) compared to the cost of learning the actual demand parameters for each product (*i.e.*, the Bayes regret of Algorithm 1).

Non-anticipating Policies: We restrict ourselves to the family of non-anticipating policies $\Pi : \pi = \{\pi_{i,t}\}$ that form a sequence of random functions $\pi_{i,t}$ that depend only on price and demand observations collected until time t in epoch i (including all times $t \in [T]$ from prior epochs), and feature vector observations up to time $t+1$ in epoch i. In particular, let $\mathcal{H}_{0,0} = (\mathbf{x}_{1,1})$, and $\mathcal{H}_{i,t} = (p_{1,1}, p_{1,2}, \cdots, p_{i,t}, D_{1,1}, D_{1,2}, \cdots, D_{i,t}, \mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \cdots, \mathbf{x}_{i,t+1})$ denote the history of prices and corresponding demand realizations from prior epochs and time periods, as well as the observed feature vectors up to the next time period; let $\mathcal{F}_{i,t}$ denote the σ-field generated by $\mathcal{H}_{i,t}$. Then, we impose that $\pi_{i,t+1}$ is $\mathcal{F}_{i,t}$ measurable.

The values of the prior mean θ_* as well as the actual product demand parameter vectors $\{\theta_i\}_{i=1}^N$ are unknown; we consider two settings — known and unknown Σ_* (covariance of the prior).

2.4. Assumptions

We now describe some mild assumptions on the parameters of the problem for our regret analysis.

Assumption 1 **(Boundedness)**. *The support of the features are bounded, i.e.,*

$$\forall i \in [N], \forall t \in [T] \quad \|\mathbf{x}_{i,t}\| \le x_{\max}.$$

Furthermore, there exists a positive constant S such that $\|\theta_*\| \le S$.

[2] We use the term meta oracle to distinguish from the oracle in the classical formulation.

Our first assumption is that the observed feature vectors $\{\mathbf{x}_{i,t}\}$ as well as the mean of the product demand parameters θ_* are bounded. This is a standard assumption made in the bandit and dynamic pricing literature, ensuring that the average regret at any time step is bounded. This is likely satisfied since features and outcomes are typically bounded in practice.

ASSUMPTION 2 **(Positive-Definite Feature Covariance)**. *The minimum eigenvalue of the feature covariance matrix* $\mathbb{E}_{\mathbf{x}_{i,t}\sim\mathcal{P}_i}\left[\mathbf{x}_{i,t}\mathbf{x}_{i,t}^\top\right]$ *in every epoch* $i \in [N]$ *is lower bounded by some positive constant* λ_0*, i.e.,*

$$\min_{i\in[N]} \lambda_{\min}\left(\mathbb{E}_{\mathbf{x}_{i,t}\sim\mathcal{P}_i}\left[\mathbf{x}_{i,t}\mathbf{x}_{i,t}^\top\right]\right) \geq \lambda_0 .$$

Our second assumption imposes that the covariance matrix of the observed feature vectors $\mathbb{E}\left[\mathbf{x}_{i,t}\mathbf{x}_{i,t}^\top\right]$ in every epoch is positive-definite. This is a standard assumption for the convergence of OLS estimators; in particular, our demand model is linear, and therefore requires that no features are perfectly collinear in order to identify each product's true demand parameters.

ASSUMPTION 3 **(Positive-Definite Prior Covariance)**. *The maximum and minimum eigenvalues of* Σ_* *are upper and lower bounded by positive constants* $\overline{\lambda}$ *and* $\underline{\lambda}$*, respectively i.e.,*

$$\lambda_{\max}(\Sigma_*) \leq \overline{\lambda}, \quad \lambda_{\min}(\Sigma_*) \geq \underline{\lambda}.$$

We further assume that the trace of Σ_* *is upper bounded by* κ*, i.e.,* $\mathrm{tr}(\Sigma_*) \leq \kappa$*.*

Our final assumption imposes that the covariance matrix of the random product demand parameter θ is also positive-definite. Again, this assumption ensures that each product's true demand parameter is identifiable using standard OLS estimators.

3. `Meta-DP` Algorithm

We begin with the case where the prior's covariance matrix Σ_* is known, and describe the Meta Dynamic Pricing (`Meta-DP`) algorithm for this setting. We will consider the case of unknown Σ_* in the next section.

3.1. Overview

The `Meta-DP` algorithm begins by using initial product epochs as an exploration phase to initialize our estimate of the prior mean θ_*. These exploration epochs use the prior-independent UCB algorithm to ensure no more than $\widetilde{O}(\sqrt{T})$ meta regret for each epoch. After this initial exploration period, our algorithm leverages the estimated prior within each subsequent epoch, and continues to sequentially update the estimated prior after each epoch. The key challenge is that the estimated prior has finite-sample estimation error, and can thus result in poor performance within a given

epoch. At the same time, we can no longer employ a prior-independent approach, since this will cause our meta regret to grow linearly in N. Our algorithm addresses this challenge by carefully widening the covariance of the prior (beyond the known covariance Σ_*) within each epoch by a term that scales as the expected error of the estimated θ_*. This correction approaches zero as N grows large, ensuring that our meta regret grows sublinearly in N.

3.2. Algorithm

The `Meta-DP` algorithm is presented in Algorithm 2. We first define some additional notation, and then describe the algorithm in detail.

Additional Notation: Throughout the rest of the paper, we use $\boldsymbol{m}_{i,t} = \left(\mathbf{x}_{i,t},\ p_{i,t}\mathbf{x}_{i,t}\right)^\top$ to denote the price and feature information of round t in epoch i for all $i \in [N]$ and $t \in [T]$. We also define the following quantities for each epoch $i \in [N]$:

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{1,1} & \cdots & \mathbf{x}_{i,1} \\ p_{1,1}\mathbf{x}_{1,1} & \cdots & p_{i,1}\mathbf{x}_{i,1} \end{pmatrix}, \quad \mathbf{D}_i = \begin{pmatrix} D_{1,1}(p_{i,1}, \mathbf{x}_{i,1}) \\ \vdots \\ D_{i,1}(p_{i,1}, \mathbf{x}_{i,1}) \end{pmatrix}. \tag{3}$$

$\mathbf{X}_i$ is the price and feature design matrix, and $\mathbf{D}_i$ is the corresponding vector of realized demands from all initialization steps ($t = 1$) in epochs $\{1, \cdots, i\}$.

Algorithm Description: The first N_0 epochs are treated as exploration epochs, where we define

$$N_0 = \max\left\{ \frac{2\log_{e/2}(2dNT)}{c_1},\ d^2,\ \left[c_2 \log_e\left(2^d N^2 T\right)\right]^2 \right\} \tag{4}$$

and the constants are given by

$$c_0 = \frac{1}{3} \min_{\|z_1\|^2 + \|z_2\|^2 = 1} \left[\left(p_{\min}\|z_2\| - \|z_1\|\right)^2 + \left(p_{\max}\|z_2\| - \|z_1\|\right)^2 \right],$$
$$c_1 = \frac{c_0 \lambda_0}{\sqrt{(1 + p_{\max}^2)} x_{\max}}, \quad \text{and} \quad c_2 = \frac{4\left(x_{\max}^2 \overline{\lambda}(1 + p_{\max}^2) + \sigma^2\right)}{\underline{\lambda} c_0 \lambda_0}.$$

As described in the overview, the `Meta-DP` algorithm proceeds in two phases. In particular, we distinguish the following two cases for all $t \geq 2$ (similar to Algorithm 1, the first period $t = 1$ of each epoch is reserved for initialization):

1. **Epoch $\mathbf{i < N_0}$**: the `Meta-DP` algorithm runs the prior-independent UCB algorithm proposed by Abbasi-Yadkori et al. (2011) for the rest of the epoch. In particular, for each $t \geq 2$, we construct the UCB estimate $\theta_{i,t}^{\mathrm{UCB}}$ using the regularized least square estimator on the price and feature data, and the corresponding demands observed so far, *i.e.*,

$$\theta_{i,t}^{\mathrm{UCB}} = \begin{pmatrix} \alpha_{i,t}^{\mathrm{UCB}} \\ \beta_{i,t}^{\mathrm{UCB}} \end{pmatrix} = \left(\sum_{\tau=1}^{t-1} \boldsymbol{m}_{i,\tau} \boldsymbol{m}_{i,\tau}^\top + I_{2d} \right)^{-1} \left(\sum_{\tau=1}^{t-1} D_{i,\tau}\left(p_{i,\tau}, \mathbf{x}_{i,\tau}\right) \boldsymbol{m}_i \right). \tag{5}$$

The `Meta-DP` algorithm then offers the price with the largest upper confidence bound, *i.e.*,

$$p_{i,t} = \underset{p\in[p_{\min},p_{\max}]}{\arg\max}\; p\left\langle \alpha_{i,t}^{\mathrm{UCB}}, \mathbf{x}_{i,t}\right\rangle + p^2\left\langle \beta_{i,t}^{\mathrm{UCB}}, \mathbf{x}_{i,t}\right\rangle + \left\| \begin{pmatrix} \mathbf{x}_{i,t} \\ p\mathbf{x}_{i,t} \end{pmatrix} \right\|_{\left(\sum_{\tau=1}^{t-1} \boldsymbol{m}_{i,\tau}\boldsymbol{m}_{i,\tau}^\top + I_{2d}\right)^{-1}}, \tag{6}$$

and observes the realized demand $D_{i,t}\left(p_{i,t}, \mathbf{x}_{i,t}\right)$.

2. **Epoch $\mathbf{i \geq N_0}$** : the `Meta-DP` algorithm utilizes the data collected from the initialization step of all past epochs and the current epoch to compute our estimate $\hat{\theta}_i$ of the prior mean θ_*. We use the ordinary least square estimator, *i.e.*,

$$\hat{\theta}_i = \left(\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\mathbf{X}_i\mathbf{D}_i. \tag{7}$$

However, as noted earlier, using the estimated prior directly can cause Thompson sampling to fail due to finite-sample estimation error. Thus, we widen the prior by increasing the covariance beyond Σ_*. In particular, we set the prior as follows:

$$\mathcal{N}\left(\theta_{i,2}^{\mathrm{MPDP}}, \Sigma_{i,2}^{\mathrm{MPDP}}\right) = \mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right) \tag{8}$$

$$\Sigma_i = \eta_i\Sigma_* = \left(1 + \frac{1}{\sqrt{i}}\right)\Sigma_*. \tag{9}$$

Note that the extent of prior widening approaches zero for later epochs (*i.e.*, i large), when we expect the estimation error of the prior mean to be small.

Next, the `Meta-DP` algorithm follows the TS algorithm armed with the widened prior $\mathcal{N}\left(\theta_{i,2}^{\mathrm{MPDP}}, \Sigma_{i,2}^{\mathrm{MPDP}}\right)$. In particular, for each time step $t \geq 2$, we (1) sample the unknown product demand parameters $\mathring{\theta}_{i,t} = \left[\mathring{\alpha}_{i,t}; \mathring{\beta}_{i,t}\right]$ from the posterior $\mathcal{N}\left(\theta_{i,t}^{\mathrm{MPDP}}, \Sigma_{i,t}^{\mathrm{MPDP}}\right)$, and (2) solve and offer the resulting optimal price based on the demand function given by the sampled parameters

$$p_{i,t} = \underset{p\in[p_{\min},p_{\max}]}{\arg\max}\; p\left\langle \mathring{\alpha}_{i,t}, \mathbf{x}_{i,t}\right\rangle + p^2\left\langle \mathring{\beta}_{i,t}, \mathbf{x}_{i,t}\right\rangle. \tag{10}$$

Upon observing the actual realized demand $D_{i,t}\left(p_{i,t}, \mathbf{x}_{i,t}\right)$ at the end of the time step, we compute the posterior $\mathcal{N}\left(\mu_{i,t+1}^{\mathrm{MPDP}}, \Sigma_{i,t+1}^{\mathrm{MPDP}}\right)$ for the next time step $t+1$.

3.3. Meta Regret Analysis

We now prove an upper bound on the meta regret of the `Meta-DP` algorithm.

We begin by noting that the prior-independent UCB algorithm employed in the exploration epochs satisfies a meta regret guarantee:

LEMMA 1. *The meta regret of the UCB algorithm in a single epoch is* $\widetilde{O}(d\sqrt{T})$.

The proof of this result is essentially the same as that of Theorem 1, and is thus omitted. Lemma 1 ensures that we accrue at most $\widetilde{O}(dN_0\sqrt{T})$ regret in the N_0 exploration epochs; from Eq. (4), we know that N_0 grows merely poly-logarithmically in N and T.

Algorithm 2 Meta-Personalized Dynamic Pricing Algorithm

Input: The prior covariance matrix Σ_*, the total number of epochs N, the length of each epoch T, the upper bound on the prior mean S, the subgaussian parameter σ, and the set of feasible prices $[p_{\min}, p_{\max}]$.
Initialization: N_0 as defined in eq. (4).
for each epoch $i = 1, \ldots, N$ **do**
    Observe feature vector $\mathbf{x}_{i,1}$, and set initial price $p_{i,1} \leftarrow \begin{cases} p_{\min} & \text{if } i \text{ is even,} \\ p_{\max} & \text{otherwise.} \end{cases}$
    Observe initial demand $D_{i,1}(p_{i,1}, \mathbf{x}_{i,1})$.
    if $i < N_0$ **then**
        for $t = 2, \ldots, T$ **do**
            Observe feature vector $\mathbf{x}_{i,t}$ and update $\theta_{i,t}^{\mathrm{UCB}}$ according to eq. (5)
            Choose price $p_{i,t}$ according to eq. (6), and observe demand $D_{i,t}(p_{i,t}, \mathbf{x}_{i,t})$.
        end for
    else
        Update $\hat{\theta}_i$ according to eq. (7), and set $\eta_i \leftarrow 1 + 1/\sqrt{i}$, $\Sigma_i \leftarrow \eta_i \Sigma_*$.
        Construct prior $\mathcal{N}\left(\theta_{i,2}^{\mathrm{MPDP}}, \Sigma_{i,2}^{\mathrm{MPDP}}\right) \leftarrow \mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)$.
        for $t = 2, \ldots, T$ **do**
            Observe feature vector $\mathbf{x}_{i,t}$, and sample parameter $\mathring{\theta}_{i,t} \sim \mathcal{N}\left(\theta_{i,t}^{\mathrm{MPDP}}, \Sigma_{i,t}^{\mathrm{MPDP}}\right)$.
            Choose price $p_{i,t}$ according to eq. (10), observe demand $D_{i,t}(p_{i,t}, \mathbf{x}_{i,t})$, and compute the posterior $\mathcal{N}\left(\theta_{i,t+1}^{\mathrm{MPDP}}, \Sigma_{i,t+1}^{\mathrm{MPDP}}\right)$.
        end for
    end if
end for

Next, after the exploration epochs conclude, we begin using the estimated prior mean, which we greedily update at the end of each subsequent epoch. The following theorem bounds the error of this estimate with high probability:

THEOREM 2. *For any fixed $i \geq 2$, with probability at least $1 - \delta - 2d\left(\frac{\exp(-\zeta)}{(1-\zeta)^{1-\zeta}}\right)^{c_1 i}$, the ℓ_2 distance between $\hat{\theta}_i$ and θ_* satisfies*

$$\left\|\hat{\theta}_i - \theta_*\right\| \leq \frac{2R\sqrt{d\log_e 2 - \log_e \delta}}{\sqrt{(1-\zeta)c_0\lambda_0 i}},$$

where c_0 and c_1 are constants that depends only on λ_0, $p_{\min}$, $p_{\max}$, and $x_{\max}$.

Proof Sketch. The complete proof is provided in Appendix B. Let $M_i = [\mathbf{x}_{i,1};\ p_{i,1}\mathbf{x}_{i,1}]$ be the initial feature and price vector of the first round of each epoch i. Then, for an epoch $i \in [N]$, the initial demand realization satisfies

$$\begin{aligned} D_{i,1} &= \langle \theta_i, M_i \rangle + \varepsilon_{i,1} \\ &= \langle \theta_*, M_i \rangle + \langle \Delta_i, M_i \rangle + \varepsilon_{i,1}, \end{aligned}$$

where $\Delta_i \sim \mathcal{N}(\mathbf{0}, \Sigma_*)$. Note that M_i is an independent random variable across different epochs, since the feature vectors $\mathbf{x}_{i,1}$ are drawn i.i.d. from $\mathcal{P}_i$, and the prices alternate between $p_{\min}$ and

$p_{\max}$ by construction. Thus, we can equivalently view the demand realization as the mean demand $\langle \theta_*, M_i \rangle$ corrupted by the price dependent (or heteroscedastic) noise $\langle \Delta_i, M_i \rangle + \varepsilon_{i,1}$. It can be verified that $\langle \Delta_i, M_i \rangle + \varepsilon_{i,1}$ is R-subgaussian with $R = \sqrt{x_{\max}^2 \overline{\lambda}(1+p_{\max}^2) + \sigma^2}$.

Next, we can bound the difference between $\widehat{\theta}$ and θ with high probability, *i.e.*,

$$\Pr\left(\|\widehat{\theta} - \theta\| \geq \frac{2R\sqrt{d \log_e 2 - \log_e \delta}}{\lambda_{\min}(V_i)} \right) \leq \delta \,.$$

Thus, it suffices to lower bound the smallest eigenvalue of V_i to ensure that $\widehat{\theta}$ is close to θ. To this end, we employ matrix Chernoff bounds by Tropp (2011). First, we show that there exists a positive constant c_0 (that depends only on λ_0, $p_{\min}$, and $p_{\max}$) such that the minimum eigenvalue of the expectation $\mathbb{E}[V_i] = \sum_{\iota=1}^{i} \mathbb{E}\left[M_\iota M_\iota^\top \middle| \mathcal{F}_{(\iota-1),T} \right]$ is lower bounded by $c_0 \lambda_0 i$, *i.e.*,

$$\lambda_{\min}\left(\sum_{\iota=1}^{i} \mathbb{E}\left[M_\iota M_\iota^\top \middle| \mathcal{F}_{(\iota-1),T} \right] \right) \geq c_0 \lambda_0 i.$$

We apply the matrix Chernoff inequality (Tropp 2011) to provide a high probability lower bound on the minimum eigenvalue of the random matrix V_i, *i.e.*,

$$\Pr\left[\lambda_{\min}(V_i) \geq (1-\zeta) c_0 \lambda_0 i \right] \;\geq\; 1 - 2d \left(\frac{\exp(-\zeta)}{(1-\zeta)^{1-\zeta}} \right)^{c_1 i},$$

Finally, by a simple union bound, we conclude the proof. □

We now state our main result upper bounding the meta regret of the `Meta-DP` algorithm.

THEOREM 3. *If the number of products is at least $N = \widetilde{\Omega}(d^2)$, then the meta regret of the proposed `Meta-DP` algorithm satisfies*

$$\mathcal{R}_{N,T}(\textit{Meta-DP algorithm}) = \tilde{O}\left(d^2 \sqrt{NT} \right).$$

Proof Sketch. The complete proof is provided in Appendix C.

We begin by defining some helpful notation. First, let $\texttt{REV}\left(\theta, \hat{\theta}, \Sigma\right)$ be the expected revenue obtained by running the Thompson sampling algorithm in Algorithm 1 with the (possibly incorrect) prior $\mathcal{N}\left(\hat{\theta}, \Sigma\right)$ after initialization in an epoch whose true parameter is θ. Second, let $\texttt{REV}_*(\theta)$ be the maximum expected revenue that can be obtained from an epoch parametrized by θ after initialization. We also define the clean event $\mathcal{E}$ over all non-exploration epochs:

$$\forall i \geq N_0 \qquad \left\| \hat{\theta}_i - \theta_* \right\| \leq \frac{2R\sqrt{2d \log_e 2 + 2\log_e (N^2 T)}}{\sqrt{c_0 \lambda_0 i}}.$$

When $\mathcal{E}$ holds, our estimate of the prior mean has bounded error from the true prior mean in all non-exploration epochs. Theorem 2 implies that $\mathcal{E}$ holds with probability at least $1 - \frac{9}{NT}$. Note that the meta regret over non-exploration epochs is trivially bounded by $O(NT)$. Then, the cumulative

contribution to the expected meta regret when the clean event $\mathcal{E}$ is violated is $O(NT)\Pr(\neg\mathcal{E}) = O(1)$. We then proceed to analyze the regret of each epoch conditioned on the clean event $\mathcal{E}$. For an epoch $i \geq N_0$, the expected meta regret $\mathcal{R}_{N,T}(i)$ of this epoch can be written as

$$\mathcal{R}_{N,T}(i) = \mathbb{E}_{\theta_i}\mathbb{E}_{\hat{\theta}_i}\left[\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i, \hat{\theta}_i, \Sigma_i\right)\right] - \mathbb{E}_{\theta_i}\left[\left(\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i, \theta_*, \Sigma_*\right)\right)\right].$$

Now, from Section 3 of Russo and Van Roy (2014), we upper bound the first term as

$$\mathbb{E}_{\theta_i}\mathbb{E}_{\hat{\theta}_i}\left[\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i, \hat{\theta}_i, \Sigma_i\right)\right] \leq \mathbb{E}_{\hat{\theta}_i}\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} \mathbb{E}_{\theta_i}\left[\left(\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i, \theta_*, \Sigma_*\right)\right)\right],$$

where $\frac{d\mathcal{N}(\theta_*, \Sigma_*)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}$ is the Radon-Nikodym derivative of $\mathcal{N}\left(\theta_*, \Sigma_*\right)$ with respect to $\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)$, and $\|\cdot\|_{\mathcal{N}\left(\bar{\theta}_n, \Sigma_n\right), \infty}$ is the essential supremum magnitude with respect to $\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)$. Therefore,

$$\mathcal{R}_{N,T}(i) \leq \mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1\right] \mathbb{E}_{\theta_i}\left[\left(\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i, \theta_*, \Sigma_*\right)\right)\right],$$

and, by applying Theorem 1, the total meta regret can be upper bounded as

$$\mathcal{R}_{N,T} \leq \sum_{i=1}^{N_0-1} \mathcal{R}_{N,T}(i) + \sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1\right] \widetilde{O}\left(d\sqrt{T}\right). \tag{11}$$

The first term in (11) is simply the regret accrued by UCB in the first N_0 exploration epochs. Applying Lemma 1 and the definition of N_0 from Eq. (4), we can bound this term as

$$\sum_{i=1}^{N_0-1} \mathcal{R}_{N,T}(i) = \widetilde{O}\left(dN_0\sqrt{T}\right) = \widetilde{O}\left(d^3\sqrt{T}\right) = \widetilde{O}\left(d^2\sqrt{NT}\right).$$

For the second term in (11), we use the definition of the multivariate normal to compute

$$\begin{aligned}
&\sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1\right] \\
&= \sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i}\left[\sup_{\theta} \frac{\det\left(2\pi\Sigma_*\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\theta-\theta_*)^\top \Sigma_*^{-1}(\theta-\theta_*)\right)}{\det\left(2\pi\Sigma_i\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\theta-\hat{\theta}_i)^\top \Sigma_i^{-1}(\theta-\hat{\theta}_i)\right)} - 1\right] \\
&= \sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d \sup_{\theta} \exp\left(\frac{\frac{\eta_i}{\eta_i - 1}\Delta_i^\top \Sigma_*^{-1}\Delta_i - (\eta_i - 1)\left(\theta - \theta_* - \frac{\Delta_i}{\eta_i - 1}\right)^\top \Sigma_*^{-1}\left(\theta - \theta_* - \frac{\Delta_i}{\eta_i - 1}\right)}{2\eta_i}\right) - 1\right].
\end{aligned} \tag{12}$$

Since we have assumed that Σ_* is positive definite, it follows that Σ_*^{-1} is positive definite as well. Recalling that $\eta_i = 1 + 1/\sqrt{i} > 1$, note that

$$-(\eta_i - 1)\left(\theta - \theta_* - \frac{\Delta_i}{\eta_i - 1}\right)^\top \Sigma_*^{-1}\left(\theta - \theta_* - \frac{\Delta_i}{\eta_i - 1}\right) \;\leq\; 0\,.$$

Furthermore, since we have conditioned on the clean event $\mathcal{E}$, Eq. (12) does not exceed

$$\sum_{i=N_0}^{N}\left[\left(1+\frac{1}{\sqrt{i}}\right)^d \exp\left(\frac{c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}}\right) - 1\right].$$

Finally, using the identity that $(1+1/a)^a \leq e$ for all $a > 0$, we can simplify

$$\left(1+\frac{1}{\sqrt{i}}\right)^d = \left(\left(1+\frac{1}{\sqrt{i}}\right)^{\sqrt{i}}\right)^{\frac{d}{\sqrt{i}}} \leq \exp\left(\frac{d}{\sqrt{i}}\right).$$

Therefore, the second term in (11) can be bounded as

$$\sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1\right] \leq \sum_{i=N_0}^{N}\left[\exp\left(\frac{d + c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}}\right) - 1\right].$$

By definition of N_0 in Eq. (4), we can write $\sqrt{i} \geq d + c_2 \log_e\left(2^d N^2 T\right)$. Using the identity that $\exp(a) \leq 1 + 2a$ for any $a \in [0,1]$, it follows that

$$\begin{aligned}
\sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1\right] &\leq \sum_{i=N_0}^{N}\left[\left(1 + \frac{2d + 2c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}}\right) - 1\right] \\
&\leq \sum_{i=N_0}^{N} \frac{2d + 2c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}} \\
&= \widetilde{O}(d\sqrt{N}).
\end{aligned}$$

Combining the expressions above yields the result. □

REMARK 3. Note that if we are in the regime where $N \lesssim N_0$ prescribed by Eq. (4), then the decision-maker can choose N_0 to instead be

$$N_0 = \max\left\{\frac{\log_{e/2}\left(2dNT\right)}{c_1},\; d^2,\; \left[\frac{c_2}{\rho}\log_e\left(2^d N^2 T\right)\right]^2\right\}$$

and set $\eta_i = 1 + \rho/\sqrt{i}$ for any choice of $\rho \geq 1$, without affecting the theoretical guarantee stated in Theorem 3. In other words, we can trade off the number of exploration epochs (N_0) with the extent of prior widening (η_i) in non-exploration epochs.

3.4. Prior Widening

We now pause to comment on the necessity of our prior widening technique. An immediate and tempting alternative to the `Meta-DP` algorithm is the the following "greedy" algorithm: it is identical to the `Meta-DP` algorithm, but in each non-exploration epoch ($i \geq N_0$), the greedy approach uses the updated prior directly without any prior widening, *i.e.*, setting $\eta_i = 1$ for all $i \geq N_0$ in Algorithm 2. In other words, after the initial exploration epochs, the algorithm greedily applies Thompson sampling with the current estimated prior (which is updated at the end of every epoch) in each subsequent epoch.

However, the estimated prior naturally has finite-sample estimation error. Empirical evidence from Lattimore and Szepesvári (2018) shows that even a small misspecification in the prior can lead to significant performance degradation of the Thompson Sampling algorithm. This raises the concern that the simple greedy approach may fail to perform well in some epochs due to estimation error. In Section 5, we compare the performance of the greedy approach described above to our proposed approach on a range of numerical experiments on both synthetic and real auto loan data. We consistently find that our proposed approach performs better, suggesting that prior widening is in fact necessary. In what follows, we provide intuition from our theoretical analysis on why the greedy approach may fail, and explain how prior widening helps overcome this challenge.

Consider inequality (11) in the proof sketch of Theorem 3. When applied to the greedy approach, the upper bound for the meta regret becomes

$$\mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_*\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_*\right), \infty} - 1\right] \widetilde{O}\left(d\sqrt{T}\right). \tag{13}$$

Following the same steps as in Eq. (12), we can write

$$\left\|\frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_*\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_*\right), \infty} = \sup_{\theta} \ \exp\left((\theta_* - \hat{\theta}_i)^\top \Sigma_*^{-1}(\theta - \theta_*) + \frac{1}{2}(\theta_* - \hat{\theta}_i)^\top \Sigma_*^{-1}(\theta_* - \hat{\theta}_i)\right). \tag{14}$$

Suppose we take θ to be the form $\hat{\theta}_i + \nu(\hat{\theta}_i - \theta_*)$ for some $\nu \in \Re$, then Eq. (14) becomes

$$\sup_{\nu} \ \exp\left(-(\nu+1)(\theta_* - \hat{\theta}_i)^\top \Sigma_*^{-1}(\theta_* - \hat{\theta}_i) + \frac{1}{2}(\theta_* - \hat{\theta}_i)^\top \Sigma_*^{-1}(\theta_* - \hat{\theta}_i)\right).$$

Note that Σ_*^{-1} is positive definite, so the quadratic form $(\theta_* - \hat{\theta}_i)^\top \Sigma_*^{-1}(\theta_* - \hat{\theta}_i)$ is positive as long as $\hat{\theta}_i \neq \theta_*$, *i.e.*, there exists *any* estimation error in $\hat{\theta}_i$. It is thus easy to verify that as $\nu \to \infty$, $\left\|\frac{d\mathcal{N}(\theta_*, \Sigma_*)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_*\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_*\right), \infty} \to \infty$ as well. This suggests that for *some* realizations of θ_i, the Thompson algorithm with the greedy prior estimate can fail to converge and achieve worst-case performance. In contrast, by widening the prior, we ensure that this term is bounded above with high probability

(see Eq. (12)), thereby ensuring convergence within every epoch. The `Meta-DP` algorithm provides an exact prior correction path over time to ensure low meta regret in every non-exploration epoch.

We note that the above argument simply indicates that the same analysis of the `Meta-DP` algorithm cannot be applied to the greedy approach; we were unable to prove a lower bound that a greedy approach that does not apply prior widening achieves poor meta regret. This is because the cost of prior misspecification in Thompson sampling is difficult to characterize in general (see, *e.g.*, Honda and Takemura 2014, Liu and Li 2015, for analogous results in very simplified settings.) Thus, although it is clear that a greedy approach can employ a prior that does not place sufficient weight on the true parameter of interest (due to finite-sample estimation error), it is unclear how this affects the resulting regret. However, the empirical evidence from Lattimore and Szepesvári (2018) and our numerical experiments in Section 5 together suggest that the greedy approach in fact performs poorly. We believe this is an interesting direction for future research.

4. `Meta-DP++` Algorithm

In this section, we consider the setting where the prior covariance matrix Σ_* is also unknown. We propose the `Meta-DP++` algorithm, which builds on top of the `Meta-DP` algorithm and additionally estimates the unknown prior covariance Σ_*.

4.1. Overview

The key challenge compared to the previous section is that the `Meta-DP` algorithm required only the *initial* samples from each epoch to estimate the unknown prior mean θ_*. In particular, when Σ_* was known, the algorithm did not need to recover the actual unknown product demand parameters $\{\theta_i\}$ across epochs to estimate the prior. However, when Σ_* is unknown, we will need to estimate the unknown product parameters to an acceptable degree of accuracy for at least some epochs. Therefore, the `Meta-DP++` algorithm additionally performs random price exploration for several time steps (instead of prior-independent UCB) in the initial exploration epochs to collect enough data to reconstruct the prior covariance matrix Σ_*.

4.2. Algorithm

The `Meta-DP++` algorithm is presented in Algorithm 3. We first define some additional notation, and then describe the algorithm in detail.

Additional Notation: As with the `Meta-DP` algorithm, at the end of each epoch $i \in [N]$, we update our estimate $\hat{\theta}_i$ of the prior mean θ_*. In addition, to estimate Σ_*, we also estimate the unknown parameter *realizations* $\{\theta_i\}$ in the exploration epochs; we refer to these estimates as $\{\tilde{\theta}_i\}$.

Algorithm Description: The first N_1 epochs are treated as exploration epochs. Recall that, in the `Meta-DP` algorithm, we employed the prior-independent UCB algorithm throughout the exploration epochs. However, in the `Meta-DP++` algorithm, we perform random price exploration for the first N_2 time steps in each exploration epoch, and perform prior-independent UCB for the remaining time steps $[N_2+1, T]$. We define these quantities as

$$N_1 = \max\left\{4c_4^2(d^2 + d\log_e(NT))N^{\frac{1}{2}}, N_0\right\} \tag{15}$$

$$N_2 = \max\left\{2c_4^2 dN^{\frac{1}{4}}, \frac{2\log_{e/2}(2dN^2T)}{c_1}\right\}, \tag{16}$$

where the constants are given by

$$c_3 = \frac{48\sigma^2}{c_0\lambda_0} \quad c_4 = \max\left\{\frac{2c_3}{3\underline{\lambda}}, \frac{12}{\underline{\lambda}}\sqrt{\overline{\lambda}+c_3}\right\}. \tag{17}$$

Note that we now require $\widetilde{O}(\sqrt{N})$ exploration epochs, whereas we only required $\widetilde{O}(d^2)$ exploration epochs for the `Meta-DP` algorithm.

As described in the overview, the `Meta-DP++` algorithm proceeds in two phases:

1. **Epoch $\mathbf{i \leq N_1}$:** In each exploration epoch, we first perform random price exploration during time steps $t \leq N_2$. For convenience, our exploration strategy is to alternate between the prices $p_{\min}$ and $p_{\max}$. This choice is arbitrary and one could alternatively randomly sample from any set of fixed prices without affecting the order of the regret. After price exploration ($t = N_2$), we compute our estimate $\tilde{\theta}_i$ of the unknown product realization θ_i using the OLS estimator, *i.e.*,

$$\tilde{\theta}_i = \left(\sum_{\tau=1}^{N_2} \boldsymbol{m}_{i,\tau}\boldsymbol{m}_{i,\tau}^\top\right)^{-1}\left(\sum_{\tau=1}^{N_2} D_{i,\tau}\left(p_{i,\tau}, \mathbf{x}_{i,\tau}\right)\boldsymbol{m}_i\right), \tag{18}$$

where we recall that $\boldsymbol{m}_{i,\tau} = \left(\mathbf{x}_{i,\tau}, p_{i,\tau}\mathbf{x}_{i,\tau}\right)^\top$ is the price and feature information of round τ. For the remaining $T - N_2$ rounds in the exploration epoch i, the `Meta-DP++` algorithm runs the prior-independent UCB algorithm described earlier in Eq. (5)-(6) to ensure low regret.

At the end of all N_1 exploration epochs, the `Meta-DP++` algorithm computes the empirical covariance matrix using the estimated realizations $\{\tilde{\theta}_i\}_{i=1}^{N_1}$ as follows:

$$\hat{\Sigma}_* = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1}\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)^\top. \tag{19}$$

2. **Epoch $\mathbf{i > N_1}$:** In all non-exploration epochs, the `Meta-DP++` algorithm proceeds similarly as the `Meta-DP` algorithm, except that it uses the estimated prior covariance matrix $\hat{\Sigma}_*$ (rather than the true Σ_*) with an additional correction term (to account for uncertainty in the estimated $\hat{\Sigma}_*$):

$$\Sigma_{\text{correction}} = 12\sqrt{\left(\bar{\lambda} + \frac{c_3 d}{N_2}\right)\left(\frac{\log_e(17)d + \log_e(NT)}{N_1}\right)}I_{2d}. \tag{20}$$

In particular, after an initialization step ($t=1$), we employ the TS algorithm using the prior:

$$\mathcal{N}\left(\theta_{i,2}^{\mathrm{MPDP}}, \Sigma_{i,2}^{\mathrm{MPDP}}\right) = \mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)$$
$$\Sigma_i = \eta_i\left(\hat{\Sigma}_* + \Sigma_{\mathrm{correction}}\right) = \left(1+\frac{1}{\sqrt{i}}\right)\left(\hat{\Sigma}_* + \Sigma_{\mathrm{correction}}\right),$$

where the estimated prior mean $\hat{\theta}_i$ is computed as before using Eq. (7). As before, the extent of prior widening decreases for later epochs (*i.e.*, i large), when we expect the estimation error of the prior mean to be small. However, there is now a component that is a fixed constant throughout the non-exploration epochs ($\Sigma_{\mathrm{correction}}$) due to uncertainty in the estimated $\hat{\Sigma}_*$.

Algorithm 3 Meta-Personalized Dynamic Pricing++ Algorithm

Input: The total number of products N, the length of each epoch T, the upper bound on the prior mean S, the subgaussian parameter σ, and the set of feasible prices $[p_{\min}, p_{\max}]$.
Initialization: N_1 and N_2 as defined in eq. (15) and (16), $\Sigma_{\mathrm{correction}}$ as defined in eq. (20)
for epoch $i=1,\ldots,N_1$ **do**
  for $t=1,\ldots,N_2$ **do**
    Observe feature $\mathbf{x}_{i,t}$, and set $p_{i,t} \leftarrow \begin{cases} p_{\min} & \text{if } i \text{ is even and } t \le \frac{N_2}{2} \text{ or } i \text{ is odd and } t > \frac{N_2}{2}, \\ p_{\max} & \text{otherwise.} \end{cases}$.
    Observe demand $D_{i,t}\left(p_{i,t}, \mathbf{x}_{i,t}\right)$.
  end for
  Compute $\tilde{\theta}_i$ according to eq. (18).
  for $t=N_2+1,\ldots,T$ **do**
    Observe feature $\mathbf{x}_{i,t}$ and update $\theta_{i,t}^{\mathrm{UCB}}$ according to eq. (5)
    Choose price $p_{i,t}$ according to eq. (6), and observe demand $D_{i,t}(p_{i,t}, \mathbf{x}_{i,t})$.
  end for
end for
Compute the empirical covariance matrix $\hat{\Sigma}_*$ according to eq. (19).
for epoch $i=N_1+1,\ldots,N$ **do**
  Observe the feature $\mathbf{x}_{i,1}$, and set $p_{i,1} \leftarrow \begin{cases} p_{\min} & \text{if } i \text{ is even,} \\ p_{\max} & \text{otherwise.} \end{cases}$
  Update $\hat{\theta}_i$ according to eq. (7), and set $\eta_i \leftarrow 1+1/\sqrt{i}, \Sigma_i \leftarrow \eta_i\left(\hat{\Sigma}_* + \Sigma_{\mathrm{correction}}\right)$.
  Construct prior $\mathcal{N}\left(\theta_{i,2}^{\mathrm{MPDP}}, \Sigma_{i,2}^{\mathrm{MPDP}}\right) \leftarrow \mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)$.
  for $t=2,\ldots,T$ **do**
    Observe feature $\mathbf{x}_{i,t}$, and sample parameter $\mathring{\theta}_{i,t} \sim \mathcal{N}\left(\theta_{i,t}^{\mathrm{MPDP}}, \Sigma_{i,t}^{\mathrm{MPDP}}\right)$.
    Choose price $p_{i,t}$ according to eq. (10), observe demand $D_{i,t}\left(p_{i,t}, \mathbf{x}_{i,t}\right)$, and compute the posterior $\mathcal{N}\left(\theta_{i,t+1}^{\mathrm{MPDP}}, \Sigma_{i,t+1}^{\mathrm{MPDP}}\right)$.
  end for
end for

REMARK 4. The `Meta-DP++` algorithm does not update its estimate $\hat{\Sigma}_*$ of the prior covariance matrix Σ_* after the initial exploration epochs. This is because estimating Σ_* requires accurate estimates $\{\tilde{\theta}_i\}$ of unknown parameter realizations $\{\theta_i\}$, which we can only obtain in the exploration

epochs (where we perform random price exploration). In other words, we rely on an *explore-then-commit* strategy, which has been shown to be near-optimal in a variety of bandit problems (Lattimore and Szepesvari 2018).

4.3. Meta Regret Analysis

We now prove an upper bound on the meta regret of the `Meta-DP++` algorithm.

Using Lemma 1 and Assumption 1, we can easily bound the meta regret from the N_1 exploration epochs by $\widetilde{O}(N_1N_2 + dN_1\sqrt{T})$: the first term captures the meta regret from N_2 steps of price exploration, and the second term captures the meta regret of prior-independent UCB in the remaining steps. Next, after the exploration epochs conclude, we estimate the prior covariance matrix $\hat{\Sigma}_*$. The following theorem bounds the error of this estimate with high probability:

THEOREM 4. *For any $\delta > 0$, with probability at least $1 - 2\delta - 4dN_1(e/2)^{-c_1 N_2/2}$, the operator norm of $\hat{\Sigma}_* - \Sigma_*$ is upper bounded as*

$$\left\|\hat{\Sigma}_* - \Sigma_*\right\|_{op} \leq \frac{2c_3 d}{3N_2} + 12\sqrt{\left(\bar{\lambda} + \frac{c_3 d}{N_2}\right)}\left[\left(\frac{d\log_e 17 - \log_e \delta}{N_1}\right) \vee \sqrt{\frac{d\log_e 17 - \log \delta}{N_1}}\right]$$

and

$$\max_{v\in\mathbb{R}^{2d}:\|v\|\leq 1} v^\top\left(\Sigma_* - \hat{\Sigma}_*\right)v \geq 12\sqrt{\left(\bar{\lambda} + \frac{c_3 d}{N_2}\right)}\left[\left(\frac{d\log_e 17 - \log_e \delta}{N_1}\right) \vee \sqrt{\frac{d\log_e 17 - \log_e \delta}{N_1}}\right].$$

Proof Sketch. The complete proof is provided in Appendix D. From Wainwright (2019), for any constant $c > 0$, we have that

$$\Pr\left(\left\|\hat{\Sigma}_* - \Sigma_*\right\|_{op} \geq c\right) \leq 17^d \Pr\left(\max_{v\in\Re^d:\|v\|\leq 1} v^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v \geq \frac{c}{2}\right).$$

Now for any fixed v in the d-dimensional unit ball and any exploration epoch $i \in [N_1]$, we can decompose

$$\begin{aligned} v^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v &= \frac{1}{N_1}\sum_{i=1}^{N_1} v^\top\left(\frac{N_1}{N_1 - 1}\left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1}\theta_j + \Delta_j}{N_1}\right)\left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1}\theta_j + \Delta_j}{N_1}\right)^\top - \Sigma_*\right)v \\ &= \frac{1}{N_1}\left(\sum_{i=1}^{N_1} Z_i Z_i^\top - v^\top\Sigma_* v\right) \end{aligned}$$

where we let $\Delta_i = \tilde{\theta}_i - \theta_i$. Since the OLS estimator is unbiased, note that $\mathbb{E}[\Delta_i] = 0$. Defining

$$Z_i = \sqrt{\frac{N_1}{N_1 - 1}}\left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1}\theta_j + \Delta_j}{N_1}\right)^\top v,$$

we note that $\mathbb{E}[Z_i] = 0$, and $\mathbb{E}[Z_i^2] = v^\top \left(\Sigma_* + \mathbb{E}\left[\Delta_i \Delta_i^\top\right]\right) v$. Furthermore, we observe that its moment generating function can be upper bounded as follows: for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda Z_i\right)\right] \leq \exp\left(\lambda^2 \left(\frac{\bar{\lambda}}{2} + \frac{22\sigma^2 d}{c_0 \lambda_0 N_2}\right)\right).$$

Then, Lemma 1.12 of Rigollet and Hütter (2018) implies that $Z_i^2 - \mathbb{E}[Z_i^2]$ is subexponential with parameter $16\left(\bar{\lambda} + \frac{44\sigma^2 d}{c_0 \lambda_0 N_2}\right)$. The result then follows by applying Bernstein's inequality. $\square$

The above theorem yields the following performance guarantee for the `Meta-DP++` algorithm.

THEOREM 5. *If $N = \widetilde{\Omega}\left(d^4\right)$ and $T = \widetilde{\Omega}\left(dN^{1/4}\right)$, the meta regret of the proposed `Meta-DP++` algorithm is upper bounded as*

$$\mathcal{R}_{N,T}(\textit{`Meta-DP++` algorithm}) = \tilde{O}\left(d^2 N^{\frac{3}{4}} T^{\frac{1}{2}}\right)$$

The proof of Theorem 5 is provided in Appendix E.

REMARK 5. The requirement $N = \widetilde{\Omega}\left(d^4\right)$ is purely for the brevity of presentation, and the meta regret bound still holds even if $N = \widetilde{\Omega}\left(d^2\right)$, *i.e.*, the same condition as Theorem 3, but the exponent of d in the regret expression will be slightly larger due to different choices of N_1 and N_2.

4.4. Additional Remarks

Knowledge of N, T: Our formulation assumes knowledge of N and T. However, this assumption can easily be removed using the well-known "doubling trick". In particular, we can initially fix any values N_0 and T_0, and iteratively double the length of the respective horizons; we refer the interested reader to Cesa-Bianchi and Lugosi (2006) for details. For the `Meta-DP` algorithm, we would simply continue to update the estimated prior mean and follow the prior widening schedule; for the `Meta-DP++` algorithm, we would need to also perform additional random price exploration to ensure that we have sufficient data to reconstruct the prior covariance matrix Σ_*. It is easy to see that our regret bounds are preserved up to logarithmic terms under such an approach.

Overlapping Epochs: We model epochs as fully sequential for simplicity; if epochs overlap, we would need to additionally model a customer arrival process for each epoch. Our algorithms straightforwardly generalize to a setting where arrivals are randomly distributed across overlapping epochs. In particular, since the `Meta-DP` algorithm only uses the *initial* sample from each epoch for estimating the prior mean, the algorithm and analysis are not affected. For the `Meta-DP++` algorithm, we would need to employ random price exploration until we observe at least $N_2 = \widetilde{O}(d\sqrt{N})$ samples from at least $N_1 = \widetilde{O}(dN^{\frac{1}{4}})$ epochs to estimate the prior covariance Σ_*; after this, we again only require the initial sample from the remaining epochs for estimating the prior mean.

5. Numerical Experiments

We now validate our theoretical results by empirically comparing the performance of our proposed algorithms against algorithms that ignore shared structure and a greedy approach that does not employ prior widening (see discussion in Section 3.4). In particular, we compare the `Meta-DP` algorithm and the `Meta-DP++` algorithm against two benchmarks:

1. *Prior-free:* This algorithm runs a separate prior-independent Thompson sampling algorithm in each epoch; we use the algorithm proposed by Agrawal and Goyal (2013). This approach ignores learning shared structure (the prior) across products, and achieves $\widetilde{O}(N)$ meta regret.

2. *Greedy:* This algorithm is identical to the `Meta-DP` algorithm when the prior covariance is known, and the `Meta-DP++` algorithm when the prior covariance is unknown, with the exception that it does not employ prior widening in both cases. In particular, $\eta_i = 1$ for all $i \in [N]$.

We perform numerical experiments on both synthetic data as well as a real dataset on auto loans provided by the Columbia University Center for Pricing and Revenue Management.

5.1. Synthetic Data

We begin with the case where the prior covariance Σ_* is known.

Parameters: We consider $N = 1000$ products, each with a selling horizon of $T = 1000$ periods. We set the feature dimension $d = 5$, the prior mean $\theta_* = 10^{-1} \times [\mathbf{1}_d; -\mathbf{1}_d]^\top$, and the prior covariance $\Sigma_* = 10^{-2} \times I_{2d}$. In each epoch $i \in [N]$ and each round $t \in [T]$, each entry of the observed feature vector $\mathbf{x}_{i,t}$ is drawn i.i.d. from the uniform distribution over $[0, 1/\sqrt{d}]^d$; note that this ensures the ℓ_2 norm of each feature vector is upper bounded by 1. For each product $i \in [N]$, we randomly draw a demand parameter θ_i i.i.d. from the true prior $\mathcal{N}(\theta_*, \Sigma_*)$. The allowable prices are given by the set $(0, 1]$,. Finally, the noise distribution is the standard normal distribution, *i.e.*, $\sigma = 1$.

Results: We plot the cumulative meta regret of each algorithm, averaged over 10 random trials, as a function of the number of epochs N. (Recall that each epoch lasts for T periods.) The results are shown in Fig. 1. As expected, the prior-independent approach performs poorly, since it ignores shared structure; it achieves meta regret that scales linearly in N, since each epoch is treated independently. The `Meta-DP` algorithm and the greedy algorithm are identical during the exploration epochs. Thus, we see that both algorithms achieve linear meta regret in these first few epochs, while collecting initial data to form an estimate of the prior mean θ_*. After the exploration epochs end, we see the `Meta-DP` algorithm with prior widening achieves much slower growth of its cumulative meta regret compared to the greedy algorithm. In particular, when $N = 1000$, the meta regret of the `Meta-DP` algorithm is $\geq 25\%$ less than that of the greedy algorithm. This result suggests that prior widening is indeed critical for achieving good empirical performance.

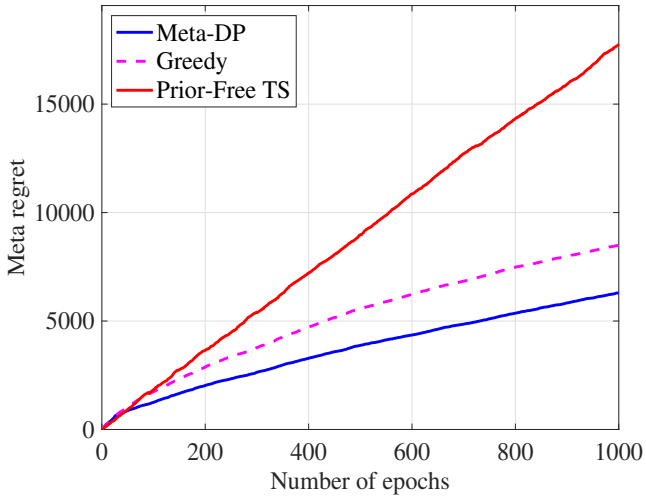
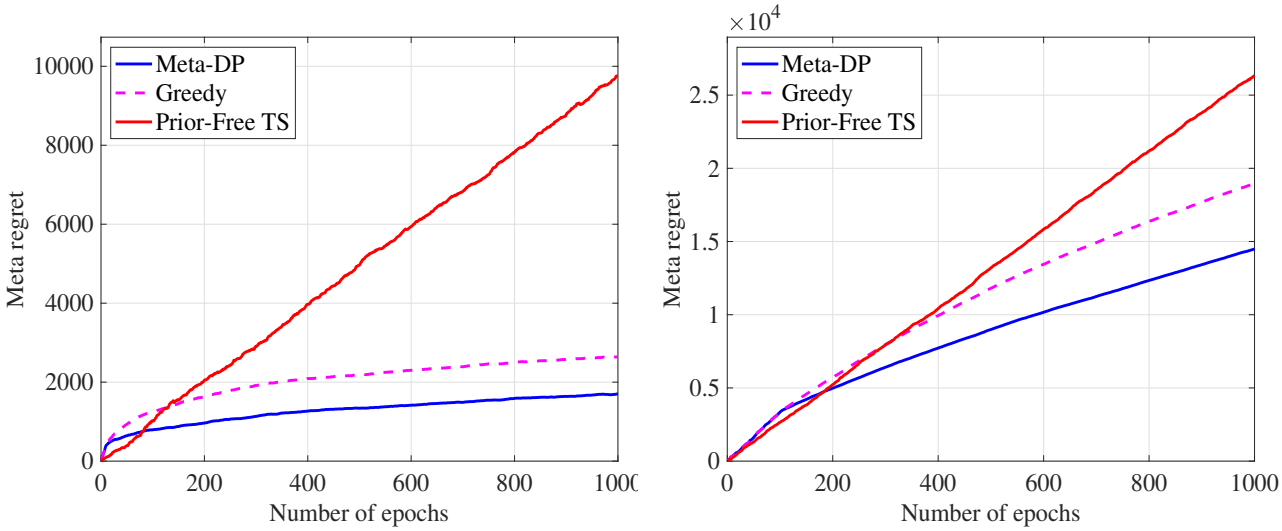


Figure 1 **Cumulative meta regret for the `Meta-DP` algorithm and benchmark algorithms.**



(a) $d = 1$ (b) $d = 10$

Figure 2 **Cumulative meta regret for the `Meta-DP` algorithm and benchmark algorithms for different values of d.**

Varying the feature dimension d: We now explore how our results vary as we change the dimension of the observed features. Our previous results considered $d = 5$. We now additionally consider:

1. *No features, $d = 1$:* We set $\mathbf{x}_{i,t} = 1$ for all $i \in [N]$ and $t \in [T]$.
2. *Many features, $d = 10$:* Each entry of the observed feature vector $\mathbf{x}_{i,t}$ is again drawn i.i.d. from the uniform distribution over $[0, 1/\sqrt{d}]^d$ for all $i \in [N]$ and $t \in [T]$.

The results for both cases, averaged over 10 random trials, are shown in Fig. 2(a) and 2(b) respectively. Again, we see that the performance of the `Meta-DP` algorithm is significantly better than the other two benchmarks, regardless of the choice of feature dimension d. Note that we require more exploration epochs when d is larger (recall that N_0 scales as d^2).

Interestingly, we also note that the gap between the greedy approach and our proposed approach appears higher when the dimension is smaller. In particular, when $d = 1$, the `Meta-DP` algorithm lowers meta regret by over 35% compared to the greedy approach when $N = 1000$. But when $d = 5$ or $d = 10$, this improvement reduces to roughly 25%. This finding matches empirical results by Bastani et al. (2017), which suggest that greedy approaches are less likely to fail or "get stuck" in sub-optimal fixed points when the feature dimension is larger.

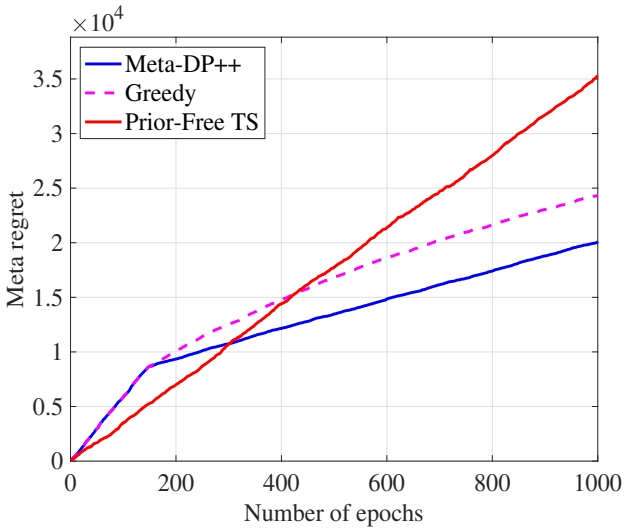


Figure 3 Cumulative meta regret for the `Meta-DP++` algorithm and benchmark algorithms.

Unknown prior covariance Σ_**:* We now consider the setting where Σ_* is unknown. Thus, we shift our attention to the `Meta-DP++` algorithm, and we adapt our greedy benchmark to follow `Meta-DP++` algorithm as well (again, we drop the prior widening step and take $\eta_i = 1$ for all $i \in [N]$). We follow the same setup described earlier, but we increase T to 2000 since the algorithms need more time to recover the underlying parameters. The results, averaged over 10 random trials, are shown in Fig. 3. We again see that the `Meta-DP++` algorithm significantly outperforms the other two benchmarks. The relative performance of prior-independent Thompson sampling demonstrates that learning shared structure can improve performance in experiment-rich environments (large N), even when nothing about the prior is known in advance. Moreover, we find that prior widening is still a critical ingredient of the algorithm, since the `Meta-DP++` algorithm lowers meta regret by approximately 15% compared to the greedy approach when $N = 1000$.

5.2. Real Data on Online Auto-Lending

We now turn to the on-line auto lending dataset. This dataset was first studied by Phillips et al. (2015), and subsequently used to evaluate dynamic pricing algorithms by Ban and Keskin (2017). We will follow a similar set of modeling assumptions.

The dataset records all auto loan applications received by a major online lender in the United States from July 2002 through November 2004. It contains 208,085 loan applications. For each application, we observe some loan-specific features (*e.g.*, date of application, the term and amount of loan requested, and the borrower's personal information), the lender's pricing decision (*i.e.*, the monthly payment required of the borrower), and the resulting demand (*i.e.*, whether or not this offer was accepted by the borrower). We refer the interested reader to Columbia University Center for Pricing and Revenue Management (Columbia 2015) for a detailed description of the dataset.

Products: We first define a set of related products. We segment loans by the borrower's state (there are 50 states), the term class of the loan (0-36, 37-48, 49-60, or over 60 months), and the car type (new, used, or refinanced). The expected demand and loan decisions offered for each type of loan is likely different based on these attributes. We consider loans that share all three attributes as a single "product" offered by the online lender. We thus obtain a total of $N = 589$ unique products. The number of applicants in the data for each loan type determines T for each product; importantly, note that T is not identical across products.

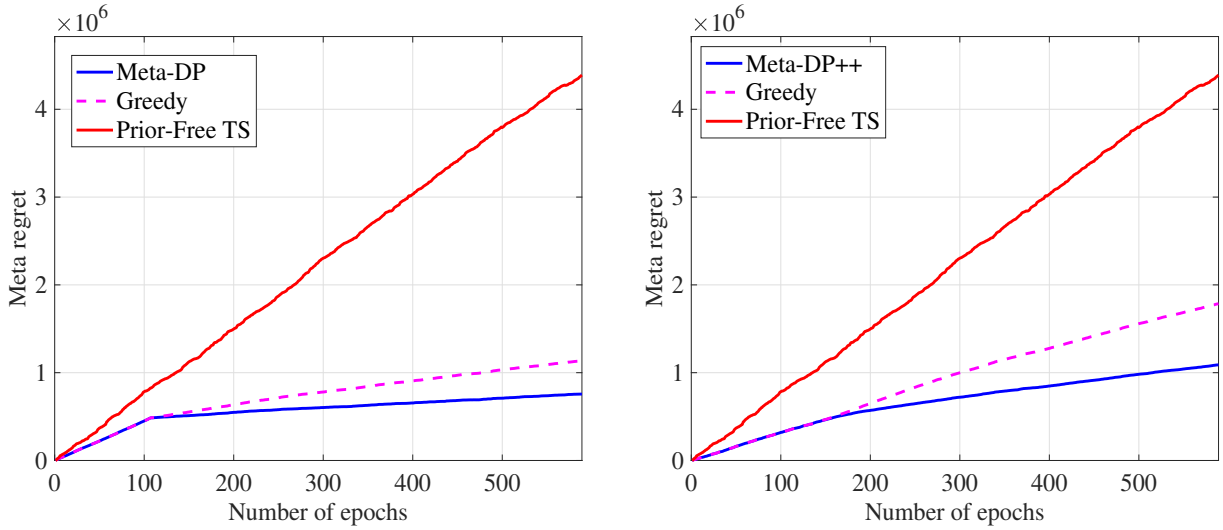
REMARK 6. Following our model, we simulate each epoch sequentially. In reality, customers will likely arrive randomly for each loan type at different points of time. We note that the `Meta-DP` algorithm only uses the initial sample from each epoch for estimating the prior mean, and thus, in principle, it can be adapted to a setting where arrivals are randomly distributed across overlapping epochs as well (see discussion in §4.4).

Features: We consider two cases: (i) the non-contextual case (*i.e.*, $d = 1$) and (ii) the contextual case ($d = 4$), where additional loan and customer features are observed as well. In the latter case, we use the feature selection results from Ban and Keskin (2017), which yields the following features: FICO score, the loan amount approved, prime rate, and the competitor's rate.

Setup: Following the approach of Phillips et al. (2015) and Ban and Keskin (2017), we impute the price of a loan as the net present value of future payments (a function of the monthly payment, customer rate, and term approved; we refer the reader to the cited references for details). The allowable price range in our experiment is $[0, 300]$.

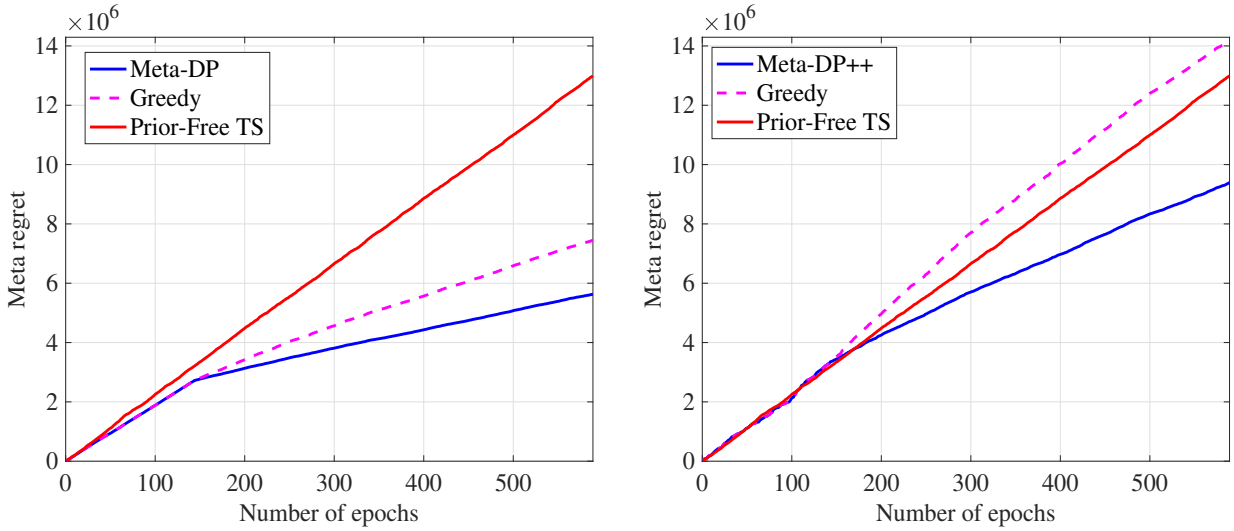
We note that, although we use a linear demand model, our responses are binary (*i.e.*, whether a customer accepts the loan). This approach is common in the literature (see, e.g., Li et al. 2010). Besbes and Zeevi (2015) provide theoretical justification for this approach by showing that we may still converge to the optimal price despite the demand model being misspecified.

Finally, unlike our model and analysis, the true distribution over loan demand parameters across products may not be a multivariate gaussian. We fit a multivariate gaussian over our data to inform the "oracle," and to provide the `Meta-DP` algorithm with the "true" Σ_*. However, the meta regret is otherwise evaluated with respect to the true data. Thus, this experiment can provide a check on whether our proposed algorithms are robust to model misspecification of the prior.



(a) Known Σ_* (b) Unknown Σ_*

Figure 4 Results for the online auto-lending dataset: non-contextual case.



(a) Known Σ_* (b) Unknown Σ_*

Figure 5 Results for the online auto-lending dataset: contextual case.

Results: We average our results over 50 random permutations (within each epoch) of the data. The results for the non-contextual case are shown in Fig. 4 while the results for the contextual case are shown in Fig. 5. In both cases, we consider the two settings where the prior covariance Σ_* is known and unknown. We again see that the `Meta-DP` algorithm and the `Meta-DP++` algorithm significantly outperform the other two benchmarks in all cases. Interestingly, despite potential misspecification of the prior's model class, we find that we can still achieve improved meta regret by leveraging shared structure. In particular, we outperform prior-independent Thompson sampling and our meta regret appears to grow sublinearly in N, even though the multivariate gaussian prior

that we estimate may not be the true prior. This result suggests that our proposed algorithms may be robust to model misspecification of the prior. Furthermore, we see the importance of prior widening, since the `Meta-DP` algorithm and the `Meta-DP++` algorithm lowers the meta regret by at least 20% when compared to the greedy approach for $N = 589$.

6. Discussion & Conclusions

Firms are increasingly performing experimentation. This provides an opportunity for decision-makers to learn not just *within* experiments, but also *across* experiments. In this paper, we consider the multi-product dynamic pricing setting where a decision-maker must learn a sequence of related unknown parameters through experimentation; we capture the relationship across these unknown parameters by imposing that they arise from a shared distribution (the prior). We propose meta-learning policies that efficiently learn both the shared distribution across experiments and the individual unknown parameters within experiments.

Our meta-learning approach can easily be adapted beyond dynamic pricing applications to classical multi-armed and contextual bandit problems as well. For instance, consider clinical trials, which were the original motivation for bandit problems (Thompson 1933, Lai and Robbins 1985). Many have argued the benefits of Bayesian clinical trials, which allow for the use of historical information and for synthesizing results of past relevant trials, *e.g.*, past clinical trials on the same disease may indicate that patients with certain biomarkers or concomitant medications are less likely to benefit from standard therapy. Such information can be encoded in a Bayesian prior to potentially allow for more informative clinical trials and improved treatment allocations to patients within the trial (see, e.g., Berry 2006, Chick et al. 2018). Our meta-learning approach can inform how such priors are constructed. Importantly, prior widening gracefully transitions from an uninformative to an informative prior as we accrue data from more related clinical trails.

Our prior widening technique is inspired by the emerging literature studying prior misspecification in Thompson sampling. In general, adopting a more conservative prior allows Thompson sampling to still achieve the optimal theoretical guarantee, while a less conservative prior may cause failure to converge (Honda and Takemura 2014, Liu and Li 2015). However, the use of a conservative prior often results in poor empirical performance, and can erode the benefit of using Thompson sampling over UCB and other prior-free approaches (see, *e.g.*, Russo and Van Roy 2014, Bastani et al. 2017). We take the view that a successful implementation of Thompson sampling *requires* learning an appropriate prior, and propose meta-learning policies to achieve this goal across a sequence of learning problems.

Acknowledgments

The authors gratefully acknowledge Columbia University Center for Pricing and Revenue Management for providing us the dataset on auto loans.

References

Abbasi-Yadkori, Yasin, David Pál, Csaba. Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *NIPS*.

Agrawal, Shipra, Nikhil R Devanur. 2014. Bandits with concave rewards and convex knapsacks. *EC*. ACM, 989–1006.

Agrawal, Shipra, Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. *International Conference on Machine Learning*. 127–135.

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

Auer, Peter. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3**(Nov) 397–422.

Ban, Gah-Yi, N Bora Keskin. 2017. Personalized dynamic pricing with machine learning .

Bastani, Hamsa, Mohsen Bayati, Khashayar Khosravi. 2017. Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011* .

Berry, Donald A. 2006. Bayesian clinical trials. *Nature reviews Drug discovery* **5**(1) 27.

Besbes, Omar, Yonatan Gur, Assaf Zeevi. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. *NIPS*. 199–207.

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

Besbes, Omar, Assaf Zeevi. 2015. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science 61(4):723–739*.

Bhat, Nikhil, Vivek F Farias, Ciamac C Moallemi, Deeksha Sinha. 2019. Near optimal ab testing. *Management Science* .

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

Bubeck, Sébastien, Che-Yu Liu. 2013. Prior-free and prior-dependent regret bounds for thompson sampling. *NIPS*. 638–646.

Cesa-Bianchi, Nicolò, Gábor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Chapelle, Olivier, Lihong Li. 2011. An empirical evaluation of thompson sampling. *NIPS*. 2249–2257.

Chick, Stephen E, Noah Gans, Ozge Yapar. 2018. Bayesian sequential learning for clinical trials of multiple correlated medical interventions .

Cohen, Maxime, Ilan Lobel, Renato Paes Leme. 2016. Feature-based dynamic pricing .

Columbia. 2015. Center for pricing and revenue management datasets. URL `https://www8.gsb.columbia.edu/cprm/sites/cprm/files/files/CPRM_AutoLoan_Data%20dictionary%283%29.pdf`.

Dani, Varsha, Thomas Hayes, Sham Kakade. 2008. Stochastic linear optimization under bandit feedback. *COLT* .

den Boer, Arnoud V, Bert Zwart. 2013. Simultaneously learning and optimizing using controlled variance pricing. *Management science* **60**(3) 770–783.

Farias, Vivek F, Benjamin Van Roy. 2010. Dynamic pricing with a prior on market response. *Operations Research* **58**(1) 16–29.

Ferreira, Kris, David Simchi-Levi, He Wang. 2018. Online network revenue management using thompson sampling. *Operations Research*.

Ferreira, Kris Johnson, Bin Hong Alex Lee, David Simchi-Levi. 2015. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* **18**(1) 69–88.

Finn, Chelsea, Pieter Abbeel, Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*. 1126–1135.

Finn, Chelsea, Kelvin Xu, Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. *NIPS*.

Fisher, Marshall, Santiago Gallino, Jun Li. 2017. Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management Science* **64**(6) 2496–2514.

Harrison, J Michael, N Bora Keskin, Assaf Zeevi. 2012. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science* **58**(3) 570–586.

Hartland, Cédric, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, Michéle Sebag. 2006. Multi-armed bandit, dynamic environments and meta-bandits .

Honda, Junya, Akimichi Takemura. 2014. Optimality of thompson sampling for gaussian bandits depends on priors. *AISTATS*. 375–383.

Javanmard, Adel, Hamid Nazerzadeh. 2019. Dynamic pricing in high-dimensions. *JMLR* .

Johari, Ramesh, Leo Pekelis, David J Walsh. 2015. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922* .

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

Kleinberg, Robert, Tom Leighton. 2003. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *FOCS*. IEEE, 594.

Lai, Tze Leung, Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* **6**(1) 4–22.

Lattimore, T., C. Szepesvári. 2018. *Bandit Algorithms*. Cambridge University Press.

Lattimore, Tor, Csaba Szepesvari. 2018. Bandit algorithms. *Cambridge University Press, Available at: http://banditalgs.com*.

Li, Lihong, Wei Chu, John Langford, Robert Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. *WWW*.

Liu, Che-Yu, Lihong Li. 2015. On the prior sensitivity of thompson sampling. *arXiv preprint arXiv:1506.03378* .

Maes, Francis, Louis Wehenkel, Damien Ernst. 2012. Meta-learning of exploration/exploitation strategies: The multi-armed bandit case. *International Conference on Agents and Artificial Intelligence*. Springer, 100–115.

Optimizely. 2019. Online. URL `https://www.optimizely.com/optimization-glossary/ab-testing/`. [Last accessed January 21, 2019].

Phillips, Robert, A. Serdar Simsek, Garrett van Ryzin. 2015. The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science 61(8):1741–1759*.

Qiang, Sheng, Mohsen Bayati. 2016. Dynamic pricing with demand covariates .

Raina, Rajat, Andrew Y Ng, Daphne Koller. 2006. Constructing informative priors using transfer learning. *ICML*. ACM, 713–720.

Rigollet, R., J. Hütter. 2018. *High Dimensional Statistics*. Lecture Notes.

Rusmevichientong, Paat, John N Tsitsiklis. 2010. Linearly parameterized bandits. *Mathematics of Operations Research* **35**(2) 395–411.

Russo, Daniel, Benjamin Van Roy. 2014. Learning to optimize via posterior sampling. *Mathematics of Operations Research 39(4):1221–1243. https://doi.org/10.1287/moor.2014.0650*.

Russo, Daniel J, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning* **11**(1) 1–96.

Scott, Steven L. 2015. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry* **31**(1) 37–45.

Sharaf, Amr, Hal Daumé III. 2019. Meta-learning for contextual bandit exploration. *arXiv preprint arXiv:1901.08159* .

Thompson, William R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**(3/4) 285–294.

Tropp, Joel. 2011. User-friendly tail bounds for matrix martingales. *Available at: http://www.dtic.mil/dtic/tr/fulltext/u2/a555817.pdf*.

Wainwright, Martin. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wang, Zi, Beomjoon Kim, Leslie Pack Kaelbling. 2018. Regret bounds for meta bayesian optimization with an unknown gaussian process prior. *NIPS*. 10498–10509.

Xu, Joseph, Peter Fader, Senthil K Veeraraghavan. 2019. Designing and evaluating dynamic pricing policies for major league baseball tickets. *MSOM* .

Yoon, Jaesik, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, Sungjin Ahn. 2018. Bayesian model-agnostic meta-learning. *NIPS*. 7343–7353.

Zhang, Dennis J, Hengchen Dai, Lingxiu Dong, Fangfang Qi, Nannan Zhang, Xiaofei Liu, Zhongyi Liu. 2017. How does dynamic pricing affect customer behavior on retailing platforms? evidence from a large randomized experiment on alibaba .

Zhu, Ruihao, Eytan Modiano. 2018. Learning to route efficiently with end-to-end feedback: The value of networked structure. *Available at: https://arxiv.org/abs/1810.10637*.

Appendix. Proofs

We begin by defining some helpful notation. First, let $\texttt{REV}\left(\theta, \hat{\theta}, \Sigma\right)$ be the expected revenue obtained by running the Thompson sampling algorithm in Algorithm 1 with the (possibly incorrect) prior $\mathcal{N}\left(\hat{\theta}, \Sigma\right)$ after initialization in an epoch whose true parameter is θ. Second, let $\texttt{REV}_*(\theta)$ be the maximum expected revenue that can be obtained from an epoch parametrized by θ after initialization.

A. Proof of Theorem 1

To analyze the quantity $\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\left(\texttt{REV}_*(\theta) - \texttt{REV}(\theta, \theta_*, \Sigma_*)\right)\right]$, we construct a mapping between the dynamic pricing setting and the linear bandit setting, and try to leverage the results of TS algorithm and UCB algorithm for linear bandits (Russo and Van Roy 2014, Abbasi-Yadkori et al. 2011). Conditioned on the feature vector $\mathbf{x}$, we can map $\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\left(\texttt{REV}_*(\theta) - \texttt{REV}(\theta, \theta_*, \Sigma_*)\right)\right]$, the Bayes regret of an epoch, to the Bayes regret of the Thompson sampling algorithm (Russo and Van Roy 2014) for a linear bandit instance as follows: it has parameter $\theta = [\alpha; \beta]$ with prior $\mathcal{N}(\theta_*, \Sigma_*)$ and decision set $A_t = \{(p\mathbf{x}_t; p^2\mathbf{x}_t) : p \in [p_{\min, p_{\max}}]\}$, where $\mathbf{x}_t$ is the feature vector drawn i.i.d from the feature distribution. The magnitude of the ℓ_2-norm of the actions is at most $p_{\max}\sqrt{1+p_{\max}^2}x_{\max}$. The noise terms are conditionally $p_{\max}\sigma$-sub-Gaussian.

By Lemma 9 in Appendix F, the Bayes regret of an epoch is upper bounded as

$$\begin{aligned}\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\left(\texttt{REV}_*(\theta) - \texttt{REV}(\theta, \theta_*, \Sigma_*)\right)\right] &= \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\widetilde{O}\left(\|\theta\|\sqrt{dT}\left(\|\theta\| + \sqrt{d}\right)\right)\right] &(21)\\ &= \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\widetilde{O}\left(\|\theta\|^2\sqrt{dT} + \|\theta\| d\sqrt{T}\right)\right]. &(22)\end{aligned}$$

where eq. (21) follows from 1) The regret upper bound of a linear bandit instance scales linearly with the maximum absolute value of the rewards. 2) The absolute value of the expected reward (revenue) for each round is upper bounded as

$$\max_{p\in[p_{\min},p_{\max}]}\|\langle \boldsymbol{m}, \theta\rangle\|_1 \le \max_{p\in[p_{\min},p_{\max}]}\|\boldsymbol{m}\|\|\theta\| = \sqrt{1+p_{\max}^2}x_{\max}\|\theta\| = O\left(\|\theta\|\right) \tag{23}$$

by Cauchy-Schwarz inequality.

To proceed, we analyze the terms $\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\|\theta\|^2\right]$ and $\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\|\theta\|\right]$ separately. By the "trace trick", we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\|\theta\|^2\right] =& \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\mathrm{tr}\left(\|\theta\|^2\right)\right]\\ =& \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\mathrm{tr}\left(\theta\theta^\top\right)\right]\\ =& \mathrm{tr}\left(\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\theta\theta^\top\right]\right) &(24)\\ =& \mathrm{tr}\left(\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[(\theta-\theta_*)(\theta-\theta_*)^\top + \theta_*\theta^\top + \theta\theta_*^\top - \theta_*\theta_*^\top\right]\right)\\ =& \mathrm{tr}\left(\Sigma_* + \theta_*\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\theta^\top\right] + \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\theta\right]\theta_*^\top - \theta_*\theta_*^\top\right) &(25)\\ =& \mathrm{tr}\left(\Sigma_* + 2\theta_*\theta_*^\top - \theta_*\theta_*^\top\right)\\ =& \mathrm{tr}\left(\Sigma_*\right) + \mathrm{tr}\left(\theta_*\theta_*^\top\right)\\ =& \mathrm{tr}\left(\Sigma_*\right) + \mathrm{tr}\left(\|\theta_*\|^2\right)\end{aligned}$$

$$\leq \kappa + S^2 \tag{26}$$

Here, eq. (24) and (25) follow from the linearity of expectation, eq. (25) also makes use of the definition of the covariance matrix $\Sigma_* = \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[(\theta-\theta_*)(\theta-\theta_*)^\top + \theta_*\theta^\top\right]$, and the last step follows from Assumptions 1 and 3. Moreover, by Cauchy-Schwarz inequality, we have

$$\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}[\|\theta\|] = \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}[\|\theta\|\cdot 1] \leq \sqrt{\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}[\|\theta\|^2]\,\mathbb{E}[1]} \leq \sqrt{\kappa + S^2}. \tag{27}$$

Putting eq. (26) and (27) into eq. (22), we conclude the proof.

B. Proof of Theorem 2

For any epoch index $i \in [N]$, we begin with the following decomposition:

$$\begin{aligned} D_{i,1} =& \langle \theta_i, M_i\rangle + \varepsilon_{i,1} \\ =& \langle \theta_* + \Delta_i, M_i\rangle + \varepsilon_{i,1} \\ =& \langle \theta_*, M_i\rangle + \langle \Delta_i, M_i\rangle + \varepsilon_{i,1}, \end{aligned} \tag{28}$$

where $\Delta_i \sim \mathcal{N}(\mathbf{0}, \Sigma_*)$. Since M_i is i.i.d. across different epochs, we can equivalently view the demand realization as the mean demand $\langle \theta_*, M_i\rangle$ corrupted by the price dependent (or heteroscedastic) noise $\langle \Delta_i, M_i\rangle + \varepsilon_{i,1}$. We thus need to understand the variance proxy of the noise.

LEMMA 2. *For any $i \in [N]$, the noise $\langle \Delta_i, M_i\rangle + \varepsilon_{i,1}$ is R-subgaussian, i.e.,*

$$\forall \lambda \in \mathbb{R} \quad \mathbb{E}\left[\exp\left(\lambda\left(\langle \Delta_i, M_i\rangle + \varepsilon_{i,1}\right)\right)\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right),$$

where $R = \sqrt{x_{\max}^2 \overline{\lambda}(1+p_{\max}^2) + \sigma^2}$.

Proof of Lemma 2. From the moment generating function of multivariate normal distributions, we have that $\forall \lambda \in \mathbb{R}$

$$\begin{aligned} \mathbb{E}\left[\exp\left(\lambda\langle \Delta_i, M_i\rangle\right)\right] &= \exp\left(\frac{\lambda^2 M_i^\top \Sigma^* M_i}{2}\right) \\ &\leq \exp\left(\frac{\lambda^2 \overline{\lambda}\|M_i\|^2}{2}\right) \\ &= \exp\left(\frac{\lambda^2 \overline{\lambda}(1+p_{i,1}^2)\|\mathbf{x}_{i,1}\|^2}{2}\right) \\ &\leq \exp\left(\frac{\lambda^2 x_{\max}^2 \overline{\lambda}(1+p_{\max}^2)}{2}\right), \end{aligned} \tag{29}$$

where we have use the fact that $M_i^\top \Sigma_* M_i \leq \|M_i\|^2 \overline{\lambda}$ as Σ_* is positive semi-definite. Note that $\varepsilon_{i,1}$ is σ-subgaussian variable, we can conclude the statement.

$$\begin{aligned} \mathbb{E}\left[\exp\left(\lambda\left(\langle \Delta_i, M_i\rangle + \varepsilon_{i,1}\right)\right)\right] &\leq \exp\left(\frac{\lambda^2 x_{\max}^2 \overline{\lambda}(1+p_{\max}^2)}{2}\right)\mathbb{E}\left[\exp\left(\lambda\varepsilon_{i,1}\right)\right] \\ &\leq \exp\left(\frac{\lambda^2 (x_{\max}^2 \overline{\lambda}(1+p_{\max}^2) + \sigma^2)}{2}\right). \end{aligned}$$

□

We are now ready to analyze the convergence property of the OLS estimate $\hat{\theta}_i$. First is a lemma on the convergence of the OLS.

LEMMA 3 **(Lattimore and Szepesvári (2018), Zhu and Modiano (2018))**. *The probability that the difference between $\widehat{\theta}$ and θ under the $V_i = \mathbf{X}_i\mathbf{X}_i^\top$ norm is not less than $2R\sqrt{d\log_e 2 - \log_e \delta}$ is at most δ, i.e.,*

$$\Pr\left(\|\widehat{\theta} - \theta\|_{V_i} \geq 2R\sqrt{d\log_e 2 - \log_e \delta}\right) \leq \delta,$$

The proof of Lemma 3 can be adapted easily from (Lattimore and Szepesvári 2018, Zhu and Modiano 2018), and it is thus omitted. In order to bound the estimation error of $\widehat{\theta}$ coordinate-wisely, we further need a lower bound on the smallest eigenvalue of V_i. Since V_i is a random matrix, we appeal to the matrix Chernoff inequality (Tropp 2011). To this end, we first need a lower bound on the smallest eigenvalue of $\sum_{i\in\mathcal{Q}} \mathbb{E}\left[M_i M_i^\top\right]$.

LEMMA 4. *There exists some positive constants c_0 depends only on $\lambda_0, p_{\min}$, and $p_{\max}$, such that the minimum eigenvalue of $\sum_{\iota=1}^{i} \mathbb{E}\left[M_\iota M_\iota^\top\right]$ is lower bounded by $c_0\lambda_0 i$, i.e.,*

$$\lambda_{\min}\left(\sum_{\iota=1}^{i} \mathbb{E}\left[M_\iota M_\iota^\top\right]\right) \geq c_0\lambda_0 i.$$

Proof of Lemma 4. From linearity of expectation, we have

$$\begin{aligned}
\sum_{\iota=1}^{i} \mathbb{E}\left[M_\iota M_\iota^\top\right] &= \sum_{\iota \text{ even},\ \iota\in i} \mathbb{E}\left[M_i M_i^\top\right] + \sum_{\iota \text{ odd},\ \iota\leq i} \mathbb{E}\left[M_i M_i^\top\right] \\
&= \left\lfloor \frac{i}{2} \right\rfloor \left(\begin{pmatrix} \mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] & p_{\min}\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] \\ p_{\min}\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] & p_{\min}^2\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] \end{pmatrix} + \begin{pmatrix} \mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] & p_{\max}\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] \\ p_{\max}\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] & p_{\max}^2\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] \end{pmatrix} \right) \\
&= \left\lfloor \frac{i}{2} \right\rfloor \begin{pmatrix} 2\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] & (p_{\min}+p_{\max})\,\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] \\ (p_{\min}+p_{\max})\,\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] & (p_{\min}^2+p_{\max}^2)\,\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top] \end{pmatrix}
\end{aligned}$$

Now from the fact that for any positive semi-definite matrix $A \in \mathbb{R}^{2d\times 2d}$,

$$\lambda_{\min}(A) = \min_{z\in\mathbb{R}^{2d}:\|z\|^2=1} z^\top A z, \tag{30}$$

we have

$$\begin{aligned}
&\lambda_{\min}\left(\sum_{\iota=1}^{i} \mathbb{E}\left[M_\iota M_\iota^\top\right]\right) \\
=&\left\lfloor \frac{i}{2} \right\rfloor \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \begin{pmatrix} z_1^\top & z_2^\top \end{pmatrix} \left(\sum_{\iota=1}^{i} \mathbb{E}\left[M_\iota M_\iota^\top\right]\right) \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \\
=&\left\lfloor \frac{i}{2} \right\rfloor \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} 2z_1^\top\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top]z_1 + 2(p_{\min}+p_{\max})z_2^\top\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top]z_1 + (p_{\min}^2+p_{\max}^2)z_2^\top\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top]z_2 \\
=&\left\lfloor \frac{i}{2} \right\rfloor \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} (p_{\min}z_2+z_1)^\top\,\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top]\,(p_{\min}z_2+z_1) + (p_{\max}z_2+z_1)^\top\,\mathbb{E}[\mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top]\,(p_{\max}z_2+z_1) \\
\geq&\left\lfloor \frac{i}{2} \right\rfloor \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \lambda_0\,(p_{\min}z_2+z_1)^\top\,(p_{\min}z_2+z_1) + \lambda_0\,(p_{\max}z_2+z_1)^\top\,(p_{\max}z_2+z_1) \qquad (31) \\
=&\lambda_0\left\lfloor \frac{i}{2} \right\rfloor \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \left[\left(p_{\min}^2+p_{\max}^2\right)\|z_2\|^2 + 2\|z_1\|^2 + 2(p_{\min}+p_{\max})z_2^\top z_1\right]
\end{aligned}$$

$$
\begin{aligned}
&\geq \lambda_0 \left\lfloor \frac{i}{2} \right\rfloor \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \left[\left(p_{\min}^2 + p_{\max}^2\right) \|z_2\|^2 + 2\|z_1\|^2 - 2(p_{\min}+p_{\max})\|z_2\|\|z_1\| \right] \qquad (32)\\
&= \lambda_0 \left\lfloor \frac{i}{2} \right\rfloor \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \left[\left(p_{\min}\|z_2\| - \|z_1\|\right)^2 + \left(p_{\max}\|z_2\| - \|z_1\|\right)^2 \right]\\
&\geq \frac{\lambda_0 i}{3} \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \left[\left(p_{\min}\|z_2\| - \|z_1\|\right)^2 + \left(p_{\max}\|z_2\| - \|z_1\|\right)^2 \right],
\end{aligned}
$$

where inequality (31) follows again from equation (30) and inequality (32) follows from Cauchy-Schwarz inequality. Now we see that if $\min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \left[\left(p_{\min}\|z_2\| - \|z_1\|\right)^2 + \left(p_{\max}\|z_2\| - \|z_1\|\right)^2 \right] \leq 0$, then both $p_{\min}\|z_2\| - \|z_1\|$ and $p_{\max}\|z_2\| - \|z_1\|$ should be 0. However, this can hold if and only if $\|z_1\| = \|z_2\| = 0$, which contradicts the constraint $\|z_1\|^2 + \|z_2\|^2 = 1$. Therefore, we can take

$$
c_0 = \frac{1}{3} \min_{z_1,z_2\in\mathbb{R}^d:\|z_1\|^2+\|z_2\|^2=1} \left[\left(p_{\min}\|z_2\| - \|z_1\|\right)^2 + \left(p_{\max}\|z_2\| - \|z_1\|\right)^2 \right] > 0.
$$

to conclude the statement. □

We are now ready to apply the matrix Chernoff inequality (Tropp 2011) to arrive at the following result

LEMMA 5 **(Tropp (2011))**. *The probability that the minimal eigenvalue of $V_i = \mathbf{X}_i\mathbf{X}_i^\top$ is larger than $(1-\zeta)c_0\lambda_0 i$ with probability at least $1 - 2d\left(\frac{\exp(-\zeta)}{(1-\zeta)^{1-\zeta}}\right)^{c_1 i}$ for any $\zeta \in [0,1)$, i.e.,*

$$
\Pr\left(\lambda_{\min}(V_i) \geq (1-\zeta)c_0\lambda_0 i\right) \geq 1 - 2d\left(\frac{\exp(-\zeta)}{(1-\zeta)^{1-\zeta}}\right)^{c_1 i},
$$

where

$$
c_1 = \frac{c_0\lambda_0}{\sqrt{(1+p_{\max}^2)}x_{\max}}.
$$

The proof of this lemma is a straightforward result from Lemma 4 and Theorem 3.1 in (Tropp 2011), and it is thus omitted. Finally by a union bound between Lemma 3 and Lemma 5, we conclude the proof of Theorem 2.

C. Proof of Theorem 3

First, we define the clean event $\mathcal{E}$:

$$
\forall i \geq N_0 \qquad \left\|\hat{\theta}_i - \theta_*\right\| \leq \frac{2R\sqrt{2\log_e(2)d + 2\log_e\left(N^2T\right)}}{\sqrt{c_0\lambda_0 i}}. \qquad (33)
$$

The meta regret can then be decomposed as follows:

$$
\mathcal{R}_{N,T} = (\mathcal{R}_{N,T}|\mathcal{E})\Pr(\mathcal{E}) + (\mathcal{R}_{N,T}|\neg\mathcal{E})\Pr(\neg\mathcal{E}) \leq (\mathcal{R}_{N,T}|\mathcal{E}) + (\mathcal{R}_{N,T}|\neg\mathcal{E})\Pr(\neg\mathcal{E}). \qquad (34)
$$

Applying a union bound over the epochs $i \geq N_0$ to Theorem 1 (with $\delta = 1/(N^2T)$ and $\zeta = 1/2$) to obtain that the clean event $\mathcal{E}$ holds with probability at least

$$
\begin{aligned}
\Pr(\mathcal{E}) &\geq 1 - (N+1-N_0)\,\delta - 2d\sum_{i=N_0}^{N}\left(\frac{e}{2}\right)^{-c_1 i/2}\\
&\geq 1 - \frac{1}{NT} - 2d\frac{(e/2)^{-\log_{e/2}(2dNT)}}{1-(e/2)^{-1/2}}\\
&\geq 1 - \frac{1}{NT} - \frac{8}{NT}\\
&= 1 - \frac{9}{NT}, \qquad (35)
\end{aligned}
$$

When the clean event $\mathcal{E}$ is violated, the meta regret is at most $O(NT)$, and its contribution to the expected meta regret is

$$O(NT)\Pr(\neg\mathcal{E}) = O(1). \tag{36}$$

We then proceed to analyze the regret of each epoch conditioned on the clean event $\mathcal{E}$. For an epoch $i \geq N_0$, the meta regret $\mathcal{R}_{N,T}(i)|\mathcal{E}$ of this epoch can be written as

$$\begin{aligned}
\mathcal{R}_{N,T}(i)|\mathcal{E} =& \mathbb{E}_{\theta_i}\mathbb{E}_{\hat{\theta}_i}\left[\texttt{REV}\left(\theta_i,\theta_*,\Sigma_*\right) - \texttt{REV}\left(\theta_i,\hat{\theta}_i,\Sigma_i\right)\middle|\mathcal{E}\right] \\
=& \mathbb{E}_{\theta_i}\mathbb{E}_{\hat{\theta}_i}\left[\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i,\hat{\theta}_i,\Sigma_i\right) - \left(\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i,\theta_*,\Sigma_*\right)\right)\middle|\mathcal{E}\right] \\
=& \mathbb{E}_{\theta_i}\mathbb{E}_{\hat{\theta}_i}\left[\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i,\hat{\theta}_i,\Sigma_i\right)\right] - \mathbb{E}_{\theta_i}\left[\left(\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i,\theta_*,\Sigma_*\right)\right)|\mathcal{E}\right].
\end{aligned} \tag{37}$$

Now from Lemma 10 in Appendix F, we have that if the parameter θ_i follows a multivariate normal distribution $\mathcal{N}\left(\theta_*,\Sigma_*\right)$, the regret of running the Thompson sampling algorithm of (Russo and Van Roy 2014) with a multivariate normal distribution $\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)$ as prior, *i.e.*, the first term of equation (37), is upper bounded as

$$\begin{aligned}
&\mathbb{E}_{\theta_i}\mathbb{E}_{\hat{\theta}_i}\left[\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i,\hat{\theta}_i,\Sigma_i\right)\middle|\mathcal{E}\right] \\
\leq& \mathbb{E}_{\hat{\theta}_i}\left\|\frac{d\mathcal{N}\left(\theta_*,\Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty} \mathbb{E}_{\theta_i}\left[\left(\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i,\theta_*,\Sigma_*\right)\right)|\mathcal{E}\right],
\end{aligned} \tag{38}$$

where $\frac{d\mathcal{N}(\theta_*,\Sigma_*)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}$ is the Radon-Nikodym derivative of $\mathcal{N}\left(\theta_*,\Sigma_*\right)$ with respect to $\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)$, $\|\cdot\|_{\mathcal{N}\left(\bar{\theta}_n,\Sigma_n\right),\infty}$ is the essential supremum magnitude with respect to $\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)$. With inequality (38), it is evident that (37) is upper bounded as

$$\mathcal{R}_{N,T}(i)|\mathcal{E} \leq \mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*,\Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty} - 1\right]\mathbb{E}_{\theta_i}\left[\left(\texttt{REV}_*\left(\theta_i\right) - \texttt{REV}\left(\theta_i,\theta_*,\Sigma_*\right)\right)|\mathcal{E}\right].$$

and the conditional meta regret can thus be upper bounded as

$$\begin{aligned}
&\mathcal{R}_{N,T}|\mathcal{E} \\
=& \sum_{i=1}^{N}\mathcal{R}_{N,T}(i)|\mathcal{E} \\
\leq& \sum_{i=1}^{N_0-1}\mathcal{R}_{N,T}(i)|\mathcal{E} + \sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*,\Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty} - 1\right]\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}\left(\theta_*,\Sigma_*\right)}\left[\left(\texttt{REV}_*\left(\theta\right) - \texttt{REV}\left(\theta,\theta_*,\Sigma_*\right)\right)|\mathcal{E}\right] \\
\leq& \sum_{i=1}^{N_0-1}\mathcal{R}_{N,T}(i)|\mathcal{E} + \sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*,\Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty} - 1\right]\widetilde{O}\left(d\sqrt{T}\right),
\end{aligned} \tag{39}$$

where the last step follows from Theorem 1. We then analyze the terms on the RHS of inequality (39) separately.

C.1. Analyzing $\sum_{i=1}^{N_0-1}\mathcal{R}_{N,T}(i)|\mathcal{E}$

This part is an immediate corollary of Theorem 1. For each of the epochs $i\in[N_0-1]$, the expected regret is $\widetilde{O}(d\sqrt{T})$, and the total meta regret is

$$\sum_{i=1}^{N_0-1}\mathcal{R}_{N,T}(i)|\mathcal{E}=\widetilde{O}\left(dN_0\sqrt{T}\right)=\widetilde{O}\left(d^3\sqrt{T}\right)=\widetilde{O}\left(d^2\sqrt{NT}\right). \tag{40}$$

C.2. Analyzing $\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}(\theta_*,\Sigma_*)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty}-1\right]$

By definition of the Radon-Nikodym derivative and the multivariate normal distribution, we have

$$\begin{aligned}&\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*,\Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty}-1\right]\\&=\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\sup_{\theta}\frac{\det\left(2\pi\Sigma_*\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)\right)}{\det\left(2\pi\Sigma_i\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\theta-\hat{\theta}_i)^\top\Sigma_i^{-1}(\theta-\hat{\theta}_i)\right)}-1\right].\end{aligned} \tag{41}$$

Recall that $\Sigma_i=\eta_i\Sigma_*\ \left(\in\mathbb{R}^{2d\times 2d}\right)$, equation (41) can be rewritten as

$$\begin{aligned}&\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\sup_{\theta}\exp\left(\frac{(\theta-\hat{\theta}_i)^\top\Sigma_i^{-1}(\theta-\hat{\theta}_i)-(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)}{2}\right)-1\right]\\&=\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\sup_{\theta}\exp\left(\frac{(\theta-\theta_*+\theta_*-\hat{\theta}_i)^\top\Sigma_*^{-1}(\theta-\theta_*+\theta_*-\hat{\theta}_i)-\eta_i(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)}{2\eta_i}\right)-1\right]\\&=\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\sup_{\theta}\exp\left(\frac{\Delta_i^\top\Sigma_*^{-1}\Delta_i-(\eta_i-1)(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)+2(\theta-\theta_*)^\top\Sigma_*^{-1}\Delta_i}{2\eta_i}\right)-1\right],\end{aligned}$$

where we have used Δ_i to denote $\theta_*-\hat{\theta}_i$. We then complete the square for the last two terms in $\exp(\cdot)$ to arrive at

$$\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\sup_{\theta}\exp\left(\frac{\frac{\eta_i}{\eta_i-1}\Delta_i^\top\Sigma_*^{-1}\Delta_i-(\eta_i-1)\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)^\top\Sigma_*^{-1}\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)}{2\eta_i}\right)-1\right]. \tag{42}$$

By definition, Σ_* is positive semi-definite, and so is Σ_*^{-1}. Also recalling that $\eta_i=1+1/\sqrt{i}>1$, we have

$$-(\eta_i-1)\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)^\top\Sigma_*^{-1}\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)\le 0,$$

and equation (42) is thus upper bounded by

$$\begin{aligned}\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\exp\left(\frac{\Delta_i^\top\Sigma_*^{-1}\Delta_i}{2(\eta_i-1)}\right)-1\right]&\le\sum_{i=1+N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\exp\left(\frac{\|\Delta_i\|^2\lambda_{\max}\left(\Sigma_*^{-1}\right)}{2(\eta_i-1)}\right)-1\right]\\&\le\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\exp\left(\frac{\|\Delta_i\|^2}{2(\eta_i-1)\underline{\lambda}}\right)-1\right].\end{aligned} \tag{43}$$

Since we have conditioned on the clean event $\mathcal{E}$ (defined in (33)), equation (43) does not exceed

$$\sum_{i=N_0}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\eta_i^d\exp\left(\frac{c_2(\log_e(2)d+\log_e\left(N^2T\right))}{(\eta_i-1)i}\right)-1\right]=\sum_{i=N_0}^{N}\left[\eta_i^d\exp\left(\frac{c_2\log_e\left(2^dN^2T\right)}{(\eta_i-1)i}\right)-1\right], \tag{44}$$

where we recall c_2 as $\frac{4R^2}{\underline{\lambda} c_0 \lambda_0}$, for brevity. If we plug in $\eta_i = 1 + 1/\sqrt{i}$ for all $i \geq N_0$, equation (44) then becomes

$$\sum_{i=N_0}^{N} \left[\left(1 + \frac{1}{\sqrt{i}}\right)^d \exp\left(\frac{c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}}\right) - 1 \right] \tag{45}$$

By definition of e (or $\exp(1)$), one has $(1+1/a)^a \leq e$ for all $a > 0$, and thus

$$\left(1 + \frac{1}{\sqrt{i}}\right)^d = \left(\left(1 + \frac{1}{\sqrt{i}}\right)^{\sqrt{i}}\right)^{\frac{d}{\sqrt{i}}} \leq \exp\left(\frac{d}{\sqrt{i}}\right). \tag{46}$$

Eq. (45) and inequality (46) jointly lead to

$$\sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i} \left[\left\| \frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)} \right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1 \right] \leq \sum_{i=N_0}^{N} \left[\exp\left(\frac{d}{\sqrt{i}}\right) \exp\left(\frac{c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}}\right) - 1 \right]. \tag{47}$$

To move forward, we prove the following lemma

LEMMA 6. *For any number $a > 1$,*

$$\exp\left(\frac{1}{a}\right) \leq 1 + \frac{2}{a}.$$

Proof of Lemma 6. We note that the function $f(x) = \exp(x) - 1 - 2x$ is a convex function as

$$f''(x) = e^x > 0, \tag{48}$$

as well as that $f(0) = 1 - 1 = 0$ and $f(1) = e - 3 < 0$, so $f(x) \leq 0$ for all $x \in [0, 1]$. The statement follows from the observation that $1/a \in [0, 1]$ for any $a > 1$. □

By definition of N_0 in eq. (33), *i.e.*,

$$\forall i \geq N_0 \quad \sqrt{i} \geq d \text{ and } \sqrt{i} \geq c_2 \log_e\left(2^d N^2 T\right), \tag{49}$$

we can then apply Lemma 6 to eq. (47):

$$\begin{aligned}
&\sum_{i=N_0}^{N} \left[\left(1 + \frac{2d}{\sqrt{i}}\right) \left(1 + \frac{2c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}}\right) - 1 \right] \\
\leq &\sum_{i=N_0}^{N} \left(\frac{2d + 2c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}} + \frac{4c_2 d \log_e\left(2^d N^2 T\right)}{i} \right) \\
\leq &\sum_{i=1}^{N} \left(\frac{2d + 2c_2 \log_e\left(2^d N^2 T\right)}{\sqrt{i}} + \frac{4c_2 d \log_e\left(2^d N^2 T\right)}{i} \right) \\
\leq &4\left[d + c_2 \log_e\left(2^d N^2 T\right)\right] \sqrt{N} + 4c_2 d^2 \log_e(N) + 4c_2 d \log_e(N^2 T) \log_e(N).
\end{aligned} \tag{50}$$

where inequality (50) follows from the fact that $\sum_{i=1}^{N} 1/\sqrt{i} \leq 2\sqrt{i}$ and $\sum_{i=1}^{N} 1/i \leq \log_e(N)$ Note that $d = O\left(\sqrt{N}\right)$ by assumption, we can thus derive

$$\sum_{i=N_0}^{N} \mathbb{E}_{\hat{\theta}_i} \left[\left\| \frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)} \right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1 \right] = \widetilde{O}\left(d\sqrt{N}\right). \tag{51}$$

We can now combine eq. (36), (39), (40), and (51) to arrive at the conclusion of the statement.

D. Proof of Theorem 4

Denoting $\tilde{V}_i = \left(\sum_{\tau=1}^{N_2} \boldsymbol{m}_{i,\tau}\boldsymbol{m}_{i,\tau}^\top\right)$ for every $i \in [N_1]$, we conditioned our discussion on

$$\forall i \in [N_1] \quad \lambda_{\min}\left(\tilde{V}_i\right) \geq \frac{\sqrt{c_0\lambda_0 N_2}}{2}, \tag{52}$$

which happens with probability at least $1 - 2dN_1(e/2)^{-c_1 N_2/2}$ by Lemma 5. Let $\mathcal{N}_{1/8} = \left\{v_1, \ldots, v_{\|\mathcal{N}_{1/8}\|_1}\right\}$ be the 1/8-covering of the $2d$-dimensional unit ball, then for any vector $v \in \Re^{2d}$ such that $\|v\| \leq 1$, there exists some $v_j \in \mathcal{N}_{1/8}$ such that $\|v_j - v\| \leq 1/8$, and thus

$$\begin{aligned}
v^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v =& (v_j + v - v_j)^\top\left(\hat{\Sigma}_* - \Sigma_*\right)(v_j + v - v_j) \\
=& v_j^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_j + 2(v - v_j)^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_j + (v - v_j)^\top\left(\hat{\Sigma}_* - \Sigma_*\right)(v - v_j) \\
\leq& v_j^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_j + \frac{\left\|\hat{\Sigma}_* - \Sigma_*\right\|_{op}}{4} + \frac{\left\|\hat{\Sigma}_* - \Sigma_*\right\|_{op}}{64} \qquad (53) \\
\leq& v_j^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_j + \frac{\left\|\hat{\Sigma}_* - \Sigma_*\right\|_{op}}{2},
\end{aligned}$$

where inequality (53) follows by definition of the operator norm. Rearranging the terms, we can conclude

$$\left\|\hat{\Sigma}_* - \Sigma_*\right\|_{op} \leq 2 \max_{l \in \left[\left\|\mathcal{N}_{1/8}\right\|_1\right]} v_l^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_l \tag{54}$$

and for any constant $c > 0$,

$$\begin{aligned}
\Pr\left(\left\|\hat{\Sigma}_* - \Sigma_*\right\|_{op} \geq c\right) \leq& \Pr\left(\max_{l \in [|\mathcal{N}_{1/4}|]} v_l^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_l \geq \frac{c}{2}\right) \\
\leq& \sum_{l=1}^{|\mathcal{N}_{1/4}|} \Pr\left(v_l^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_l \geq \frac{c}{2}\right) \\
\leq& \left\|\mathcal{N}_{1/4}\right\|_1 \Pr\left(\max_{v \in \mathcal{B}^d} v^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v \geq \frac{c}{2}\right) \\
\leq& 17^{2d} \Pr\left(\max_{v \in \mathcal{B}^d} v^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v \geq \frac{c}{2}\right) \qquad (55)
\end{aligned}$$

by the union bound and Lemma 11 in Appendix F. We proceed to bound the term $\Pr\left(v_l^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v_l \geq \frac{c}{2}\right)$ individually. For any fixed v in the d-dimensional unit ball, we have

$$\begin{aligned}
& v^\top\left(\hat{\Sigma}_* - \Sigma_*\right)v \\
=& v^\top\left(\frac{1}{N_1 - 1}\sum_{i=1}^{N_1}\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)^\top - \Sigma_*\right)v \\
=& v^\top\left(\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{N_1}{N_1 - 1}\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)^\top - \Sigma_*\right)v \\
=& \frac{1}{N_1}\sum_{i=1}^{N_1} v^\top\left(\frac{N_1}{N_1 - 1}\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)\left(\tilde{\theta}_i - \frac{\sum_{j=1}^{N_1}\tilde{\theta}_j}{N_1}\right)^\top - \Sigma_*\right)v \\
=& \frac{1}{N_1}\sum_{i=1}^{N_1} v^\top\left(\frac{N_1}{N_1 - 1}\left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1}\theta_j + \Delta_j}{N_1}\right)\left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1}\theta_j + \Delta_j}{N_1}\right)^\top - \Sigma_*\right)v
\end{aligned}$$

$$=\frac{\sum_{i=1}^{N_1} Z_i^2}{N_1} - v^\top \Sigma_* v, \tag{56}$$

where we have denote Δ_i as the difference between $\tilde{\theta}_i$ and θ_i for every $i \in [N_1]$ and for every $i \in [N_1]$

$$Z_i = \sqrt{\frac{N_1}{N_1 - 1}} \left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1} \theta_j + \Delta_j}{N_1} \right)^\top v, \tag{57}$$

Note that $\forall i \in [N_1]$ $\mathbb{E}[\Delta_i] = 0$ by virtue of the OLS estimator. To study the convergence property of the empirical covariance estimate, we need to bound its tail probability. We have $\mathbb{E}[Z_i] = 0$, and

$$\begin{aligned}
&\mathbb{E}[Z_i^2] \\
=&v^\top \mathbb{E}\left[\frac{N_1}{N_1 - 1} \left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1} \theta_j + \Delta_j}{N_1} \right) \left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1} \theta_j + \Delta_j}{N_1} \right)^\top \right] v \\
=&\frac{N_1}{N_1 - 1} v^\top \mathbb{E}\left[\theta_i \theta_i^\top + \theta_i \Delta_i^\top + \Delta_i \theta_i^\top + \Delta_i \Delta_i^\top - \frac{\sum_{j=1}^{N_1} \left(\theta_i \theta_j^\top + \theta_i \Delta_j^\top \right)}{N_1} - \frac{\sum_{j=1}^{N_1} \left(\Delta_i \theta_j^\top + \Delta_i \Delta_j^\top \right)}{N_1} \right. \\
&\left. - \frac{\sum_{j=1}^{N_1} \left(\theta_j \theta_i^\top + \Delta_j \theta_i^\top \right)}{N_1} - \frac{\sum_{j=1}^{N_1} \left(\theta_j \Delta_i^\top + \Delta_j \Delta_i^\top \right)}{N_1} + \frac{\left(\sum_{j=1}^{N_1} \theta_j + \Delta_j \right) \left(\sum_{j=1}^{N_1} \theta_j + \Delta_j \right)^\top}{N_1^2} \right] v \\
=&\frac{N_1}{N_1 - 1} v^\top \mathbb{E}\left[\theta_i \theta_i^\top + \Delta_i \Delta_i^\top - \frac{2\theta_i \theta_i^\top}{N_1} - \frac{2\Delta_i \Delta_i^\top}{N_1} - \frac{2(N_1 - 1)\theta_* \theta_*^\top}{N_1} + \frac{\theta_i \theta_i^\top + \Delta_i \Delta_i^\top + 2(N_1 - 1)\theta_* \theta_*^\top}{N_1} \right] v \qquad (58) \\
=&v^\top \left(\mathbb{E}\left[\theta_i \theta_i^\top - \theta_* \theta_*^\top \right] + \mathbb{E}\left[\Delta_i \Delta_i^\top \right] \right) v \\
=&v^\top \left(\Sigma_* + \mathbb{E}\left[\Delta_i \Delta_i^\top \right] \right) v, \qquad (59)
\end{aligned}$$

where we have make use of the fact that the θ_i's and the Δ_i's are mutually independent, the θ_i's are i.i.d. drawn from the feature distribution, and the Δ_i's are also i.i.d. in eq. (58) as well as the definition of Σ_* in eq. (59).

We also consider its moment generating function: for any $\lambda \in \mathbb{R}$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda Z_i \right) \right] \\
=&\mathbb{E}\left[\exp\left(\lambda \sqrt{\frac{N_1}{N_1 - 1}} \left(\theta_i + \Delta_i - \frac{\sum_{j=1}^{N_1} \theta_j + \Delta_j}{N_1} \right)^\top v \right) \right] \\
=&\mathbb{E}\left[\exp\left(\lambda \frac{\sum_{j \in [N_1], j \neq i} \left(\Delta_i - \Delta_j \right)^\top v}{\sqrt{N_1 (N_1 - 1)}} \right) \prod_{j \in [N_1], j \neq i} \exp\left(\lambda \frac{\left(\theta_i - \theta_j \right)^\top v}{\sqrt{N_1 (N_1 - 1)}} \right) \right] \\
=&\mathbb{E}\left[\exp\left(\lambda \frac{\sum_{j \in [N_1], j \neq i} \left(\Delta_i - \Delta_j \right)^\top v}{\sqrt{N_1 (N_1 - 1)}} \right) \right] \mathbb{E}\left[\prod_{j \in [N_1], j \neq i} \exp\left(\lambda \frac{\left(\theta_i - \theta_j \right)^\top v}{\sqrt{N_1 (N_1 - 1)}} \right) \right] \\
=&\mathbb{E}\left[\exp\left(\lambda \frac{\sum_{j \in [N_1], j \neq i} \left(\Delta_i - \Delta_j \right)^\top v}{\sqrt{N_1 (N_1 - 1)}} \right) \right] \mathbb{E}\left[\prod_{j \in [N_1], j \neq i} \exp\left(\lambda \frac{\left(\psi_i - \psi_j \right)^\top v}{\sqrt{N_1 (N_1 - 1)}} \right) \right],
\end{aligned}$$

where we have defined $\psi_i = (\theta_i - \theta_*) \sim \mathcal{N}(0, \Sigma_*)$ independently for every $i \in [N_1]$.

D.1. Analyzing $\mathbb{E}\left[\exp\left(\lambda \frac{\sum_{j \in [N_1], j \neq i} \left(\Delta_i - \Delta_j \right)^\top v}{\sqrt{N_1 (N_1 - 1)}} \right) \right]$

For any $j \in [N_1]$, by Cauchy-Schwarz inequality and Lemma 3, we have with probability at least $1 - \delta$,

$$\left\| \Delta_j^\top v \right\|_1 \leq \left\| \Delta_j \right\|_{\tilde{V}_i} \left\| v \right\|_{\tilde{V}_i^{-1}} \leq \frac{2\sigma \sqrt{2 \log_e(2) d + 2 \log_e \delta^{-1}}}{\sqrt{c_0 \lambda_0 N_2}} \leq \frac{2\sigma \sqrt{2d + 2 \log_e \delta^{-1}}}{\sqrt{c_0 \lambda_0 N_2}}. \tag{60}$$

In other words, for any $s > 0$,

$$\Pr\left(\left\|\Delta_j^\top v\right\|_1 \geq s\right) \leq \min\left\{1, \exp\left(d - \frac{s^2 c_0 \lambda_0 N_2}{8\sigma^2}\right)\right\}. \tag{61}$$

We then prove the following lemma regarding the moments of $\left\|\Delta_j^\top v\right\|_1$.

LEMMA 7. *For any positive integer k,*

$$\mathbb{E}\left[\left\|\Delta_j^\top v\right\|_1^k\right] \leq \left(\frac{8\sigma^2}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} \left(d^{\frac{k}{2}} + k\Gamma\left(\frac{k}{2}\right)\right).$$

Proof of Lemma 7. From inequality (61),

$$\begin{aligned}
\mathbb{E}\left[\left\|\Delta_j^\top v\right\|_1^k\right] &= \int_0^\infty \Pr\left(\left\|\Delta_j^\top v\right\|_1^k > s\right) ds \\
&= \int_0^\infty \Pr\left(\left\|\Delta_j^\top v\right\|_1 > s^{1/k}\right) ds \\
&\leq \int_0^\infty \min\left\{1, \exp\left(d - \frac{s^{2/k} c_0 \lambda_0 N_2}{8\sigma^2}\right)\right\} ds \\
&= \int_0^{\left(\frac{8\sigma^2}{c_0\lambda_0 N_2}\right)^{k/2}} 1 ds + \int_0^\infty \exp\left(-\frac{s^{2/k} c_0 \lambda_0 N_2}{8\sigma^2}\right) ds \\
&= \left(\frac{8\sigma^2 d}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} + \left(\frac{8\sigma^2}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} k \int_0^\infty \exp\left(-s'\right)\left(s'\right)^{k/2-1} ds' \qquad (62) \\
&= \left(\frac{8\sigma^2 d}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} + \left(\frac{8\sigma^2}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} k\Gamma\left(\frac{k}{2}\right) \\
&= \left(\frac{8\sigma^2}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} \left(d^{\frac{k}{2}} + k\Gamma\left(\frac{k}{2}\right)\right).
\end{aligned}$$

Here, we have made the substitution $s' = s^{2/k} c_0 \lambda_0 N_2/(8\sigma^2)$ in eq. (62) $\square$

We are now ready to provide an upper bound on the moment generating function of $\pm\Delta_j^\top v$.

LEMMA 8. *For any λ*

$$\mathbb{E}\left[\exp\left(\pm\lambda\Delta_j^\top v\right)\right] \leq \exp\left(\frac{22\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right).$$

Proof of Lemma 8. We use the Taylor expansion of the exponential function as follows: by the dominated convergence theorem and Lemma 7,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\pm\lambda\Delta_j^\top v\right)\right] \\
\leq& 1 + \sum_{k=2}^\infty \frac{\lambda^k \mathbb{E}\left[\left\|\Delta_j^\top v\right\|_1^k\right]}{k!} \\
\leq& 1 + \sum_{k=2}^\infty \left(\frac{8\lambda^2\sigma^2}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} \frac{\left(d^{\frac{k}{2}} + k\Gamma\left(\frac{k}{2}\right)\right)}{k!} \\
=& 1 + \sum_{k=2}^\infty \left(\frac{8\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} \frac{1}{k!} + \sum_{k=2}^\infty \left(\frac{8\lambda^2\sigma^2}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} \frac{k\Gamma\left(\frac{k}{2}\right)}{k!} \\
=& 1 + \sum_{k=2}^\infty \left(\frac{8\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right)^{\frac{k}{2}} \frac{1}{k!} + \sum_{k=1}^\infty \left(\frac{8\lambda^2\sigma^2}{c_0\lambda_0 N_2}\right)^{k} \frac{2k\Gamma\left(k\right)}{(2k)!} + \sum_{k=1}^\infty \left(\frac{8\lambda^2\sigma^2}{c_0\lambda_0 N_2}\right)^{k} \frac{(2k+1)\Gamma\left(k+1/2\right)}{(2k+1)!}
\end{aligned}$$

$$
\begin{aligned}
\leq & 1+\left(\frac{1}{2}+\sqrt{\frac{16\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}}\right)\sum_{k=1}^{\infty}\left(\frac{16\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right)^k\frac{1}{k!}+\left(1+\sqrt{\frac{16\lambda^2\sigma^2}{c_0\lambda_0 N_2}}\right)\sum_{k=1}^{\infty}\left(\frac{16\lambda^2\sigma^2}{c_0\lambda_0 N_2}\right)^k\frac{k!}{(2k)!} \\
\leq & 1+\left(\frac{1}{2}+\sqrt{\frac{16\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}}\right)\sum_{k=1}^{\infty}\left(\frac{16\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right)^k\frac{1}{k!}+\left(\frac{1}{2}+\sqrt{\frac{4\lambda^2\sigma^2}{c_0\lambda_0 N_2}}\right)\sum_{k=1}^{\infty}\left(\frac{16\lambda^2\sigma^2}{c_0\lambda_0 N_2}\right)^k\frac{1}{k!} \qquad (63) \\
\leq & 1+\left(1+5\sqrt{\frac{\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}}\right)\sum_{k=1}^{\infty}\left(\frac{16\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right)^k\frac{1}{k!} \\
= & \exp\left(\frac{16\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right)+6\sqrt{\frac{\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}}\left(\exp\left(\frac{16\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right)-1\right) \\
\leq & \exp\left(\frac{22\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right) \qquad (64)
\end{aligned}
$$

where inequality (63) is a consequence of the fact that $2k! \leq (2k)!$ and the last step follows from $\sqrt{a} \leq \exp(a)$.

□

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\frac{\sum_{j\in[N_1],j\neq i}(\Delta_i-\Delta_j)^\top v}{\sqrt{N_1(N_1-1)}}\right)\right] = & \mathbb{E}\left[\exp\left(\lambda\frac{(N_1-1)\Delta_i^\top v}{\sqrt{N_1(N_1-1)}}\right)\right]\prod_{j\in[N_1],j\neq i}\mathbb{E}\left[\exp\left(-\lambda\frac{\Delta_j^\top v}{\sqrt{N_1(N_1-1)}}\right)\right] \\
& \leq \exp\left(\frac{22\lambda^2\sigma^2 d}{c_0\lambda_0 N_2}\right). \qquad (65)
\end{aligned}
$$

D.2. Analyzing $\mathbb{E}\left[\prod_{j\in[N_1],j\neq i}\exp\left(\lambda\frac{(\psi_i-\psi_j)^\top v}{\sqrt{N_1(N_1-1)}}\right)\right]$

For the second term,

$$
\begin{aligned}
& \mathbb{E}\left[\prod_{j\in[N_1],j\neq i}\exp\left(\lambda\frac{(\psi_i-\psi_j)^\top v}{\sqrt{N_1(N_1-1)}}\right)\right] \\
= & \mathbb{E}\left[\exp\left(\lambda\sqrt{\frac{N_1-1}{N_1}}\psi_i^\top v\right)\right]\prod_{j\in[N_1],j\neq i}\mathbb{E}\left[\exp\left(\lambda\frac{-\psi_j^\top v}{\sqrt{N_1(N_1-1)}}\right)\right] \\
= & \exp\left(\frac{\lambda^2(N_1-1)v^\top\Sigma_* v}{2N_1}\right)\prod_{j\in[N_1],j\neq i}\exp\left(\frac{\lambda^2 v^\top\Sigma_* v}{2N_1(N_1-1)}\right) \qquad (66) \\
= & \exp\left(\frac{\lambda^2 v^\top\Sigma_* v}{2}\right) \\
\leq & \exp\left(\frac{\lambda^2\bar{\lambda}}{2}\right) \qquad (67)
\end{aligned}
$$

where eq. (66) follows from the moment generating function of multivariate normal distributions while eq. (67) follows from the fact that $v^\top\Sigma_* v \leq \|v\|^2\lambda_{\max}(\Sigma_*) \leq \bar{\lambda}$ as Σ_* is positive semi-definite and v belongs to the d-dimensional unit ball.

Combining eq. (65) and eq. (67), we know that

$$
\mathbb{E}\left[\exp\left(\lambda Z_i\right)\right] \leq \exp\left(\lambda^2\left(\frac{\bar{\lambda}}{2}+\frac{22\sigma^2 d}{c_0\lambda_0 N_2}\right)\right) \qquad (68)
$$

and $Z_i^2-\mathbb{E}[Z_i^2]$ is thus $16\left(\bar{\lambda}+\frac{44\sigma^2 d}{c_0\lambda_0 N_2}\right)$ subexponential following Lemma 12 in Appendix F, *i.e.*,

$$
\forall\|\lambda\|_1 \leq \frac{1}{4\sqrt{\left(\bar{\lambda}+\frac{44\sigma^2 d}{c_0\lambda_0 N_2}\right)}} \qquad \mathbb{E}\left[\exp\left(\lambda\left(Z_i^2-\mathbb{E}\left[Z_i^2\right]\right)\right)\right] \leq \exp\left(8\lambda^2\left(\bar{\lambda}+\frac{44\sigma^2 d}{c_0\lambda_0 N_2}\right)\right). \qquad (69)
$$

Therefore, for any v belongs to the d-dimensional unit ball and any $c>0$,

$$\begin{aligned}
&\Pr\left(v^\top\left(\hat{\Sigma}_*-\Sigma_*\right)v\geq\frac{c}{2}\right)\\
=&\Pr\left(\frac{\sum_{i=1}^{N_1}\left(Z_i^2-\mathbb{E}\left[Z_i^2\right]\right)}{N_1}+\frac{\sum_{i=1}^{N_1}\mathbb{E}\left[v^\top\Delta_i\Delta_i^\top v\right]}{N_1}\geq\frac{c}{2}\right) &(70)\\
\leq&\Pr\left(\frac{\sum_{i=1}^{N_1}\left(Z_i^2-\mathbb{E}\left[Z_i^2\right]\right)}{N_1}+\frac{\sum_{i=1}^{N_1}\mathbb{E}\left[\|v\|^2\|\Delta_i\|^2\right]}{N_1}\geq\frac{c}{2}\right) &(71)\\
\leq&\Pr\left(\frac{\sum_{i=1}^{N_1}\left(Z_i^2-\mathbb{E}\left[Z_i^2\right]\right)}{N_1}\geq\frac{c}{2}-\frac{8\sigma^2\left(d+2\right)}{c_0\lambda_0N_2}\right) &(72)\\
\leq&\exp\left(-\left(\frac{N_1\left(\frac{c}{2}-\frac{8\sigma^2(d+2)}{c_0\lambda_0N_2}\right)^2}{32\left(\bar{\lambda}+\frac{44\sigma^2d}{c_0\lambda_0N_2}\right)}\wedge\frac{N_1\left(\frac{c}{2}-\frac{8\sigma^2(d+2)}{c_0\lambda_0N_2}\right)}{\sqrt{32\left(\bar{\lambda}+\frac{44\sigma^2d}{c_0\lambda_0N_2}\right)}}\right)\right), &(73)
\end{aligned}$$

where eq. (70) makes use of the definition of Z_i's in eq. (57) as well as eq. (59), inequality (71) is Cauchy-Schwarz inequality, inequality (72) follows from the condition in eq. (52), and inequality (73) holds by Bernstein's inequality (Lemma 13 in Appendix F).

Plugging the above into inequality (55), we have

$$\Pr\left(\left\|\hat{\Sigma}_*-\Sigma_*\right\|_{op}\geq c\right)\leq 17^{2d}\exp\left(-\left(\frac{N_1\left(\frac{c}{2}-\frac{8\sigma^2(d+2)}{c_0\lambda_0N_2}\right)^2}{32\left(\bar{\lambda}+\frac{44\sigma^2d}{c_0\lambda_0N_2}\right)}\wedge\frac{N_1\left(\frac{c}{2}-\frac{8\sigma^2(d+2)}{c_0\lambda_0N_2}\right)}{\sqrt{32\left(\bar{\lambda}+\frac{44\sigma^2d}{c_0\lambda_0N_2}\right)}}\right)\right), \tag{74}$$

and this yields for any $\delta>0$,

$$\Pr\left(\left\|\hat{\Sigma}_*-\Sigma_*\right\|_{op}\geq\frac{2c_3d}{3N_2}+12\sqrt{\left(\bar{\lambda}+\frac{c_3d}{N_2}\right)}\left[\left(\frac{\log_e(17)d+\log_e\delta^{-1}}{N_1}\right)\vee\sqrt{\frac{\log_e(17)d+\log_e\delta^{-1}}{N_1}}\right]\right)\leq\delta \tag{75}$$

with

$$c_3=\frac{48\sigma^2}{c_0\lambda_0}.$$

Similarly,

$$\Pr\left(\max_{v\in\mathbb{R}^{2d}:\|v\|\leq1}v^\top\left(\Sigma_*-\hat{\Sigma}_*\right)v\geq12\sqrt{\left(\bar{\lambda}+\frac{c_3d}{N_2}\right)}\left[\left(\frac{\log_e(17)d+\log_e\delta^{-1}}{N_1}\right)\vee\sqrt{\frac{\log_e(17)d+\log_e\delta^{-1}}{N_1}}\right]\right)\leq\delta.$$

E. Proof of Theorem 5

Similar to the proof of Theorem 3 in Section C, we set $\delta=1/NT$ in Theorem 4, and define the clean event $\mathcal{E}'$:

$$\begin{aligned}
&\forall i\geq N_1\qquad\left\|\hat{\theta}_i-\theta_*\right\|\leq\frac{2R\sqrt{2\log_e(2)d+2\log_e\left(N^2T\right)}}{\sqrt{c_0\lambda_0i}},\\
&\left\|\hat{\Sigma}_*-\Sigma_*\right\|_{op}\leq\frac{4c_3d}{N_2}+12\sqrt{\left(\bar{\lambda}+\frac{c_3d}{N_2}\right)}\sqrt{\frac{\log_e(17)d+\log_e(NT)}{N_1}},\\
&\max_{v\in\mathbb{R}^{2d}:\|v\|\leq1}v^\top\left(\Sigma_*-\hat{\Sigma}_*\right)v\geq12\sqrt{\left(\bar{\lambda}+\frac{c_3d}{N_2}\right)}\sqrt{\frac{\log_e(17)d+\log_e(NT)}{N_1}}.
\end{aligned}\tag{76}$$

The meta regret can then be decomposed as follows:

$$\mathcal{R}_{N,T} = (\mathcal{R}_{N,T}|\mathcal{E}')\Pr(\mathcal{E}') + (\mathcal{R}_{N,T}|\neg\mathcal{E})\Pr(\neg\mathcal{E}') \leq (\mathcal{R}_{N,T}|\mathcal{E}') + (\mathcal{R}_{N,T}|\neg\mathcal{E}')\Pr(\neg\mathcal{E}'). \tag{77}$$

From the proof of Theorem 4 and by virtue of our choice of N_1 (*i.e.*, $N_1 \geq N_0$), we can easily see that the first part of the clean event $\mathcal{E}'$ holds with probability at most $9/(NT)$ from inequality (35). For the second part, we can apply Theorem 4, and it does not hold with probability at most

$$\begin{aligned} &\frac{2}{NT} + 4dN_1(e/2)^{-c_1 N_2/2} \\ \leq &\frac{2}{NT} + 4dN_1(e/2)^{-\log_{e/2}(2dN^2T)} \qquad (78) \\ \leq &\frac{2}{NT} + \frac{2}{NT} \\ = &\frac{4}{NT}, \end{aligned}$$

Here, inequality (78) follows by definition of N_2. A simple union bound tells us that $\mathcal{E}'$ is violated with probability at most $13/(NT)$ When the clean event $\mathcal{E}'$ is violated, the meta regret is at most $O(NT)$, and its contribution to the expected meta regret is

$$O(NT)\Pr(\neg\mathcal{E}') = O(1). \tag{79}$$

We shall condition our discussion on the clean event $\mathcal{E}'$ from now on. Similar to the proof of Theorem 3, we decompose the conditional meta regret of `Meta-DP++` algorithm as follows:

$$\mathcal{R}_{N,T}|\mathcal{E}' \leq \sum_{i=1}^{N_1} \mathcal{R}_{N,T}(i)|\mathcal{E}' + \sum_{i=N_1+1}^{N} \mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}(\theta_*, \Sigma_*)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1\right] \widetilde{O}\left(d\sqrt{T}\right). \tag{80}$$

We again analyze the two terms separately.

E.1. Analyzing $\sum_{i=1}^{N_1-1} \mathcal{R}_{N,T}(i)|\mathcal{E}'$

We begin by considering the meta regret from learning the covariance matrix. From inequality 27, the meta regret can be upper bounded as

$$\widetilde{O}\left(N_1 N_2 \|\theta\|\right) = \widetilde{O}\left(d^3 N^{\frac{3}{4}}\right) \tag{81}$$

For the rest of the rounds, the meta regret can be upper bounded as

$$N_1 \widetilde{O}\left(d\sqrt{T}\right) = \widetilde{O}\left(d^3\sqrt{NT}\right). \tag{82}$$

In total, the meta regret is of order

$$\widetilde{O}\left(d^3 N^{\frac{3}{4}} + d^3 N^{\frac{1}{2}} T^{\frac{1}{2}}\right). \tag{83}$$

E.2. Analyzing $\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}(\theta_*,\Sigma_*)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty}-1\right]\widetilde{O}\left(d\sqrt{T}\right)$

By definition of the Radon-Nikodym derivative and the multivariate normal distribution, we have

$$\begin{aligned}
&\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\left\|\frac{d\mathcal{N}\left(\theta_*,\Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right)}\right\|_{\mathcal{N}\left(\hat{\theta}_i,\Sigma_i\right),\infty}-1\right]\\
=&\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\sup_{\theta}\frac{\det\left(2\pi\Sigma_*\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)\right)}{\det\left(2\pi\Sigma_i\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\theta-\hat{\theta}_i)^\top\Sigma_i^{-1}(\theta-\hat{\theta}_i)\right)}-1\right].
\end{aligned}\tag{84}$$

Recall that $\Sigma_i=\eta_i\left(\hat{\Sigma}_*+\Sigma_{\text{correction}}\right)\succeq\eta_i\Sigma_*\ \left(\in\mathbb{R}^{2d\times 2d}\right)$, equation (84) does not exceed

$$\begin{aligned}
&\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\sup_{\theta}\exp\left(\frac{(\theta-\hat{\theta}_i)^\top\Sigma_*^{-1}(\theta-\hat{\theta}_i)/\eta_i-(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)}{2}\right)-1\right]\\
=&\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\sup_{\theta}\exp\left(\frac{(\theta-\theta_*+\theta_*-\hat{\theta}_i)^\top\Sigma_*^{-1}(\theta-\theta_*+\theta_*-\hat{\theta}_i)-\eta_i(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)}{2\eta_i}\right)-1\right]\\
=&\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\sup_{\theta}\exp\left(\frac{\Delta_i^\top\Sigma_*^{-1}\Delta_i-(\eta_i-1)(\theta-\theta_*)^\top\Sigma_*^{-1}(\theta-\theta_*)+2(\theta-\theta_*)^\top\Sigma_*^{-1}\Delta_i}{2\eta_i}\right)-1\right],
\end{aligned}$$

where we have used Δ_i to denote $\theta_*-\hat{\theta}_i$. We then complete the square for the last two terms in $\exp(\cdot)$ to arrive at

$$\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\sup_{\theta}\exp\left(\frac{\frac{\eta_i}{\eta_i-1}\Delta_i^\top\Sigma_*^{-1}\Delta_i-(\eta_i-1)\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)^\top\Sigma_*^{-1}\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)}{2\eta_i}\right)-1\right].\tag{85}$$

By definition, Σ_* is positive semi-definite, and so is Σ_*^{-1}. Also recalling that $\eta_i=1+1/\sqrt{i}>1$, we have

$$-(\eta_i-1)\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)^\top\Sigma_*^{-1}\left(\theta-\theta_*-\frac{\Delta_i}{\eta_i-1}\right)\le 0,$$

and equation (85) is thus upper bounded by

$$\begin{aligned}
\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\exp\left(\frac{\Delta_i^\top\Sigma_*^{-1}\Delta_i}{2(\eta_i-1)}\right)-1\right]&\le\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\exp\left(\frac{\|\Delta_i\|^2\lambda_{\max}\left(\Sigma_*^{-1}\right)}{2(\eta_i-1)}\right)-1\right]\\
&\le\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\exp\left(\frac{\|\Delta_i\|^2}{2(\eta_i-1)\underline{\lambda}}\right)-1\right].
\end{aligned}\tag{86}$$

Since we have conditioned on the clean event $\mathcal{E}_1$ (defined in (33)), equation (86) does not exceed

$$\begin{aligned}
&\sum_{i=N_1+1}^{N}\mathbb{E}_{\hat{\theta}_i}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\exp\left(\frac{c_2(\log_e(2)d+\log_e\left(N^2T\right))}{(\eta_i-1)i}\right)-1\right]\\
=&\sum_{i=N_1+1}^{N}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\exp\left(\frac{c_2(\log_e\left(2^dN^2T\right))}{(\eta_i-1)i}\right)-1\right]\\
\le&\sum_{i=N_1+1}^{N}\left[\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}}\exp\left(\frac{c_2(\log_e\left(2^dN^2T\right))}{\sqrt{i}}\right)-1\right],
\end{aligned}\tag{87}$$

where we have use results from Lemma 6 in the last step. Note that

$$\Sigma_i = \eta_i \left(\hat{\Sigma}_* + \Sigma_{\text{correction}} \right), \tag{88}$$

the ratio between the determinants in eq. (87) is

$$\begin{aligned}
\frac{\det\left(\Sigma_i\right)^{\frac{1}{2}}}{\det\left(\Sigma_*\right)^{\frac{1}{2}}} =& \left(\frac{\prod_{j=1}^{2d} \lambda_i(\Sigma_i)}{\prod_{j=1}^{2d} \lambda_i(\Sigma_*)} \right)^{\frac{1}{2}} \\
=& \eta_i^d \left(\frac{\prod_{j=1}^{2d} \left(\lambda_i(\Sigma_*) + \frac{2c_3 d}{3N_2} + 12\sqrt{\left(\bar{\lambda} + \frac{c_3 d}{N_2}\right)} \sqrt{\frac{\log_e(17)d + \log_e(NT)}{N_1}} \right)}{\prod_{j=1}^{2d} \lambda_i(\Sigma_*)} \right)^{\frac{1}{2}} \\
=& \eta_i^d \prod_{j=1}^{2d} \left(1 + \frac{\frac{2c_3 d}{3N_2} + 12\sqrt{\left(\bar{\lambda} + \frac{c_3}{N_2}\right)} \sqrt{\frac{\log_e(17)d + \log_e(NT)}{N_1}}}{\lambda_i(\Sigma_*)} \right)^{\frac{1}{2}} \\
\leq& \eta_i^d \prod_{j=1}^{2d} \left(1 + \frac{\frac{2c_3 d}{3N_2} + 12\sqrt{\left(\bar{\lambda} + \frac{c_3 d}{N_2}\right)} \sqrt{\frac{\log_e(17)d + \log_e(NT)}{N_1}}}{\underline{\lambda}} \right)^{\frac{1}{2}} \\
=& \exp\left(\frac{d}{\sqrt{i}}\right) \left(1 + \frac{2c_3 d}{3\underline{\lambda} N_2} + \frac{12}{\underline{\lambda}} \sqrt{\left(\bar{\lambda} + \frac{c_3 d}{N_2}\right) \left(\frac{\log_e(17)d + \log_e(NT)}{N_1} \right)} \right)^d \\
\leq& \exp\left(\frac{d}{\sqrt{i}}\right) \left(1 + \frac{2c_3 d}{3\underline{\lambda} N_2} + \frac{12}{\underline{\lambda}} \sqrt{\left(\bar{\lambda} + c_3 d\right) \left(\frac{\log_e(17)d + \log_e(NT)}{N_1} \right)} \right)^d \\
=& \exp\left(\frac{d}{\sqrt{i}}\right) \left(1 + c_4 \left(\frac{d}{N_2} + \sqrt{\frac{d^2 + d\log_e(NT)}{N_1}} \right) \right)^d,
\end{aligned} \tag{89}$$

where we recall $c_4 = \max\left\{ \frac{2c_3}{3\underline{\lambda}}, \frac{12}{\underline{\lambda}} \sqrt{\bar{\lambda} + c_3} \right\}$ and have applied the same steps as inequality (46) in Section C to the term η_i^d. With inequality (89), eq. (87) then becomes

$$\begin{aligned}
& \sum_{i=N_1+1}^{N} \left[\left(1 + c_4 \left(\frac{d}{N_2} + \sqrt{\frac{d^2 + d\log_e(NT)}{N_1}} \right) \right)^d \exp\left(\frac{d + c_2 \log_e(2^d N^2 T)}{\sqrt{i}} \right) - 1 \right] \\
=& \sum_{i=N_1+1}^{N} \left[\left(1 + \frac{1}{2N^{1/4}} + \frac{1}{2N^{1/4}} \right)^d \exp\left(\frac{d + c_2 \log_e(2^d N^2 T)}{\sqrt{i}} \right) - 1 \right] \\
=& \sum_{i=N_1+1}^{N} \left[\left(1 + \frac{1}{N^{1/4}} \right)^d \exp\left(\frac{d + c_2 \log_e(2^d N^2 T)}{\sqrt{i}} \right) - 1 \right] \\
\leq& \sum_{i=N_1+1}^{N} \left[\exp\left(\frac{d}{N^{1/4}} \right) \exp\left(\frac{d + c_2 \log_e(2^d N^2 T)}{\sqrt{i}} \right) - 1 \right].
\end{aligned} \tag{90}$$

Here, we have again made use of the same steps as inequality (46) in Section C to the term $\left(1 + \frac{1}{N^{1/4}}\right)^d$. If we further apply the results from Lemma 6 (note that $N^{1/4} \geq d$ as $N \geq d^4$), eq. (90) is no larger than

$$\sum_{i=N_1+1}^{N} \left[\left(1 + \frac{2d}{N^{1/4}} \right) \left(1 + \frac{2d + 2c_2 \log_e(2^d N^2 T)}{\sqrt{i}} \right) - 1 \right]$$

$$= \sum_{i=N_1+1}^{N} \left[\frac{2d}{N^{1/4}} + \left(1 + \frac{2d}{N^{1/4}}\right) \frac{2d + 2c_2 \log_e \left(2^d N^2 T\right)}{\sqrt{i}} \right]$$

$$\leq \sum_{i=1}^{N} \left[\frac{2d}{N^{1/4}} + \left(1 + \frac{2d}{N^{1/4}}\right) \frac{2d + 2c_2 \log_e \left(2^d N^2 T\right)}{\sqrt{i}} \right] \tag{91}$$

$$\leq 2dN^{\frac{3}{4}} + \left(1 + \frac{2d}{N^{1/4}}\right) \left[4d + 4c_2 \log_e \left(2^d N^2 T\right)\right] N^{\frac{1}{2}}, \tag{92}$$

where inequality (91) hold trivially as $N_1 \geq 1$, inequality (92) makes use of inequality $\sum_{i=1}^{N} 1/\sqrt{i} \leq 2\sqrt{i}$.

This further leads to

$$\sum_{i=N_1+1}^{N} \mathbb{E}_{\hat{\theta}_i} \left[\left\| \frac{d\mathcal{N}\left(\theta_*, \Sigma_*\right)}{d\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right)} \right\|_{\mathcal{N}\left(\hat{\theta}_i, \Sigma_i\right), \infty} - 1 \right] = \widetilde{O}\left(dN^{\frac{3}{4}}\right). \tag{93}$$

We can now combine eq. (79), (80), (83), and (93) to arrive at the conclusion of the statement.

F. Auxiliary Results

For completeness, we restate some well-known results from the literature.

The following lemma characterizes the Bayesian regret of Thompson sampling for the linear bandit.

LEMMA 9 **(Russo and Van Roy 2014)**. *Fix positive constants σ, c, and c'. Denote the set of all possible parameters as $\Theta \in \mathbb{R}^d$, the mean reward function as $f_\theta(a) = \langle \phi(a), \theta \rangle$ for some $\phi : \mathcal{A} \to \mathbb{R}$, $\sup_{\rho \in \Theta} \|\rho\| \leq c$, and $\sup_{a \in \mathcal{A}} \|\phi(a)\| \leq c'$, and for each t, the noise term is σ-subgaussian, then the Bayesian regret of the Thompson sampling algorithm is $\widetilde{O}(d\sqrt{T})$.*

The following lemma upper bounds the loss due to prior misspecification in Thompson sampling.

LEMMA 10 **(Russo and Van Roy 2014)**. *For a bandit problem parameterized by θ, if the prior over the underlying parameter θ is μ, but for convenience the decision maker selects actions as though his prior were an alternative $\tilde{\mu}$, the resulting Bayesian regret satisfies*

$$\mathbb{E}_{\theta \sim \mu} \left[\text{Regret}(\theta, \tilde{\mu})\right] \leq \left\| \frac{d\mu}{d\tilde{\mu}} \right\|_{\tilde{\mu}, \infty} \mathbb{E}_{\theta \sim \tilde{\mu}} \left[\text{Regret}(\theta, \tilde{\mu})\right]$$

where Regret(θ, ν) is the regret of the Thompson sampling algorithm Russo and Van Roy (2014) implemented with the prior ν, $d\mu/d\tilde{\mu}$ is the Radon-Nikodym derivative of μ with respect to $\tilde{\mu}$ and $\|\cdot\|_{\tilde{\mu}, \infty}$ is the essential supremum magnitude with respect to $\tilde{\mu}$.

The following lemma upper bounds the covering number of a d-dimensional unit ball.

LEMMA 11 **(Wainwright 2019)**. *For the d-dimensional unit ball, its δ covering number is upper bounded by $d \log_e(1 + 2/\delta)$.*

The following lemma makes a connection between subgaussian and subexponential random variables.

LEMMA 12 **(Rigollet and Hütter 2018)**. *If Z is a ν-subgaussian random variable, then $W = Z^2 - \mathbb{E}[Z^2]$ is 4ν-subexponential, i.e.,*

$$\forall \|s\|_1 \leq \frac{1}{4\nu} \quad \mathbb{E}\left[\exp(sW)\right] \leq \exp\left(8s^2\nu^2\right).$$

The following lemma provides a concentration inequality for subexponential random variables.

LEMMA 13 **(Bernstein Inequality)**. *Let $Z_1, \ldots, Z_m$ be independent random variables such that $\mathbb{E}[Z_i] = 0$ and each Z_i is ν-subexponential for every $i \in [m]$. Define*

$$\bar{Z} = \frac{1}{m} Z_i,$$

then for any $s > 0$,

$$\Pr\left(\left\|\bar{Z}\right\|_1 \geq s\right) \leq 2\exp\left(-\frac{m}{2}\max\left\{\frac{s^2}{\nu^2}, \frac{s}{\nu}\right\}\right).$$